International Zurich Seminar on Information and Communication (IZS 2024)

Proceedings

International Zurich Seminar on Information and Communication

March 6 – 8, 2024

Sorell Hotel Zürichberg, Zurich, Switzerland

Proceedings

Acknowledgment of Support

The International Zurich Seminar on Information and Communication (IZS) is organized by the Zurich Shannon Society and the IEEE Switzerland Chapter on Digital Communication Systems in collaboration with ETH Zurich.
The financial responsibility lies with the ZuSem-Stiftung, Zürich.

Conference Organization

General Co-Chairs

Amos Lapidoth and Stefan M. Moser

Technical Program Committee

Stephan ten Brink
Shraga Bross
David Burshtein
Yuval Cassuto
Terence H. Chan
Giuseppe Durisi
Robert Fischer
Bernard Fleury
Michael Gastpar
Albert Guillén i Fàbregas
Martin Hänggi
Franz Hlawatsch
Ashish Khisti
Tobias Koch
Yuval Kochman
Gerhard Kramer
Maël Le Treust
Hsuan-Yin Lin
Hans-Andrea Loeliger
Thomas Mittelholzer
Fernando Pérez-Cruz
Haim Permuter
Ron Roth
Igal Sason
Robert Schober
Yanina Shkel
Anelia Somekh-Baruch
Yossef Steinberg
Christoph Studer
Ido Tal
Ran Tamir
Giorgio Taricco
Emre Telatar
Pascal Vontobel
Ligong Wang
Michèle Wigger

Organizers of Invited Sessions

Thomas A. Courtade
Marco Mondelli
Gonzalo Vazquez-Vilar
Ramji Venkataramanan
Shun Watanabe

Local Organization

Olivia Bärtsch Popov (Secretary)
Michael Lerjen (Publications)
Patrick Strebel (Registration)

Table of Contents

Keynote Lectures

Session 1 Wed 09:50 – 11:30

Statistical Learning in High Dimensions

Invited session organizer: Ramji Venkataramanan (Cambridge University)

*Invited papers are marked by an asterisk.

Session 2 Wed 13:20 – 15:00
Learning and Estimation

Invited session organizer: Marco Mondelli (Institute of Science and Technology, Austria)

Session 3 Wed 15:30 – 16:50
Recent Trends in Multi-User Information Theory

Invited session organizer: Shun Watanabe (Tokyo University of Agriculture and Technology)

Session 4 Wed 17:00 – 17:40

Computation, Privacy, and Coding

Chaired by Yanina Shkel (EPFL)

Session 5 Thu 09:50 – 11:10

Hypothesis Testing and Asymptotics

Invited session organizer: Gonzalo Vazquez-Vilar (Universidad Carlos III de Madrid)

Session 6 Thu 13:20 – 15:00
Communications

Chaired by Albert Guillén i Fàbregas (Cambridge University)

Session 7 Thu 15:30 – 17:10
Decoding Mismatch and Error Exponents

Chaired by Tobias Koch (Universidad Carlos III de Madrid)

Session 8 Fri 09:50 – 11:30

Information Inequalities and Optimal Transport

Invited session organizer: Thomas A. Courtade (University of California, Berkeley)

Session 9 Fri 13:20 – 15:00

Data Science

Chaired by Christoph Studer (ETH Zurich)

Session 10 Fri 15:30 – 16:50

Statistical Theory of Communication

Chaired by Ligong Wang (ETH Zurich)

Recent-Results Posters

Wednesday, March 6

Information Velocity of Cascaded Gaussian Channels with Feedback

Elad Domanovitz (Tel Aviv University, Tel Aviv, Israel)
Anatoly Khina (Tel Aviv University, Tel Aviv, Israel)
Tal Philosof (Samsung Research, Tel Aviv, Israel)
Yuval Kochman (Hebrew University of Jerusalem, Jerusalem, Israel)

A Dominant Interferer-based Approximation for SINR Meta Distribution of UAV-assisted Wireless Communication Networks

Yujie Qin (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia)
Mustafa A. Kishk (Maynooth University, Maynooth, Ireland)
Mohamed-Slim Alouini (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia)

Thursday, March 7

Covert Authentication in MIMO Systems

Sang Wu Kim (Iowa State University, Ames, USA)

A New Error Detection Method for S-Box in Advanced Encryption Standard

Po-Hung Chen (National Yang Ming Chiao Tung University, Hsinchu, Taiwan)
Jiun-Hung Yu (National Yang Ming Chiao Tung University, Hsinchu, Taiwan)
Kai-Po Hsu (National Yang Ming Chiao Tung University, Hsinchu, Taiwan)

Friday, March 8

Near-Field Localization for Hybrid RIS-Assisted Terahertz Systems

Jiao Wu (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia)
Seungnyun Kim (Massachusetts Institute of Technology, Cambridge, USA)
Mohamed-Slim Alouini (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia)

Channel Modeling with RIS-powered Wireless Communication Systems

Kumud S. Altmayer (University of Arkansas at Little Rock, USA)
Ilya Burtakov (Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia)

Toward Immersive Underwater Cloud-Enabled Networks

Rawan Alghamdi (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia)
Hayssam Dahrouj (University of Sharjah, Sharjah, United Arab Emirates)
Tareq Y. Al-Naffouri (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia)
Mohamed-Slim Alouini (King Abdullah University of Science and Technology, Thuwal, Saudi Arabia)

A Representation-Learning Game for Classes of Prediction Tasks

Neria Uzan and Nir Weinberger
The Viterbi Faculty of Electrical and Computer Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200004, Israel
neriauzan@gmail.com, nirwein@technion.ac.il

Abstract—**We propose a game-theoretic formulation for learning dimensionality-reducing representations of feature vectors, when a prior knowledge on future prediction tasks is available. We analytically find the value of the game and optimal mixed (randomized) strategies for the case of linear representations, tasks, and the mean squared error loss, and propose an algorithm for general classes of representations, tasks, and loss functions.**

Motivation: Data of unlabeled feature vectors $\{\boldsymbol{x}_i\} \subset \mathcal{X}$ is commonly collected without a knowledge of the *specific* downstream prediction task it will be used for. When a prediction task becomes of interest, responses $\boldsymbol{y}_i \in \mathcal{Y}$ are also collected, and a learning algorithm is trained on $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$. Modern sources, such as high-definition images or genomic sequences, have high dimensionality, and this necessitates to reduce their dimensionality, either for better generalization, for storage/communication savings, or for interpretability. The goal is thus to find a low-dimensional *representation* $\boldsymbol{z} = R(\boldsymbol{x}) \in \mathbb{R}^r$, that preserves the relevant part of the features, for all possible downstream prediction tasks. Unsupervised methods for dimensionality reduction, such as principal component analysis (PCA), kernel PCA and auto-encoders [1], aim that the representation $\boldsymbol{z}$ will maximally preserve the *variation* in $\boldsymbol{x}$, and thus ignore any prior knowledge on future prediction tasks. Following a formulation proposed in [2] for the supervised learning setting, we propose a game-theoretic formulation for the case the downstream task is only known to belong to a given class.

Problem formulation: Assume that the response is drawn according to $\boldsymbol{y} \sim f(\cdot \mid \boldsymbol{x} = x)$, where $f \in \mathcal{F}$ for some known class $\mathcal{F}$. Let $\boldsymbol{z} := R(\boldsymbol{x}) \in \mathbb{R}^r$ be an r-dimensional representation of $\boldsymbol{x}$ where $R\colon \mathcal{X} \to \mathbb{R}^r$ is chosen from a class $\mathcal{R}$ of representation functions, and let $Q\colon \mathcal{X} \to \mathcal{Y}$ be a prediction rule from a class $\mathcal{Q}_{\mathcal{X}}$, with the loss function $\mathsf{loss}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. The pointwise *regret* of (R, f) is

$$\mathsf{regret}(R, f \mid P_{\boldsymbol{x}}) := \min_{Q \in \mathcal{Q}_{\mathbb{R}^r}} \mathbb{E}\left[\mathsf{loss}(\boldsymbol{y}, Q(R(\boldsymbol{x})))\right] - \min_{Q \in \mathcal{Q}_{\mathcal{X}}} \mathbb{E}\left[\mathsf{loss}(\boldsymbol{y}, Q(\boldsymbol{x}))\right].$$

The *minimax regret in mixed strategies* is the worst case response function in $\mathcal{F}$ given by

$$\mathsf{regret}_{\mathsf{mix}}(\mathcal{R}, \mathcal{F} \mid P_{\boldsymbol{x}}) := \min_{\mathsf{L}(\boldsymbol{R}) \in \mathcal{P}(\mathcal{R})} \max_{f \in \mathcal{F}} \mathbb{E}\left[\mathsf{regret}(\boldsymbol{R}, f \mid P_{\boldsymbol{x}})\right], \tag{1}$$

where $\mathcal{P}(\mathcal{R})$ is a set of probability measures on the possible set of representations $\mathcal{R}$. The *minimax regret in pure strategies* restricts $\mathcal{P}(\mathcal{R})$ to degenerated measures (deterministic), and so the expectation in (1) is removed. Our main goal is to determine the optimal representation strategy, either in pure $R^* \in \mathcal{R}$ or mixed strategies $\mathsf{L}(\boldsymbol{R}^*) \in \mathcal{P}(\mathcal{R})$.

Theoretical contribution: We address the basic setting in which the representation, the response, and the prediction are all linear functions, under the mean squared error (MSE) loss, and the class is $\mathcal{F}_S = \{\|f\|_S \leq 1\}$ for a known symmetric matrix S. Combined with the covariance matrix of the features, S determines the relevant directions of the function in the feature space, in contrast to just the features variability, as in standard unsupervised learning. We establish the optimal representation and regret in pure strategies, which shows the utility of the prior information, and in mixed strategies, which shows that randomizing the representation yields *strictly lower* regret. We prove that randomizing between merely ℓ^* different representation rules suffices, where $r + 1 \leq \ell^* \leq d$ is a precisely characterized *effective dimension*.

Algorithmic contribution: We develop an algorithm for optimizing mixed representations for general representations/response/predictors and loss functions, based only on their gradients. The algorithm operates incrementally, and at each iteration it finds the response function in $\mathcal{F}$ that is most poorly predicted by the current mixture of representation rules. An additional representation rule is added to the mixture, based on this function and the ones from previous iterations. To optimize the weights of the representation, the algorithm solves a two-player game using the classic multiplicative weights update (MWU) algorithm [3].

Further details: A full version of the paper can be found in [4].

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] Y. Dubois, D. Kiela, D. J. Schwab, and R. Vedantam, "Learning optimal representations with the decodable information bottleneck," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18674–18690, 2020.

[3] Y. Freund and R. E. Schapire, "Adaptive game playing using multiplicative weights," *Games and Economic Behavior*, vol. 29, no. 1-2, pp. 79–103, 1999.

[4] N. Uzan and N. Weinberger, "A representation-learning game for classes of prediction tasks," *In preparation.* Available at https://drive.google.com/file/d/15SAcFDRJt6qUzsausG8C3W1L1v-PKpWh/view?usp=drive_link.

The out-of-sample prediction error of the square-root-LASSO and related estimators

Cynthia Rush
Columbia University
New York, NY, USA
email: cynthia.rush@columbia.edu

Abstract—**The extent to which prediction algorithms can perform well not just on *training* data, but also on new, unseen, *testing* inputs is a central concern in machine learning. In fact, reducing a predictor's testing error—or equivalently, improving its "out-of-sample" performance or "generalization error"—possibly at the expense of increased training error, is a typical informal motivation for introducing regularization strategies in statistical estimation. More generally, the study of issues related to problems in which training and testing environments differ from one another is the subject of several recent, rapidly growing areas of research at the intersection of machine learning and statistics: transfer learning, distributional shifts, domain adaptation, adversarial attacks, learning under biased sampling and cross-domain transfer performance are some relevant examples.**

In this work, we study the classical problem of predicting an outcome variable, Y, using a linear combination of a d-dimensional covariate vector, X. We focus on linear predictors whose coefficients, $\widehat{\beta}$, solve the problem:

$$\arg\inf_{\beta\in\mathbb{R}^d}\left(\mathbb{E}_{\mathbb{P}_n}\left[\left|Y-\mathbf{X}^\top\beta\right|^r\right]\right)^{1/r}+\delta\,\rho\,(\beta)\,, \tag{1}$$

where $\delta > 0$ is a regularization parameter, $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is a convex penalty function, $\mathbb{P}_n$ is the empirical distribution of the data, and $r \geq 1$. The square-root LASSO (henceforth, $\sqrt{\text{LASSO}}$), the square-root group LASSO, the square-root sorted ℓ_1 penalized estimator (SLOPE), and the ℓ_1-penalized least absolute deviation estimator provide examples of estimators obtained by solving (1).

We are interested in studying the out-of-sample prediction error associated to such estimators; namely

$$\mathbb{E}_{\mathbb{Q}}\left[\left|Y-\mathbf{X}^\top\widehat{\beta}\right|^r\right]. \tag{2}$$

The expectation above is computed by fixing the estimated $\widehat{\beta}$, and then drawing new covariates and outcomes according to some joint distribution $\mathbb{Q}$. The distribution $\mathbb{Q}$ is similar, but not necessarily equal to, the true data generating process, $\mathbb{P}$, or the empirical distribution of the data, $\mathbb{P}_n$.

Informally, our main result is the following upper bound on the out-of-sample prediction error: If δ is chosen appropriately, then, with high probability, for any β, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}&\left[\left|Y-\mathbf{X}^\top\beta\right|^r\right]^{1/r}\leq\\ &\mathbb{E}_{\mathbb{P}_n}\left[\left|Y-\mathbf{X}^\top\beta\right|^r\right]^{1/r}+\left(\delta+\widehat{\mathcal{W}}_r(\mathbb{P},\mathbb{Q})\right)(1+\rho\,(\beta))\,,\end{aligned} \tag{3}$$

where $\widehat{\mathcal{W}}_r$ denotes a type of *max-sliced Wasserstein metric*. We will present a formal definition of this metric and explain how distributions that are close in this metric are required to have similar prediction errors (in a sense we make precise). The proof of the above is based on three intermediate results, which bring together ideas related to *distributionally robust optimization* (DRO), finite sample analysis of the max-sliced Wasserstein metric, and empirical process theory.

First, we show that estimators constructed using (1) are equivalent to those that solve a DRO problem based on a $\widehat{\mathcal{W}}_r$-ball around $\mathbb{P}_n$. The DRO representation naturally yields finite-sample bounds for (2) in terms of (1), provided that distributions $\mathbb{Q}$ are close to $\mathbb{P}_n$ in terms of our suggested metric. Thus, our first result provides theoretical support for the claim that predictors based on estimators obtained via (1) (such as the $\sqrt{\text{LASSO}}$ and related estimators) have good out-of-sample performance.

Second, we provide a detailed statistical analysis of the balls of distributions based on our suggested metric. More precisely, we determine the required size of a ball centered on $\mathbb{P}_n$ to guarantee that it contains $\mathbb{P}$ with high probability. We present both finite-sample results and large-sample approximations. Our analysis suggests that our balls are *statistically larger* than those based on the standard Wasserstein metric. Because the balls we consider are statistically larger, their radii can shrink to zero faster than order $n^{-1/d}$ (the usual rates for Wasserstein balls), and still contain $\mathbb{P}$.

Third, we use the DRO representation of (1) and the statistical analysis of our max-sliced Wasserstein balls to i) derive oracle recommendations for the penalization parameter δ that guarantee good out-of-sample prediction error; and ii) present a test statistic to rank the out-of-sample performance of two different linear estimators.

None of our results rely on sparsity assumptions about the true data generating process; thus, they broaden the scope of use of the square-root lasso and related estimators in prediction problems.

References

[1] J. Montiel Olea, C. Rush, A. Velez, and J. Wiesel, *The out-of-sample prediction error of the square-root-LASSO and related estimators,* available online: https://arxiv.org/abs/2211.07608.

Quantitative Group Testing and Pooled Data with Sublinear Number of Tests

Nelvin Tan, Pablo Pascual Cobo, and Ramji Venkataramanan
Department of Engineering, University of Cambridge

***Abstract*—In the *pooled data* problem, the goal is to identify the categories associated with a large collection of items via a sequence of pooled tests. Each pooled test reveals the number of items of each category within the pool. A prominent special case is *quantitative group testing* (QGT), which is the case of pooled data with two categories. We consider these problems in the linear regime, where the fraction of items in each category is of constant order. We propose a spatially coupled test matrix, and prove that a suitable approximate message passing (AMP) algorithm achieves *almost-exact* recovery with the number of tests sublinear in the number of items. For both QGT and pooled data, this is the first efficient scheme that provably achieves recovery in the linear regime with a sublinear number of tests.**

I. Introduction

Group testing is a problem where items are either defective or non-defective and the goal is to estimate the defective set via pooled tests, where groups of items are tested together. The original model considers binary tests where the test returns a positive outcome if there is at least one defective item present in it, and a negative outcome otherwise. Its variant, the *quantitative group testing* (QGT) model [1] is useful when tests are more informative: each test reveals the number of defective items in that pool. A more general version of QGT is the *pooled data* problem [2] where the goal is to identify the categories associated with a large collection of items via a sequence of pooled tests. Each pooled test reveals the number of items of each category within the pool.

Quantitative group testing: There are p items, whose status is denoted by the binary vector $\beta \in \{0,1\}^p$, where one represents a defective item and zero a non-defective item. Items are allocated to tests using a design (or test matrix) $X \in \{0,1\}^{n\times p}$ where n is the number of tests and p is the number of items. The ith row X_i determines the pooling design of the ith test where $X_{ij} = 1$ indicates that the jth item will participate in the ith test, and $X_{ij} = 0$ indicates otherwise. Let k be the number of defective items with $k < p$. We consider the linear regime, where each item is independently defective with a constant probability $\pi \in (0,1)$ Mathematically, the QGT model is

$$y_i = \beta^\top (X_{i,:}) \quad \text{for } i \in \{1,\dots,n\}, \tag{1}$$

where $y_i \in \mathbb{R}$ is the output of the ith test, and $X_{i,:}$ is the ith row of X represented as a column vector. The goal to is to recover β with as few tests as possible. We define the *almost-exact* recovery criterion where we have the estimate $\tilde{\beta}$ of β satisfying

$$\frac{1}{p}\sum_{j=1}^{p} \mathbb{1}\{\tilde{\beta}_j \neq \beta_j\} \to 0 \text{ as } p \to \infty. \tag{2}$$

This is a weaker notion of recovery compared to the exact recovery criterion where we want $\mathbb{P}[\tilde{\beta} \neq \beta] \to 0$ as $p \to 0$. We note that an almost-exact recovery criterion is meaningful in the linear regime but not in the sublinear regime, where $k = o(p)$, since setting $\tilde{\beta}$ to the all-zero vector satisfies (2).

Pooled data problem: The signal to be estimated is $B \in \{0,1\}^{p\times L}$, where each row is a one-hot vector. For example $B_j = [0,1,0,\dots,0]$ represents the jth item belonging category 2 (the position of one in B_j). We consider the linear regime where each item's category is independently generated from Categorical(π), where $\pi \in \mathbb{R}^L$ has positive entries that sum to 1. The model is

$$Y_{i,:} = B^\top (X_{i,:}) \in \mathbb{R}^L \text{ for } i \in \{1,\dots,n\}, \tag{3}$$

where $Y_{i,:}$ is the ith row of Y represented as a column vector. The output of each test $Y_{i,:}$ tells us the number of items from each category present in the test, which can be viewed as a histogram. Similar to QGT, denoting the estimate by $\widehat{B}$, the almost-exact recovery criterion is

$$\frac{1}{pL}\sum_{j=1}^{p}\sum_{l=1}^{L} \mathbb{1}\{\widehat{B}_{jl} \neq B_{jl}\} \to 0 \text{ as } p \to \infty.$$

The number of categories L does not grow with p.

In recent work [3], we analyzed an Approximate Message Passing (AMP) algorithm for an i.i.d. Bernoulli test design, obtaining rigorous performance guarantees for both pooled data and QGT. In this work, we use a spatially coupled Bernoulli test design and show that a suitable AMP algorithm can achieve almost-exact recovery with $n = o(p)$ tests. To our knowledge, for both QGT and pooled data, this is the first efficient scheme that provably achieves recovery in the linear regime with a sublinear number of tests.

References

[1] E. Karimi, F. Kazemi, A. Heidarzadeh, K. R. Narayanan, and A. Sprintson, "Non-adaptive quantitative group testing using irregular sparse graph codes," in *Annu. Allerton Conf. Commun. Control Comput.*, 2019.

[2] I.-H. Wang, S.-L. Huang, K.-Y. Lee, and K.-C. Chen, "Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms," in *IEEE Int. Symp. Inf. Theory*, 2016.

[3] N. Tan, P. P. Cobo, J. Scarlett, and R. Venkataramanan, "Approximate message passing with rigorous guarantees for pooled data and quantitative group testing," 2023, arXiv:2309.15507.

Neural Compression with Lattice Transform Coding

Eric Lei, Hamed Hassani, Shirin Saeedi Bidokhti
Department of Electrical and Systems Engineering, University of Pennsylvania

Neural compression has brought tremendous progress in designing lossy compressors with good rate-distortion (RD) performance at low complexity. Thus far, neural compression design involves transforming the source to a latent vector, which is then rounded to integers and entropy coded. While this approach has been shown to be optimal in a one-shot sense on certain sources, we show that it is highly sub-optimal on i.i.d. sequences, and in fact always recovers scalar quantization of the original source sequence. We demonstrate that the sub-optimality is due to the choice of quantization scheme in the latent space, and not the transform design. By employing lattice quantization instead of scalar quantization in the latent space, we demonstrate that Lattice Transform Coding (LTC) is able to recover optimal vector quantization at various dimensions and approach the rate-distortion function, with complexity polynomial in the dimension and rate. More generally, LTC also improves upon standard neural compressors in vector quantization on real-world sources.

Nonlinear Transform Coding (NTC) [1], the standard paradigm in lossy neural compression, operates on a source realization $\boldsymbol{x}$ by first transforming it to a latent vector $\boldsymbol{y}$ via an analysis transform g_a. The latent vector is then scalar quantized with Q and entropy coded. To decode, a synthesis transform g_s transforms the latent into the reconstruction $\hat{\boldsymbol{x}}$. A theory on why NTC performs well is that while the source $\boldsymbol{x}$ is high-dimensional, it typically has an intrinsic low-dimensional latent representation, which we refer to as the latent source. For theoretical sources that have this property in [2], [3], the authors show that NTC can (i) successfully recover the latent source, (ii) optimally quantize it, and (iii) map back to the original space, providing an optimal one-shot coding scheme.

However, most of the sources they analyze have a one-dimensional uniform latent source $U \sim \mathrm{Unif}([a,b])$. While these are sufficient to analyze the role of g_a and g_s, it does not provide insights on the role of the quantizer Q. This is because a uniform source U can be optimally quantized using uniform scalar quantization, which is exactly what NTC uses in the latent space. For sources with higher-dimensional latent sources, where uniform scalar quantization is not necessarily optimal, it is unclear whether NTC can still provide optimal one-shot coding schemes.

In this paper, we investigate the role of the quantization in the latent space in NTC. We first consider the challenging case where $\boldsymbol{x}$ consists of an i.i.d. sequence of some one-dimensional source $S \sim P_S$, which has no low-dimensional latent structure, and show that NTC performs no better than simply scalar quantizing the i.i.d. sequence itself. This is highly suboptimal compared to the best one-shot scheme for $\boldsymbol{x}$, whose performance should approach the (asymptotic) rate-distortion function $R(D)$ of P_S when the sequence length is large. This can be achieved through vector quantization (VQ), a classical technique that has recently been investigated for neural compression [4], [5]. However, VQ requires computational complexity that is exponential in the rate and dimension. Thus, we desire a method that can achieve rate-distortion limits, yet maintain the low complexity of scalar quantization.

To resolve this, we propose to replace Q with lattice quantization [6], which we refer to as Lattice Transform Coding (LTC). We demonstrate that LTC is able to achieve optimal coding schemes for i.i.d. sequences, while avoiding the exponential complexity of a direct codebook search under vector quantization. Our contributions are as follows.

1) We first demonstrate the inability of NTC to optimally compress i.i.d. sequences. We show that this is due to the choice of scalar quantization in the latent space.
2) We propose Lattice Transform Coding (LTC), and show that it is able to optimally compress i.i.d. sequences, yet still maintain reasonable complexity. We discuss various design choices in the transform design as well as entropy modelling that are required to recover optimality.
3) We demonstrate LTC on i.i.d. blocks of vector sources, using lattice transform coding in the latent space, and show LTC's ability to approach the rate-distortion function of the vector source. We additionally demonstrate on general sources, such as correlated vector sources, and real-world data such as image patches and audio.

References

[1] J. Ballé, P. A. Chou, D. Minnen, *et al.*, "Nonlinear transform coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2021.

[2] A. B. Wagner and J. Ballé, "Neural networks optimally compress the sawbridge," in *2021 Data Compression Conference (DCC)*, IEEE, 2021, pp. 143–152.

[3] S. Bhadane, A. B. Wagner, and J. Ballé, "Do neural networks compress manifolds optimally?" In *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 582–587.

[4] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, 6309–6318, ISBN: 9781510860964.

[5] A. El-Nouby, M. J. Muckley, K. Ullrich, I. Laptev, J. Verbeek, and H. Jegou, "Image compression with product quantized masked image modeling," *Transactions on Machine Learning Research*, 2023, ISSN: 2835-8856.

[6] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups* (Grundlehren der mathematischen Wissenschaften). New York, NY: Springer, 1999, ISBN: 978-0-387-98585-5.

The Sample Complexity of Simple Binary Hypothesis Testing

Varun Jog
University of Cambridge
Department of Pure Mathematics and Mathematical Statistics
Cambridge, United Kingdom
email: vj270@cam.ac.uk

***Abstract*—The sample complexity of simple binary hypothesis testing is the smallest number of i.i.d. samples required to distinguish between two distributions p and q such that the Type-I and Type-II errors are smaller than some pre-specified thresholds α and β, respectively. Our main contribution is deriving, under mild technical conditions, a formula for the sample complexity in terms of parameters p, q, α, and β.**

I. Introduction

Simple binary hypothesis testing is one of the most fundamental problems in statistics. Given that one of two hypotheses H_0 and H_1 is true, the statistician observes n i.i.d. samples $X_1, X_2, \ldots, X_n$ with a common distribution p under H_0, or q under H_1, supported on a discrete set $\mathcal{X}$. The statistician's goal is identify the correct hypothesis by using a decision rule ϕ that generates an output in $\{0, 1\}$ upon observing $X_1, \ldots, X_n$. There are two kinds of errors a statistician can make: Type-I error is when $\phi(X_1, \ldots, X_n) = 1$ when H_0 is true, and Type-II error is when $\phi(X_1, \ldots, X_n) = 0$ when H_1 is true. If the probabilities of the Type-I and Type-II errors are required to be at most by α and β, then the sample complexity of the decision rule ϕ is the n required to guarantee such a performance.

A fundamental result from statistics, the Neyman–Pearson theorem, states that the optimal decision rule for the statistician is the likelihood-ratio test. For some $\eta \geq 0$, the likelihood-ratio test with threshold η is to declare 0 if $\frac{p(x)}{q(x)} \geq \eta$, otherwise declare 1 (ties are broken arbitrarily). This simple test is optimal in the sense that the any other test with the same (or smaller) Type-I error than a Neyman–Pearson must have a Type-II error that larger than or equal to that of the Neyman–Pearson test.

Characterizing the sample complexity of the optimal Neyman–Pearson test up to universal multiplicative constants was first addressed in the theoretical computer science community [1] (although it was probably folklore within the statistics community before that [2]). Specifically [1] showed that under mild technical conditions, the sample complexity for obtaining Type-I and Type-II errors of at most δ is

$$\frac{c\log(1/\delta)}{d_h^2(p,q)} \leq n^*(p,q,\delta,\delta) \leq \frac{C\log(1/\delta)}{d_h^2(p,q)},$$

where c and C are some universal constants (that is, they don't depend on p, q, or δ), and d_h^2 is the Hellinger divergence between p and q given by

$$d_h^2(p,q) = \sum_{x \in \mathcal{X}} (\sqrt{p(x)} - \sqrt{q(x)})^2.$$

We also express this as $n^*(p,q,\delta,\delta) \asymp \frac{\log(1/\delta)}{d_h^2(p,q)}$. The above bound for sufficient for applications in theoretical computer science, and so the question of determining the sample complexity when the desired Type-I and Type-II errors are unequal was left unaddressed.

In our working manuscript [3], we show that Hellinger divergence is no longer the correct measure of sample complexity when errors are unequal. In particular, when α is a constant, say $1/4$, and $\beta < 1/4$ is allowed to be arbitrary small, the sample complexity is given by

$$n^*(p,q,1/4,\beta) \asymp \frac{\beta\log(1/\beta)}{JS_\beta(p,q)},$$

where the denominator is the *skewed Jensen–Shannon divergence* between p and q, defined by

$$\begin{aligned} JS_\beta(p,q) &= \beta D(p\|\beta p + (1-\beta)q) \\ &\quad + (1-\beta)D(p\|\beta p + (1-\beta)q), \end{aligned}$$

where $D(p\|q)$ is the usual Kullback–Leibler divergence. We also discuss extensions when α is allowed to vary, and to Bayesian hypothesis testing settings.

References

[1] Bar-Yossef, Ziv. “The complexity of massive data set computations.” Thesis (2002).

[2] Donoho, David L., and Richard C. Liu. "Geometrizing rates of convergence, II." The Annals of Statistics (1991): 633-667.

[3] Pensia, Ankit and Loh, Po-Ling and Jog, Varun. “Working paper.” (2023).

Information-Theoretic Limits for Sublinear-Rank Symmetric Matrix Factorization

Jean Barbier*, Justin Ko†, and Anas A. Rahman*
*International Centre for Theoretical Physics, Trieste, Italy
†University of Waterloo, Canada and ENS Lyon, France

***Abstract*—We consider symmetric matrix factorization with additive Gaussian noise when the rank M scales with the signal-matrix size N as $M_N = \mathrm{o}(N^{1/10})$. Allowing for a growing rank offers new challenges and requires new methods. Working in the Bayes-optimal setting, we show that whenever the log-concave prior for the matrix elements takes a factorized form, then the limiting mutual information between signal and observation is the very same as when the rank equals one (namely, the standard spike Wigner model), as first conjectured in [1]. The proof is primarily based on a novel "two-step" application of the cavity method allowing for growing rank.**

Spiked matrix models of the form "signal+noise" were introduced as simple statistical models of PCA [2] and have now become a model of choice for the development of novel theoretical and algorithmic approaches [3]–[8]. In this work, we consider the archetypal *spiked Wigner model*, where the data $\boldsymbol{Y}$ is generated as

$$\boldsymbol{Y} = \sqrt{\lambda/N}\boldsymbol{X}_0\boldsymbol{X}_0^\intercal + \boldsymbol{Z}$$

where $\boldsymbol{X}_0 \in \mathbb{R}^{N\times M}$ is the signal, $\boldsymbol{Z} \in \mathbb{R}^{N\times N}$ is a standard Wigner noise matrix with $\boldsymbol{Z} \sim \exp(-\frac{1}{2}\mathrm{Tr}\boldsymbol{Z}^2)$, λ is the signal to noise ratio. The task is to infer the spike $\boldsymbol{X}_0\boldsymbol{X}_0^\intercal$ given $\boldsymbol{Y}$. Most statistical analyses of this model focused on $M = 1$ or finite (i.e., N-independent) [7], [8], with recent studies venturing into the regime of growing rank. In particular, an M-dimensional variational formula for the limiting mutual information can be surmised from [9] when $M = \mathrm{o}(N^{1/20})$, which is however intractable in practice as it is still M-dependent. We instead prove a tractable low-dimensional formula for factorized priors when $M = \mathrm{o}(N^{1/10})$ — this is a significant milestone towards understanding the recent observation of [1] that the rank-M spiked Wigner model should behave as its rank-one counterpart, so long as $M = \mathrm{o}(N)$.

Our main technical contribution is a generalization of the *Aizenman–Sims–Starr scheme* [10] to accomodate for the growing rank. Using this new tool we prove that the variational formula for the mutual information reduces to the known one of the rank-one spiked Wigner model [6]–[8].

Theorem 1. *Let $M = \mathrm{o}(N^{1/10})$, let the signal $\boldsymbol{X}_0$ be made of i.i.d. entries with common law $\mathbb{P}_X$ which is even, log-concave and with bounded support. Then the mutual information per variable $I(\boldsymbol{X}_0;\boldsymbol{Y})/(NM)$ tends, as $N \to \infty$, to the same limit as in the case $M = 1$, rigorously analyzed in [6]–[8].*

We believe that the multi-step cavity method introduced in this work (which will appear soon [11]) as well as the class of sublinear-rank inference or spin glass models are crucial steps towards understanding the challenging extensive regime $M = \Theta(N)$, see the recent works [1], [12]–[15].

Acknowledgement

The authors acknowledge funding by the European Union (ERC projects 101039794 and 884584) and crucial discussions with Francesco Camilli and Koki Okajima.

References

[1] F. Pourkamali, J. Barbier, and N. Macris, "Matrix inference in growing rank regimes," *arXiv preprint arXiv:2306.01412*, 2023.

[2] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001.

[3] S. B. Korada and N. Macris, "Exact solution of the gauge symmetric p-spin glass model on a complete graph," *Journal of Statistical Physics*, vol. 136, pp. 205–230, 2009.

[4] Y. Deshpande and A. Montanari, "Information-theoretically optimal sparse pca," in *2014 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2014, pp. 2197–2201.

[5] Y. Deshpande, E. Abbe, and A. Montanari, "Asymptotic mutual information for the balanced binary stochastic block model," *Information and Inference: A Journal of the IMA*, vol. 6, no. 2, pp. 125–170, 2017.

[6] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula," in *Advances in Neural Information Processing Systems 29*. NIPS, 2016, pp. 424–432.

[7] M. Lelarge and L. Miolane, "Fundamental limits of symmetric low-rank matrix estimation," *Probability Theory and Related Fields*, vol. 173, no. 3, pp. 859–929, 2019.

[8] C. Luneau, J. Barbier, and N. Macris, "Mutual information for low-rank even-order symmetric tensor estimation," *Information and Inference: A Journal of the IMA*, vol. 10, no. 4, pp. 1167–1207, 2021.

[9] G. Reeves, "Information-theoretic limits for the matrix tensor product," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 777–798, 2020.

[10] M. Aizenman, R. Sims, and S. L. Starr, "Extended variational principle for the sherrington-kirkpatrick spin-glass model," *Phys. Rev. B*, vol. 68, p. 214403, Dec 2003. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.68.214403

[11] J. Barbier, J. Ko, and A. Rahman, "Information-theoretic limits for sublinear-rank symmetric matrix factorization," 2023.

[12] J. Barbier and N. Macris, "Statistical limits of dictionary learning: Random matrix theory and the spectral replica method," *Phys. Rev. E*, vol. 106, no. 2, p. 024136, 2022.

[13] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, "Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2022, no. 8, p. 083301, 2022.

[14] A. Bodin and N. Macris, "Gradient flow on extensive-rank positive semi-definite matrix denoising," 2023.

[15] S. Tarmoun, G. Franca, B. D. Haeffele, and R. Vidal, "Understanding the dynamics of gradient flow in overparameterized linear models," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 10 153–10 161.

Weak decoding and symmetries

Emmanuel Abbe
EPFL
Mathematics Institute

Colin Sandon
EPFL
Mathematics Institute

***Abstract*—Traditional channel coding theory requires a vanishing block error probability for the reliability criterion. In this note, we consider a *weak decoding* reliability criterion, requiring only the *local* error (i.e., decoding a bit incorrectly given the other noisy bits) to be *non-trivial* (i.e., bounded away from half). For this weak decoding criterion, we provide a simple general result: any linear transitive code achieves weak decoding below capacity on any symmetric channel. Put in the light of the new boosting framework of [1], [2], this gives a useful base case for a wide variety of codes.**

Definition 1. *For a linear code and a symmetric*[1] *channel, let X be an n-dimensional codeword, Y its output on the channel, $\hat{X}(Y)$ the maximum likelihood (ML) decoding of X based off Y, and for $i \in [n]$, let $\hat{X}_i(Y)$ and $\hat{X}_i(Y_{-i})$ be the ML decoding of X_i based off Y and $\{Y_j\}_{j \neq i}$ respectively. Define also $P_{\text{glo}} = P(X \neq \hat{X}(Y))$, $P_{\text{bit},i} = P(X_i \neq \hat{X}_i(Y))$, $P_{\text{bit}} = \max_{i\in[n]} P_{\text{bit},i}$, $P_{\text{loc},i} = P(X_i \neq \hat{X}_i(Y_{-i}))$, $P_{\text{loc}} = \max_{i\in[n]} P_{\text{loc},i}$, $\bar{P}_{\text{loc}} = (1/n)\sum_{i\in[n]} P_{\text{loc},i}$.*

Note that the above measures are independent of the codeword choice since the code is linear. Also $P_{\text{bit}} = o_n(1)$ is equivalent to $P_{\text{loc}} = o_n(1)$, and for a BSC(ϵ), $P_{\text{bit}} < \epsilon \wedge (1-\epsilon)$ implies $P_{\text{loc}} < 1/2$.

Lemma 1. *For a symmetric channel of capacity C and a linear code of rate $R < C$, there exist $\Omega(n)$ values of j such that $P_{\text{loc},j} = 1/2 - \Omega(1)$.*

Proof. If the only valid codeword is 0 then $P_{\text{loc},j} = 0$ for all j. Otherwise let i be such that not all the codewords have their i-th bit set to 0. Since the code is linear, it must be the case that $H(X_i) = 1$. Note that $H(Y) \leq H(X) + H(Y|X) \leq n(H(Y_i) - (C - R))$. Since $H(Y) = \sum_{i=1}^n H(Y_i|Y_{<i})$, there exist $\Omega(n)$ values of j such that $H(Y_j|Y_{<j}) = H(Y_i) - \Omega(1)$. For any such j, either $H(X_j|Y_{-j}) = H(X_j) = 0$ or $H(X_j) = 1$ and $I(X_j; Y_{-j}) \geq I(Y_j; Y_{-j}) \geq I(Y_j; Y_{<j}) \geq \Omega(1)$; either way, $P_{\text{loc},j} = 1/2 - \Omega(1)$. □

Corollary 1. *For a symmetric channel of capacity C, a linear code of rate $R < C$ has $\bar{P}_{\text{loc}} = 1/2 - \Omega(1)$ and a linear transitive code of rate $R < C$ has $P_{\text{loc}} = 1/2 - \Omega(1)$.*

Definition 2. *A code sequence (indexed by the blocklength n) achieves weak decoding on a channel if $P_{\text{loc}} = 1/2 - \Omega_n(1)$.*

Conclusion and implications. We showed that linear transitive codes achieve weak decoding below capacity on any symmetric channel. This is a weak notion of decoding, requiring the local error (as defined above) to have non-trivial probability. This is much weaker than considering the block error and requiring a vanishing rate. This criterion could be tackled by the rate-distortion theory (RDT) framework, but here we focus on rates below the traditional channel capacity (while RDT further says that one can obtain weak decoding even beyond the capacity). The note shows that by the sole property of being a linear code operating below capacity, this local error is non-trivial on average, and if the code is further transitive, it is non-trivial for every coordinate. While this is a simple observation, it is important once put in light of the new boosting framework of [1], [2], since this gives a non-trivial base case for a wide variety of codes. More specifically, the camellia boosting as defined in [2] (or the sunflower boosting variant of [1]) allows to boost such a weak local decoding into a strong local decoding, where the local error becomes vanishing (rather than just non-trivial). In order for such boosting to operate, one needs to be able to aggregate enough spread subcodes of the original code (with the guarantees derived here), and this can be obtained for a broad class of symmetric codes (such as Reed-Muller codes).

References

[1] Emmanuel Abbe, Colin Sandon, *A proof that Reed-Muller codes achieve Shannon capacity on symmetric channels,* In Proc. FOCS, 2023. 1

[2] Emmanuel Abbe, Colin Sandon, *Reed-Muller codes have vanishing bit-error probability below capacity: a simple tighter proof via camellia boosting,* arXiv:2312.04329, 2023. 1

[1] A symmetric channel can be viewed as a mixture of binary symmetric channels (BSCs), i.e., independently for each i, $Y_i = (\epsilon_i, X_i \oplus w_i)$ with ϵ_i in $[0, 1/2]$ drawn under some distribution independently of $w_i \sim \text{Ber}(\epsilon_i)$.

Equivalence Principles for Nonlinear Random Matrices

Yue M. Lu
Harvard University
John A. Paulson School of Engineering and Applied Sciences
Cambridge, MA 02138, United States
email: yuelu@seas.harvard.edu

***Abstract*—Nonlinear random matrices have significant applications in machine learning, statistics, and signal processing. Recent results in the field have pointed to an intriguing equivalence principle for these matrices. This principle shows that their asymptotic properties, including but not limited to their spectral characteristics, are asymptotically equivalent to those of simpler, noisy linear equivalent models. In my presentation, I will discuss these recent findings, shedding light on their implications and applications in characterizing the performance of random feature regression and kernel ridge regression in high-dimensional settings.**

I. Introduction

Nonlinear random matrices and their spectral properties play crucial roles in several problems in machine learning, statistics, and signal processing. Examples include kernel methods (such as kernel-PCA [1] and kernel-SVM [2]), covariance thresholding procedures [3], [4], nonlinear dimension reduction [5], and probabilistic matrix factorization [6].

For example, a random inner-product kernel matrix has the form of

$$A_{ij} \overset{\text{def}}{=} \begin{cases} \frac{1}{\sqrt{n}} f_d(\sqrt{d}\, \boldsymbol{x}_i^\top \boldsymbol{x}_j) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, \tag{1}$$

where $f_d : \mathbb{R} \mapsto \mathbb{R}$ is a (nonlinear) "kernel" function, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ is a set of independent data vectors drawn from a given distribution. A closely-related non-Hermitian version of (1), where $A_{ij} = \frac{1}{\sqrt{n}} f_d(\sqrt{d} \boldsymbol{x}_i \boldsymbol{y}_j)$ for two sets of vectors $\{\boldsymbol{x}_i\}_{i \le n}$ and $\{\boldsymbol{y}_j\}_{j \le p}$, appears in the random feature model [7]–[10], an interesting theoretical model for large random neural networks.

II. An Equivalence Principle

In my presentation at the seminar, I will discuss several related recent results that point to a general equivalence principle for many such nonlinear random matrices. For example, it was shown in [11] that, when $n/d^\ell \to \kappa \in (0, \infty)$ for any $\ell \in \mathbb{N}$ and when the data vectors $\{x_i\}_{i \le n}$ are sampled from the spherical distribution, the empirical spectral distribution (ESD) of A is asymptotically equivalent to that of

$$B = \frac{\mu_\ell}{\sqrt{n N_\ell}} (W^\top W - N_\ell I) + \gamma_\ell H$$

where $W \in \mathbb{R}^{N_\ell \times n}$ is an i.i.d. standard Gaussian matrix with aspect ratio $N_\ell / n \to 1/(\kappa \ell!)$, and H is a GOE matrix independent of W. Both constants μ_ℓ and γ_ℓ depend on ℓ and can be determined by expanding f_d in the orthogonal Hermite polynomial basis. As a direct consequence of this equivalence principle, the limiting ESD of A can be characterized simply as a free additive convolution between a (shifted) Marchenko-Pastur (MP) law and a semicircle law. More recently, this result was further extended in [12] to the case where the data vectors $\{x_i\}_{i \le n}$ are sampled from general distributions that are i.i.d. over the coordinates. Similar equivalence principles have also been employed to characterize the performance of random feature regression [13] and kernel ridge regression [14] in high-dimensional settings.

References

[1] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, pp. 1299–1319, July 1998.

[2] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Adaptive Computation and Machine Learning, Cambridge, Mass: MIT Press, 2002.

[3] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[4] Y. Deshpande and A. Montanari, "Sparse pca via covariance thresholding," in *Advances in Neural Information Processing Systems*, pp. 334–342, 2014.

[5] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, vol. 15, pp. 1373–1396, June 2003.

[6] A. Mnih and R. R. Salakhutdinov, "Probabilistic Matrix Factorization," in *Advances in Neural Information Processing Systems* (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), vol. 29, 2007.

[7] A. Rahimi and B. Recht, "Random Features for Large-Scale Kernel Machines," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, p. 8, 2007.

[8] C. Louart, Z. Liao, and R. Couillet, "A Random Matrix Approach to Neural Networks," *arXiv:1702.05419 [cs, math]*, June 2017.

[9] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," *arXiv:1903.08560 [cs, math, stat]*, Dec. 2020.

[10] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, p. 124005, Dec. 2019.

[11] Y. M. Lu and H.-T. Yau, "An Equivalence Principle for the Spectrum of Random Inner-Product Kernel Matrices with Polynomial Scalings," May 2023.

[12] S. Dubova, Y. M. Lu, B. McKenna, and H.-T. Yau, "Universality for the global spectrum of random inner-product kernel matrices in the polynomial regime," Oct. 2023.

[13] H. Hu and Y. M. Lu, "Universality Laws for High-Dimensional Learning with Random Features," *arXiv:2009.07669 [cs, math]*, Mar. 2021.

[14] L. Xiao, H. Hu, T. Misiakiewicz, Y. M. Lu, and J. Pennington, "Precise Learning Curves and Higher-Order Scaling Limits for Dot Product Kernel Regression," June 2023.

Hyperparameter tuning via trajectory predictions: Stochastic prox-linear methods in matrix sensing

Mengqi Lou[†], Kabir Aladin Verchand[*], and Ashwin Pananjady[†‡]
[*]Statistical Laboratory, University of Cambridge
Cambridge, UK
email: kav29@cam.ac.uk
[†]Georgia Institute of Technology, School of Industrial and Systems Engineering, {mlou30, ashwinpm}@gatech.edu
[‡]Georgia Institute of Technology, School of Electrical and Computer Engineering, ashwinpm@gatech.edu

We consider estimating a rank one matrix $\boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top \in \mathbb{R}^{d\times d}$ from i.i.d. observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ drawn in an online, mini-batched fashion according to the model $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\mu}_\star\rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu}_\star\rangle + \epsilon_i$. To do so, we consider minimizing the population loss corresponding to the negative log-likelihood, namely $\bar{L}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbb{E}\{\big(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\mu}\rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu}\rangle\big)^2\}$, which we emphasize is a non-convex function of the inputs. Towards minimizing the population loss, consider an iterate $(\boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$. We take mini-batches of size m with samples $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)_{i=1}^m$ and form the data $\boldsymbol{a}_i^\top = \begin{bmatrix}\boldsymbol{z}_i^\top \boldsymbol{\nu}_t \boldsymbol{x}_i^\top & \boldsymbol{x}_i^\top \boldsymbol{\mu}_t \boldsymbol{z}_i^\top\end{bmatrix}$ for each $1 \le i \le m$; define the pair of diagonal matrices $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{X}\boldsymbol{\mu}_t)$, $\widetilde{\boldsymbol{W}} = \mathrm{diag}(\boldsymbol{Z}\boldsymbol{\nu}_t)$; and collect the vectors $\boldsymbol{a}_i$ into a concatenated data matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \mid \boldsymbol{a}_2 \mid \dots \mid \boldsymbol{a}_m]^\top = \big[\widetilde{\boldsymbol{W}}\boldsymbol{X} \mid \boldsymbol{W}\boldsymbol{Z}\big] \in \mathbb{R}^{m\times 2d}$. We then consider the following stochastic prox-linear update to define the next iterate $(\boldsymbol{\mu}_{t+1}, \boldsymbol{\nu}_{t+1})$

$$\begin{bmatrix}\boldsymbol{\mu}_{t+1}\\ \boldsymbol{\nu}_{t+1}\end{bmatrix} = \boldsymbol{A}_\lambda^{-1}\Big(\boldsymbol{A}^\top\big(\boldsymbol{y} + \mathrm{diag}(\boldsymbol{W}\widetilde{\boldsymbol{W}})\big) + \lambda m \begin{bmatrix}\boldsymbol{\mu}_t\\ \boldsymbol{\nu}_t\end{bmatrix}\Big),$$

where λ denotes an inverse step-size parameter and $\boldsymbol{A}_\lambda = \boldsymbol{A}^\top\boldsymbol{A} + \lambda m\boldsymbol{I}$. Our main contribution is to provide a deterministic prediction of the trajectory of the iterative method defined in the previous display under the pair of assumptions $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i\ge 1} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \boldsymbol{I}_d)$ and $\|\boldsymbol{\mu}_\star\|_2 = \|\boldsymbol{\nu}_\star\|_2 = 1$. More concretely, we obtain the following.

a) Sharp, deterministic predictions which adapt to problem error: Consider running one-step of the prox-linear update starting from a pair $(\boldsymbol{\mu}_\sharp, \boldsymbol{\nu}_\sharp)$ and let $[\boldsymbol{\mu}_+^\top \mid \boldsymbol{\nu}_+^\top]^\top$ denote the next iterate. For *all* minibatch sizes $1 \le m \le d$ and a large range of step-sizes $\lambda \gtrsim (1+\sigma)d/m$, we derive an explicit, deterministic, four-dimensional prediction that closely tracks the error of its empirical counterparts. We additionally prove a non-asymptotic guarantee on our predictions, showing that its fluctuations scale as $\frac{\|\boldsymbol{\mu}_\sharp\boldsymbol{\nu}_\sharp^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F + \sigma}{\lambda\sqrt{m}}$, up to poly-logarithmic in dimension factors. Note that this guarantee—in contrast to previous work [1], [2]—provides bounds on the deviation which scale with the *current* estimation error $\|\boldsymbol{\mu}_\sharp\boldsymbol{\nu}_\sharp^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F$. This, in turn, enables a transparent convergence analysis of the iterations for all noise levels $\sigma \ge 0$.

Our proof reposes on a variant of El Karoui, et. al's leave-one-out method [3]. In particular, given the ground truth $\boldsymbol{\mu}_\star$ and a current iterate $\boldsymbol{\mu}_\sharp$, we let $\mathbb{U} = \{\boldsymbol{\mu}_\star, \boldsymbol{P}_{\boldsymbol{\mu}_\star}^\perp\boldsymbol{\mu}_\sharp/\|\boldsymbol{P}_{\boldsymbol{\mu}_\star}^\perp\boldsymbol{\mu}_\sharp\|_2, \boldsymbol{u}_3, \dots, \boldsymbol{u}_d\}$ denote an orthonormal basis of $\mathbb{R}^d$. We obtain a closed form expression for each of the projections $\langle\boldsymbol{\mu}_+, \boldsymbol{u}\rangle$, $\boldsymbol{u} \in \mathbb{U}$. We then use standard tools in random matrix theory to obtain deterministic predictions of each of these projections.

b) Fine-grained convergence analysis: We use our deterministic predictions to execute an iterate-by-iterate analysis of the stochastic prox-linear algorithm from a local initialization. This analysis reveals several fine-grained properties of the convergence behavior. In particular, for the step-size choice $\lambda^{-1} \asymp m/(d(1+\sigma^2))$ and batch size $m \gtrsim \mathsf{polylog}(d)$, we show that it takes $\tau = \Theta\Big(\frac{d(1+\sigma^2)}{m} \cdot \log\Big(\frac{1}{\sigma^2}\Big)\Big)$ many iterations in order to guarantee an error $\|\boldsymbol{\mu}_\tau\boldsymbol{\nu}_\tau^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F^2 \lesssim \sigma^2$. This reveals a linear speed-up in the batch size m for *all* noise levels $\sigma \ge 0$. As a consequence, the total sample complexity for reaching estimation error σ^2 is $O(d(1+\sigma^2)\log(1/\sigma^2))$. Moreover, for other step-size choices $\lambda^{-1} \lesssim m/(d(1+\sigma^2))$, we show that it takes $\tau = \Theta\Big(\lambda \cdot \log\Big(\frac{\lambda m}{d\sigma^2}\Big)\Big)$ many iterations to guarantee an error $\|\boldsymbol{\mu}_\tau\boldsymbol{\nu}_\tau^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F^2 \lesssim \frac{\sigma^2 d}{\lambda m}$, which in turn quantifies the dependence of the convergence behavior on the step-size λ^{-1}. That is, decreasing the step-size λ^{-1} introduces a tension between the increasing iteration complexity and decreasing eventual estimation error. Note that our guarantees on iteration complexity are sharp in the sense that our bounds provide both upper and lower bounds on the rate of convergence.

Our convergence proofs rely on properties of the deterministic predictions. In particular, we first prove that the deterministic predictions enjoy sharp linear convergence. We then apply the deviation bounds on the deterministic predictions to transfer this property to the empirical iterates.

References

[1] K.A. Chandrasekher, M. Lou, A. Pananjady, "Alternating minimization for generalized rank one matrix sensing: Sharp predictions from a random initialization," to appear in *Algorithmic Learning Theory*, San Diego, CA, 2024.

[2] K.A. Chandrasekher, A. Pananjady, C. Thrampoulidis, "Sharp global convergence guarantees for iterative nonconvex optimization with random data," *Annals of Statistics*, vol. 51, pp. 179–210, 2023.

[3] N. El Karoui, D. Bean, P.J. Bickel, C. Lim, B. Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 14557–14562, 2013.

Phase Transitions for Spectral Estimators in Generalized Linear Models via Approximate Message Passing

Marco Mondelli
Institute of Science and Technology Austria (ISTA)
email: marco.mondelli@ist.ac.at

In a generalized linear model (GLM), the goal is to estimate a d-dimensional signal $x^* \in \mathbb{R}^d$ from an n-dimensional observation $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ given by

$$y_i = q(\langle a_i, x^* \rangle, \epsilon_i), \;\; i \in \{1, \ldots, n\}. \tag{1}$$

Here, the covariate vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ are known, and the vector $(\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ contains unknown i.i.d. random variables accounting for noise in the measurements. The nonlinearity q generalizes linear regression ($q(g, \epsilon) = g + \epsilon$) and incorporates a wide range of problems, e.g., phase retrieval ($q(g, \epsilon) = |g| + \epsilon$), 1-bit compressed sensing ($q(g, \epsilon) = \text{sign}(g) + \epsilon$), and logistic regression.

Spectral methods provide a popular solution to obtain an initial estimate, and they are also commonly used as a ‘warm start’ for other algorithms. In particular, the spectral estimator processes the observations via a function $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ and outputs the principal eigenvector of the following matrix:

$$D = \sum_{i=1}^{n} a_i a_i^\top \mathcal{T}(y_i) \in \mathbb{R}^{d \times d}. \tag{2}$$

To understand the power of spectral estimators, it is crucial to: *(i)* characterize their performance in terms of, e.g., the normalized correlation between the signal and the spectral estimate, and *(ii)* design the best preprocessing function $\mathcal{T}$ that minimizes the sample complexity, i.e., the number n of observations required to attain a desired limiting overlap.

The answers to these questions are well understood when the covariates are i.i.d. Gaussian. Using tools from random matrix theory, [1], [2] obtained tight results in the proportional regime where $n, d \to \infty$ and $n/d \to \delta$ for a fixed constant $\delta \in (0, \infty)$. Specifically, a *phase transition* phenomenon was established: if δ surpasses a critical value (referred to as the “spectral threshold”), then *(i)* a spectral gap emerges between the first two eigenvalues of D, and *(ii)* the spectral estimator attains non-vanishing correlation with x^*; otherwise, *(i)* no outlier is present to the right of the spectrum of D, and *(ii)* the spectral estimator is asymptotically independent of x^*.

However, i.i.d. Gaussian covariates fail to capture the heterogeneity and structure of data typical in applications. To capture data heterogeneity, a popular solution is to consider *mixed generalized linear models*: the objective is to learn multiple signals from unlabeled observations; each sample comes from exactly one signal, but it is not known which one. To capture data structure, a popular solution is to consider *general (correlated) designs*: the covariates $a_1, \ldots, a_n$ have an arbitrary (and, often, unknown) covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$.

The strategy of [1], [2] was to reduce the spectral matrix D in Eq. (2) to a rank-1 perturbation. However, when the model is mixed, additional terms appear which are difficult to control. Furthermore, when the covariates have a covariance Σ, the model loses its rotational invariance, which is crucially exploited by existing approaches. To overcome these difficulties, we propose a novel proof strategy based on approximate message passing (AMP). AMP refers to a family of iterative algorithms that were first proposed for linear regression, but have then been applied to various statistical estimation problems, including parameter recovery in a GLM. A crucial feature of AMP is the presence of a memory term, which debiases the iterates, ensuring that their joint empirical distribution is asymptotically Gaussian. This in turn allows to track their covariance structure via a low-dimensional recursion known as *state evolution*.

Our idea is to design and analyze an AMP that simulates a power iteration and, hence, approaches the leading eigenvector(s) of D. Then, by leveraging state evolution, we provide a precise asymptotic characterization of the normalized correlation between the spectral estimators and the signals, see [3] for a mixed model and [4] for a model with a correlated design. As a consequence, we can optimize the preprocessing $\mathcal{T}$, which leads to a significant improvement in performance with respect to existing heuristic approaches. We highlight that our methodology based on AMP is broadly applicable, and it opens the way to the study of spiked matrices and of the corresponding spectral estimators in a variety of settings.

References

[1] Y. M. Lu and G. Li, “Phase transitions of spectral initialization for high-dimensional non-convex estimation,” *Information and Inference: A Journal of the IMA*, vol. 9, no. 3, pp. 507–541, 2020.

[2] M. Mondelli and A. Montanari, “Fundamental limits of weak recovery with applications to phase retrieval,” *Foundations of Computational Mathematics*, vol. 19, no. 3, pp. 703–773, 2019.

[3] Y. Zhang, M. Mondelli, and R. Venkataramanan, “Precise asymptotics for spectral methods in mixed generalized linear models,” *arXiv preprint arXiv:2211.11368*, 2022.

[4] Y. Zhang, H. C. Ji, R. Venkataramanan, and M. Mondelli, “Spectral estimators for structured generalized linear models via approximate message passing,” *arXiv preprint arXiv:2308.14507*, 2023.

Gaussian Semantic Source Coding

Peiyao Chen, Jun Chen, Yuxuan Shi, Shuo Shao, Yongpeng Wu, Timothy N. Davidson

Abstract—**Semantic source coding differs from conventional source coding in the sense that the decoder is required to reconstruct, possibly in a lossy fashion, not only the observable source realization but also an intrinsic source state that carries certain semantic information. Centralized Gaussian semantic source coding and its distributed counterpart are studied in this work. We explicitly characterize their respective rate-distortion functions for the symmetric setting and the two-component setting via the analysis of the associated convex optimization problems, which generalize several classical results on quadratic vector Gaussian source coding and Gaussian multiterminal source coding.**

I. Introduction

Direct source coding [1] aims to find an efficient representation using a bit sequence based on which the observable source realization can be reconstructed exactly or approximately. In contrast, indirect source coding [2]–[5] deals with the situation where the object of interest is not the observable part but some hidden state. These two coding problems are closely related. In fact, it is known that indirect source coding can be reduced to direct source coding for a sufficient statistic of the observable part with respect to the hidden state under a suitably constructed surrogate distortion measure.

Semantic source coding [6], [7] couples the aforementioned two coding problems by requiring the decoder to reconstruct, possibly in a lossy fashion, both the observable source realization and the hidden source state. This unification is motivated by task-oriented compression (e.g., MPEG Compact Descriptors for Video Analysis [8] and Video Coding for Machines [9]–[11]) where the coded representation has the dual responsibility of preserving the extrinsic aspect of the given data (which corresponds to the observable source realization) and capturing its intrinsic semantic feature (which is assumed to be carried by the hidden source state). Note that the two objectives of the decoder in semantic source coding are not necessarily aligned. Indeed, with the coding rate fixed, there often exists a tension between faithfully reproducing the extrinsic observation and accurately estimating the intrinsic state. Characterizing this tension in the form of a quantitative tradeoff is a fundamental problem from the information-theoretic perspective.

So far research on semantic source coding has been exclusively focused on centralized systems with a single encoder having access to all source components. However, in practice, there are many situations where the source components are not co-located and have to be processed in a distributed manner. Even when the source components are co-located, distributed processing might still be favored due to implementation constraints (e.g., small receptive fields of neural networks) or complexity considerations. This provides a strong incentive to study distributed semantic source coding and investigate how it differs from its centralized counterpart in terms of the performance limits.

In this work, we consider the quadratic Gaussian version of centralized semantic source coding and distributed semantic source coding. The Gaussian version is known to be analytically more tractable. Indeed, we are able to obtain several conclusive results regarding the fundamental rate-distortion limits, which are elusive in general. But more importantly, the Gaussian version is of special importance due to its extremal properties. As such, our results can be used widely as baselines for the non-Gaussian versions.

The rest of this paper is organized as follows. We introduce the problem definitions in Section II. The rate-distortion function of centralized Gaussian semantic source coding is explicitly characterized for the symmetric setting and the 2-component setting in Section III. The corresponding results for distributed Gaussian semantic source coding are presented in Section IV. We conclude the paper in Section V.

II. Problem Definitions

Let $\mathbf{X} := (X_1, \ldots, X_L)^T$ be an observable vector source and S be a state variable carrying certain semantic information. We assume $X_i = S + N_i$, $i = 1, \ldots, L$, where $S, N_1, \ldots, N_L$ are mutually independent zero-mean Gaussian random variables with variances $\sigma_S^2, \sigma_{N_1}^2, \ldots, \sigma_{N_L}^2$, respectively. So the covariance matrix of $\mathbf{X}$, denoted by $\mathbf{K}_{\mathbf{X}}$, can be written as

$$\mathbf{K}_{\mathbf{X}} = \begin{bmatrix} \sigma_S^2 + \sigma_{N_1}^2 & \sigma_S^2 & \cdots & \sigma_S^2 \\ \sigma_S^2 & \sigma_S^2 + \sigma_{N_2}^2 & \cdots & \sigma_S^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_S^2 & \sigma_S^2 & \cdots & \sigma_S^2 + \sigma_{N_L}^2 \end{bmatrix}.$$

It is easy to verify that

$$S = \mathbb{E}[S|\mathbf{X}] + Z = \mathbf{g}^T\mathbf{X} + Z,$$

where Z is a zero-mean Gaussian random variable, independent of $\mathbf{X}$, with variance $\sigma_Z^2 = (\frac{1}{\sigma_S^2} + \frac{1}{\sigma_{N_1}^2} + \ldots + \frac{1}{\sigma_{N_L}^2})^{-1}$, and $\mathbf{g} = (\frac{\sigma_Z^2}{\sigma_{N_1}^2}, \ldots, \frac{\sigma_Z^2}{\sigma_{N_L}^2})^T$. Let $\{(X_1(t), \ldots, X_L(t), S(t), Z(t))\}_{t=1}^{\infty}$ be a joint i.i.d. process induced by $(X_1, \ldots, X_L, S, Z)$.

Definition 1. *Rate R is said to be achievable with respect to reproduction distortion constraints $D_1, \ldots, D_L$ and semantic distortion constraint D_S via centralized coding if given any $\epsilon > 0$, there exist encoding function $f^{(n)} : \mathbb{R}^{L\times n} \to \mathcal{C}^{(n)}$ and* *decoding functions $g^{(n)}: \mathcal{C}^{(n)} \rightarrow \mathbb{R}^{L\times n}$ and $g_S^{(n)}: \mathcal{C}^{(n)} \rightarrow \mathbb{R}^n$ for all sufficiently large n such that*

$$\begin{aligned}
&\frac{1}{n}\log|\mathcal{C}^{(n)}| \leq R+\epsilon,\\
&\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[(X_i(t)-\hat{X}_i(t))^2] \leq D_i+\epsilon, \quad i=1,\ldots,L,\\
&\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[(S(t)-\hat{S}(t))^2] \leq D_S+\epsilon,
\end{aligned}$$

where $\hat{\mathbf{X}}^n := g^{(n)}(f^{(n)}(\mathbf{X}^n))$ (with $\mathbf{X}(t)$ and $\hat{\mathbf{X}}(t)$ standing for $(X_1(t),\ldots,X_L(t))^T$ and $(\hat{X}_1(t),\ldots,\hat{X}_L(t))^T$, respectively, $t=1,\ldots,n$) and $\hat{S}^n := g_S^{(n)}(f^{(n)}(\mathbf{X}^n))$. The infimum of such achievable R is denoted by $R_c(D_1,\ldots,D_L,D_S)$.

Definition 2. *Rate R is said to be achievable with respect to reproduction distortion constraints $D_1,\ldots,D_L$ and semantic distortion constraint D_S via distributed coding if given any $\epsilon>0$, there exist encoding function $f_i^{(n)}: \mathbb{R}^n \rightarrow \mathcal{C}_i^{(n)}$, $i=1,\ldots,L$, and decoding functions $g^{(n)}: \mathcal{C}_1^{(n)}\times\ldots\times\mathcal{C}_L^{(n)} \rightarrow \mathbb{R}^{L\times n}$ and $g_S^{(n)}: \mathcal{C}_1^{(n)}\times\ldots\times\mathcal{C}_L^{(n)} \rightarrow \mathbb{R}^n$ for all sufficiently large n such that*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{L}\log|\mathcal{C}_i^{(n)}| \leq R+\epsilon,\\
&\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[(X_i(t)-\hat{X}_i(t))^2] \leq D_i+\epsilon, \quad i=1,\ldots,L,\\
&\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[(S(t)-\hat{S}(t))^2] \leq D_S+\epsilon,
\end{aligned}$$

where $\hat{\mathbf{X}}^n := g^{(n)}(f_1^{(n)}(X_1^n),\ldots,f_L^{(n)}(X_L^n))$ (with $\hat{\mathbf{X}}(t)$ standing for $(\hat{X}_1(t),\ldots,\hat{X}_L(t))^T$, $t=1,\ldots,n$) and $\hat{S}^n := g_S^{(n)}(f_1^{(n)}(X_1^n),\ldots,f_L^{(n)}(X_L^n))$. The infimum of such achievable R is denoted by $R_d(D_1,\ldots,D_L,D_S)$.

Without loss of generality, we assume $D_i \in (0,\sigma_S^2+\sigma_{N_i}^2]$, $i=1,\ldots,L$, and $D_S \in (\sigma_Z^2,\sigma_S^2]$ throughout this paper.

The centralized Gaussian semantic source coding problem considered in the present work differs from that in [6] in two aspects. Firstly, we impose a reproduction distortion constraint on each source component while [6] adopts a trace distortion constraint. Secondly, we consider a special correlation structure where the observable source components are conditionally indepent given the hidden state; in constrast, [6] has no such a restriction. It is worth mentioning that the conditional independence assumption is introduced mainly to ensure that the results derived for the centralized Gaussian semantic source coding problem can be compared with those for the distributed counterpart as the latter problem is likely intractable without this assumption.

To the best of our knowledge, the distributed Gaussian semantic source coding problem formulated above is new. Nevertheless, it has rich connections with various network source coding problems [12]–[31] in the literature. In particular, it can be viewed as a coupling of the Gaussian multiterminal source coding problem [16], [21]–[25] and the Gaussian CEO problem [15], [17]–[20].

III. Centralized Gaussian Semantic Source Coding

The following result, which is a simple variant of [6, Theorem 2], provides a computable characterization of $R_c(D_1,\ldots,D_L,D_S)$.

Theorem 1. *We have*

$$\begin{aligned}
R_c(D_1,\ldots,D_L,D_S) = &\min_{\mathbf{\Delta}} \frac{1}{2}\log\frac{\det(\mathbf{K_X})}{\det(\mathbf{\Delta})} &(1)\\
\text{s.t.}\quad &\mathbf{0} \prec \mathbf{\Delta} \preceq \mathbf{K_X}, &(2)\\
&\mathrm{diag}(\mathbf{\Delta}) \preceq \mathbf{D}, &(3)\\
&\mathbf{g}^T\mathbf{\Delta}\mathbf{g}+\sigma_Z^2 \leq D_S, &(4)
\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix with the i-th diagonal entry being D_i, $i=1,\ldots,L$.

The optimization problem in (1) is a convex program and its solution can be verified using the Karush-Kuhn-Tucker conditions [32] stated in the following lemma.

Lemma 1. *$\mathbf{\Delta}^*$ is an optimal solution of the optimization problem in (1) if it satisfies the constraints (2)–(4) and there exist positive semidefinite matrix $\mathbf{U}$, positive semidefinite diagonal matrix $\mathbf{\Lambda}$, and nonnegative number ρ such that*

$$\begin{aligned}
&-(\mathbf{\Delta}^*)^{-1}+\mathbf{U}+\mathbf{\Lambda}+\rho\mathbf{g}\mathbf{g}^T=\mathbf{0},\\
&\mathbf{U}(\mathbf{\Delta}^*-\mathbf{K_X})=\mathbf{0},\\
&\mathbf{\Lambda}(\mathrm{diag}(\mathbf{\Delta}^*)-\mathbf{D})=\mathbf{0},\\
&\rho\left(\mathbf{g}^T\mathbf{\Delta}^*\mathbf{g}+\sigma_Z^2-D_S\right)=0.
\end{aligned}$$

Equipped with Theorem 1 and Lemma 1, we proceed to compute $R_c(D_1,\ldots,D_L,D_S)$ for some special cases.

We first consider the symmetric setting with $\sigma_{N_1}^2=\ldots=\sigma_{N_L}^2=\sigma_N^2$ and $D_1=\ldots=D_L=D$. An explicit characterization of $R_c(D_1,\ldots,D_L,D_S)$, abbreviated as $R_c(D,D_S)$, is provided by the following theorem.

Theorem 2. *The expression of $R_c(D,D_S)$ is given as follows:*

1) *If $D<\sigma_N^2$ and $D<\frac{(D_S-\sigma_Z^2)\sigma_N^4}{L\sigma_Z^4}$, then*

$$R_c(D,D_S)=\frac{1}{2}\log\frac{L\sigma_S^2\sigma_N^{2(L-1)}+\sigma_N^{2L}}{D^L}.$$

2) *If $D<\frac{L-1}{L}\sigma_N^2+\frac{(D_S-\sigma_Z^2)\sigma_N^4}{L^2\sigma_Z^4}$ and $D\geq\frac{(D_S-\sigma_Z^2)\sigma_N^4}{L\sigma_Z^4}$, then*

$$\begin{aligned}
&R_c(D,D_S)\\
&=\frac{1}{2}\log\frac{L\sigma_Z^4(L\sigma_S^2\sigma_N^{2(L-1)}+\sigma_N^{2L})}{(D_S-\sigma_Z^2)\sigma_N^4(\frac{L}{L-1}D-\frac{(D_S-\sigma_Z^2)\sigma_N^4}{L(L-1)\sigma_Z^4})^{L-1}}.
\end{aligned}$$

3) *If $D\geq\sigma_N^2$ and $D<\frac{L-1}{L}\sigma_N^2+\frac{(D_S-\sigma_Z^2)\sigma_N^4}{L^2\sigma_Z^4}$, then*

$$R_c(D,D_S)=\frac{1}{2}\log\frac{L\sigma_S^2+\sigma_N^2}{LD-(L-1)\sigma_N^2}.$$

4) *If* $D \geq \frac{L-1}{L}\sigma_N^2 + \frac{(D_S-\sigma_Z^2)\sigma_N^4}{L^2\sigma_Z^4}$, *then*

$$R_c(D, D_S) = \frac{1}{2}\log\frac{\sigma_S^2 - \sigma_Z^2}{D_S - \sigma_Z^2}.$$

The next result deals with the 2-component setting and provides an explicit characterization of $R_c(D_1, D_2, D_S)$. Let $\Delta_i := \sigma_S^2 + \sigma_{N_i}^2 - D_i$, $i = 1, 2$, and $\Delta_S := \sigma_S^2 - D_S$.

Theorem 3. *The expression of* $R_c(D_1, D_2, D_S)$ *is given as follows:*

1) *If* $D_2 \geq \sigma_S^2 + \sigma_{N_2}^2 - \frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}$ *and* $D_S \geq \sigma_S^2 - \frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}$, *then*

$$R_c(D_1, D_2, D_S) = \frac{1}{2}\log\frac{\sigma_S^2 + \sigma_{N_1}^2}{D_1}.$$

2) *If* $D_1 \geq \sigma_S^2 + \sigma_{N_1}^2 - \frac{\sigma_S^4\Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$ *and* $D_S \geq \sigma_S^2 - \frac{\sigma_S^4\Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$, *then*

$$R_c(D_1, D_2, D_S) = \frac{1}{2}\log\frac{\sigma_S^2 + \sigma_{N_2}^2}{D_2}.$$

3) *If* $D_i \geq \sigma_S^2 + \sigma_{N_i}^2 - \frac{\sigma_S^4\Delta_S}{(\sigma_S^2-\sigma_Z^2)^2}$, $i = 1, 2$, *then*

$$R_c(D_1, D_2, D_S) = \frac{1}{2}\log\frac{\sigma_S^2 - \sigma_Z^2}{D_S - \sigma_Z^2}.$$

4) *If* $\Delta_1\Delta_2 \geq \sigma_S^4$ *and* $D_S \geq \frac{\sigma_Z^4}{\sigma_{N_1}^2}D_1 + \frac{\sigma_Z^4}{\sigma_{N_2}^2}D_2 + \sigma_Z^2$, *then*

$$R_c(D_1, D_2, D_S) = \frac{1}{2}\log\frac{(\sigma_S^2 + \sigma_{N_1}^2)(\sigma_S^2 + \sigma_{N_2}^2) - \sigma_S^4}{D_1D_2}.$$

5) *If* $\Delta_1\Delta_S \geq \sigma_S^4$ *and* $D_2 \geq \frac{\sigma_{N_2}^4}{\sigma_{N_1}^4}D_1 + \frac{\sigma_{N_2}^4}{\sigma_Z^4}(D_S - \sigma_Z^2)$, *then*

$$R_c(D_1, D_2, D_S) = \frac{1}{2}\log\frac{(\sigma_S^2 + \sigma_{N_1}^2)(\sigma_S^2 - \sigma_Z^2) - \sigma_S^4}{D_1(D_S - \sigma_Z^2)}.$$

6) *If* $\Delta_2\Delta_S \geq \sigma_S^4$ *and* $D_1 \geq \frac{\sigma_{N_1}^4}{\sigma_{N_2}^4}D_2 + \frac{\sigma_{N_1}^4}{\sigma_Z^4}(D_S - \sigma_Z^2)$, *then*

$$R_c(D_1, D_2, D_S) = \frac{1}{2}\log\frac{(\sigma_S^2 + \sigma_{N_2}^2)(\sigma_S^2 - \sigma_Z^2) - \sigma_S^4}{D_2(D_S - \sigma_Z^2)}.$$

7) *If* $\Delta_1\Delta_2 < \sigma_S^4$, $D_1 < \sigma_S^2 + \sigma_{N_1}^2 - \frac{\sigma_S^4\Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$, $D_2 < \sigma_S^2+\sigma_{N_2}^2-\frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}$, *and* $D_S \geq \frac{\sigma_Z^4}{\sigma_{N_1}^4}D_1+\frac{2\sigma_Z^4}{\sigma_{N_1}^2\sigma_{N_2}^2}(\sigma_S^2-\sqrt{\Delta_1\Delta_2}) + \frac{\sigma_Z^4}{\sigma_{N_2}^4}D_2 + \sigma_Z^2$, *then*

$$R_c(D_1, D_2, D_S) = \frac{1}{2}\log\frac{(\sigma_S^2 + \sigma_{N_1}^2)(\sigma_S^2 + \sigma_{N_2}^2) - \sigma_S^4}{D_1D_2 - (\sigma_S^2 - \sqrt{\Delta_1\Delta_2})^2}.$$

8) *If* $\Delta_1\Delta_S < \sigma_S^4$, $D_1 < \sigma_S^2+\sigma_{N_1}^2-\frac{\sigma_S^4\Delta_S}{(\sigma_S^2-\sigma_Z^2)^2}$, $D_S < \sigma_S^2-\frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}$, *and* $D_2 \geq \frac{\sigma_{N_2}^4}{\sigma_{N_1}^4}D_1-\frac{2\sigma_{N_2}^4}{\sigma_{N_1}^2\sigma_Z^2}(\sigma_S^2-\sqrt{\Delta_1\Delta_S})+\frac{\sigma_{N_2}^4}{\sigma_Z^4}(D_S - \sigma_Z^2)$, *then*

$$\begin{aligned} &R_c(D_1, D_2, D_S)\\ &= \frac{1}{2}\log\frac{(\sigma_S^2 + \sigma_{N_1}^2)(\sigma_S^2 - \sigma_Z^2) - \sigma_S^4}{D_1(D_S - \sigma_Z^2) - (\sigma_S^2 - \sqrt{\Delta_1\Delta_S})^2}. \end{aligned}$$

9) *If* $\Delta_2\Delta_S < \sigma_S^4$, $D_2 < \sigma_S^2+\sigma_{N_2}^2-\frac{\sigma_S^4\Delta_S}{(\sigma_S^2-\sigma_Z^2)^2}$, $D_S < \sigma_S^2-\frac{\sigma_S^4\Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$, *and* $D_1 \geq \frac{\sigma_{N_1}^4}{\sigma_{N_2}^4}D_2-\frac{2\sigma_{N_1}^4}{\sigma_{N_2}^2\sigma_Z^2}(\sigma_S^2-\sqrt{\Delta_2\Delta_S})+\frac{\sigma_{N_1}^4}{\sigma_Z^4}(D_S - \sigma_Z^2)$, *then*

$$\begin{aligned} &R_c(D_1, D_2, D_S)\\ &= \frac{1}{2}\log\frac{(\sigma_S^2 + \sigma_{N_2}^2)(\sigma_S^2 - \sigma_Z^2) - \sigma_S^4}{D_2(D_S - \sigma_Z^2) - (\sigma_S^2 - \sqrt{\Delta_2\Delta_S})^2}. \end{aligned}$$

10) *Otherwise,*

$$\begin{aligned} &R_c(D_1, D_2, D_S)\\ &= \frac{1}{2}\log\frac{(\sigma_S^2 + \sigma_{N_1}^2)(\sigma_S^2 + \sigma_{N_2}^2) - \sigma_S^4}{D_1D_2 - \frac{\sigma_{N_1}^4\sigma_{N_2}^4}{4\sigma_Z^8}(D_S - \sigma_Z^2 - \frac{\sigma_Z^4}{\sigma_{N_1}^4}D_1 - \frac{\sigma_Z^4}{\sigma_{N_2}^4}D_2)^2}. \end{aligned}$$

Remark 1. *One can specialize [33, Theorem 6] and [34, Theorem III.1] from Theorem 3 by removing semantic distortion constraint D_S, i.e., by considering only Cases 1), 2), 4), and 7) where semantic distortion constraint D_S is inactive.*

IV. Distributed Gaussian Semantic Source Coding

Let $\Omega(\mathbf{K_X})$ denote the set of positive definite matrices $\mathbf{K_W}$ such that $\mathbf{K_X^{-1}} + \mathbf{K_W^{-1}}$ is a diagonal matrix. For any $\mathbf{K_W} \in \Omega(\mathbf{K_X})$,

$$\begin{aligned} &\underline{\psi}(D_1, \ldots, D_L, D_S, \mathbf{K_W})\\ &:= \min_{\mathbf{\Delta}, \gamma_1, \ldots, \gamma_L} \frac{1}{2}\log\frac{\det(\mathbf{K_X} + \mathbf{K_W})\det((\mathbf{K_X^{-1}} + \mathbf{K_W^{-1}})^{-1})}{\det(\mathbf{\Delta} + \mathbf{K_W})\det(\mathbf{\Gamma})} \end{aligned} \tag{5}$$

$$s.t. \quad \mathbf{0} \prec \mathbf{\Delta} \preceq \mathbf{K_X}, \tag{6}$$

$$\mathbf{0} \prec \mathbf{\Gamma} \preceq \left(\mathbf{\Delta}^{-1} + \mathbf{K_W^{-1}}\right)^{-1}, \tag{7}$$

$$\mathrm{diag}(\mathbf{\Delta}) \preceq \mathbf{D}, \tag{8}$$

$$\mathbf{g}^T\mathbf{\Delta}\mathbf{g} + \sigma_Z^2 \leq D_S, \tag{9}$$

where $\mathbf{\Gamma}$ is a diagonal matrix with the i-th diagonal entry being γ_i, $i = 1, \ldots, L$, and $\mathbf{D}$ is a diagonal matrix with the i-th diagonal entry being D_i, $i = 1, \ldots, L$. We define $\overline{\psi}(D_1, \ldots, D_L, D_S, \mathbf{K_W})$ in a similar way except that the constraint in (6) is replaced by

$$\mathbf{\Gamma} = (\mathbf{\Delta}^{-1} + \mathbf{K_W^{-1}})^{-1}. \tag{10}$$

Let

$$\underline{R}_d(D_1, \ldots, D_L, D_s) := \sup_{\mathbf{K_W} \in \Omega(\mathbf{K_X})} \underline{\psi}(D_1, \ldots, D_L, D_S, \mathbf{K_W}),$$

$$\overline{R}_d(D_1, \ldots, D_L, D_s) := \sup_{\mathbf{K_W} \in \Omega(\mathbf{K_X})} \overline{\psi}(D_1, \ldots, D_L, D_S, \mathbf{K_W}).$$

The following result, which is a simple variant of [22, Theorems 1 and 2], provides computable lower and upper bounds on $R_c(D_1,\ldots,D_L,D_S)$.

Theorem 4. *We have*

$$\begin{aligned}&\underline{R}_d(D_1,\ldots,D_L,D_S)\\&\le R_d(D_1,\ldots,D_L,D_S)\\&\le \overline{R}_d(D_1,\ldots,D_L,D_S).\end{aligned}$$

The optimization problem in (5) is a convex program and its solution can be verified using the Karush-Kuhn-Tucker conditions [32] stated in the following lemma, which also provides a matching condition for $\underline{R}_d(D_1,\ldots,D_L,D_S)$ and $\overline{R}_d(D_1,\ldots,D_L,D_S)$.

Lemma 2. *Given $\mathbf{K_W}\in\Omega(\mathbf{K_X})$, $(\mathbf{\Delta}^*,\gamma_1^*,\ldots,\gamma_L^*)$ is an optimal solution of the optimization problem in (5) if it satisfies the constraints (6)–(9) and there exist positive semidefinite matrices $\mathbf{U}$ and $\mathbf{V}$, positive semidefinite diagonal matrix $\mathbf{\Lambda}$, and nonnegative number γ such that*

$$\begin{aligned}&-(\mathbf{\Delta}^*+\mathbf{K_W})^{-1}+\mathbf{U}-(\mathbf{\Delta}^*)^{-1}((\mathbf{\Delta}^*)^{-1}+\mathbf{K_W^{-1}})^{-1}\mathbf{\Lambda}\\&((\mathbf{\Delta}^*)^{-1}+\mathbf{K_W^{-1}})^{-1}(\mathbf{\Delta}^*)^{-1}+\mathbf{\Lambda}^*+\rho\mathbf{g}\mathbf{g}^T=\mathbf{0}, &(11)\\&-(\mathbf{\Gamma}^*)^{-1}+\mathrm{diag}(\mathbf{V})=\mathbf{0}, &(12)\\&\mathbf{U}(\mathbf{\Delta}^*-\mathbf{K_X})=\mathbf{0}, &(13)\\&\mathbf{V}(\mathbf{\Gamma}^*-((\mathbf{\Delta}^*)^{-1}+\mathbf{K_W^{-1}})^{-1})=\mathbf{0}, &(14)\\&\mathbf{\Lambda}(\mathrm{diag}(\mathbf{\Delta}^*)-\mathbf{D})=\mathbf{0}, &(15)\\&\rho\left(\mathbf{g}^T\mathbf{\Delta}^*\mathbf{g}+\sigma_Z^2-D_S\right)=0, &(16)\end{aligned}$$

where $\mathbf{\Gamma}^*$ is a diagonal matrix with the i-th diagonal entry being γ_i^*, $i=1,\ldots,L$. Moreover, if this $(\mathbf{\Delta}^*,\gamma_1^*,\ldots,\gamma_L^*)$ further satisfies (10), then

$$\begin{aligned}&\overline{R}_d(D_1,\ldots,D_L,D_S)\\&=\overline{R}_d(D_1,\ldots,D_L,D_S)\\&=\frac{1}{2}\log\frac{\det(\mathbf{K_X})}{\det(\mathbf{\Delta}^*)}.\end{aligned}$$

Equipped with Theorem 4 and Lemma 2, we proceed to compute $R_d(D_1,\ldots,D_L,D_S)$ for some special cases.

We first consider the symmetric setting with $\sigma_{N_1}^2=\ldots=\sigma_{N_L}^2=\sigma_N^2$ and $D_1=\ldots=D_L=D$. An explicit characterization of $R_d(D_1,\ldots,D_L,D_S)$, abbreviated as $R_d(D,D_S)$, is provided by the following result. Let

$$\alpha:=\frac{\sigma_S^2}{\sigma_S^2+\sigma_N^2+\beta\sigma_N^2(L\sigma_S^2+\sigma_N^2)},$$

where β is the unique nonnegative solution to

$$\frac{\sigma_N^2}{1+\beta\sigma_N^2}+\frac{\sigma_S^2}{(1+\beta\sigma_N^2)(1+\beta(L\sigma_S^2+\sigma_N^2))}=D.$$

Theorem 5. *The expression of $R_d(D,D_S)$ is given as follows:*

1) *If $D_S\ge\frac{L\sigma_Z^4}{\sigma_N^4}(D+(L-1)\alpha D)+\sigma_Z^2$, then*

$$R_d(D,D_S)=\frac{1}{2}\log\frac{L\sigma_S^2\sigma_N^{2(L-1)}+\sigma_N^{2L}}{L\alpha D(D-\alpha D)^{L-1}+(D-\alpha D)^L}.$$

2) *If $D_S<\frac{L\sigma_Z^4}{\sigma_N^4}(D+(L-1)\alpha D)+\sigma_Z^2$, then*

$$R_d(D,D_S)=\frac{1}{2}\log\frac{L^L\sigma_S^2\sigma_Z^{2L}D_S^{L-1}}{\sigma_N^{2L}(D_S-\sigma_Z^2)^L}.$$

The next result deals with the 2-component setting and provides an explicit characterization of $R_d(D_1,D_2,D_S)$. Let $\Theta_i:=\sqrt{\sigma_{N_i}^4+4D_iD_S}$, $i=1,2$, and $\Theta_Z:=\sqrt{\frac{\sigma_{N_1}^4\sigma_{N_2}^4}{\sigma_Z^4}+4D_1D_2}$.

Theorem 6. *Without loss of generality, assume $\sigma_{N_1}^2\le\sigma_{N_2}^2$. The expression of $R_d(D_1,D_2,D_S)$ is given as follows:*

1) *If $D_2\ge\sigma_S^2+\sigma_{N_2}^2-\frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}$ and $D_S\ge\sigma_S^2-\frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}$, then*

$$R_d(D_1,D_2,D_S)=\frac{1}{2}\log\frac{\sigma_S^2+\sigma_{N_1}^2}{D_1}.$$

2) *If $D_1\ge\sigma_S^2+\sigma_{N_1}^2-\frac{\sigma_S^4\Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$ and $D_S\ge\sigma_S^2-\frac{\sigma_S^4\Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$, then*

$$R_d(D_1,D_2,D_S)=\frac{1}{2}\log\frac{\sigma_S^2+\sigma_{N_2}^2}{D_2}.$$

3) *If $D_S\ge(\frac{1}{\sigma_S^2}+\frac{1}{\sigma_{N_1}^2}-\frac{1}{\sigma_{N_2}^2})^{-1}$, $D_1\ge\frac{(\sigma_S^2+\sigma_{N_1}^2)((\sigma_S^2+\sigma_{N_1}^2)D_S-\sigma_S^2\sigma_{N_1}^2)}{\sigma_S^4}$, and $D_2\ge D_S+\sigma_{N_2}^2$, then*

$$R_d(D_1,D_2,D_S)=\frac{1}{2}\log\frac{\sigma_S^4}{D_S\sigma_S^2-\sigma_S^2\sigma_{N_1}^2+D_S\sigma_{N_1}^2}.$$

4) *If $D_S<(\frac{1}{\sigma_S^2}+\frac{1}{\sigma_{N_1}^2}-\frac{1}{\sigma_{N_2}^2})^{-1}$, $D_1\ge\frac{(D_S^2-\sigma_Z^4)\sigma_{N_1}^4}{4D_S\sigma_Z^4}$, and $D_2\ge\frac{(D_S^2-\sigma_Z^4)\sigma_{N_2}^4}{4D_S\sigma_Z^4}$, then*

$$R_d(D_1,D_2,D_S)=\frac{1}{2}\log\frac{4\sigma_S^2\sigma_Z^4D_S}{\sigma_{N_1}^2\sigma_{N_2}^2(D_S-\sigma_Z^2)^2}.$$

5) *If $D_1<\sigma_S^2+\sigma_{N_1}^2-\frac{\sigma_S^4\Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$, $D_2<\sigma_S^2+\sigma_{N_2}^2-\frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}$, and $D_S\ge\frac{\sigma_Z^4}{\sigma_{N_1}^4}D_1+\frac{\sigma_Z^4}{\sigma_{N_1}^2\sigma_{N_2}^2}\Theta_Z+\frac{\sigma_Z^4}{\sigma_{N_2}^4}D_2$, then*

$$R_d(D_1,D_2,D_S)=\frac{1}{2}\log\frac{2\sigma_S^2\sigma_Z^2}{\sigma_Z^2\Theta_Z-\sigma_{N_1}^2\sigma_{N_2}^2}.$$

6) *If $D_1<\frac{(D_S^2-\sigma_Z^4)\sigma_{N_1}^4}{4D_S\sigma_Z^4}$, $D_S<\min\{(\frac{1}{\sigma_S^2}+\frac{1}{\sigma_{N_1}^2}-\frac{1}{\sigma_{N_2}^2})^{-1},\sigma_S^2-\frac{\sigma_S^4\Delta_1}{(\sigma_S^2+\sigma_{N_1}^2)^2}\}$, and $D_2\ge\frac{\sigma_{N_2}^4}{\sigma_{N_1}^4}D_1-\frac{\sigma_{N_2}^4}{\sigma_Z^2\sigma_{N_1}^2}\Theta_1+\frac{\sigma_{N_2}^4}{\sigma_Z^4}D_S$, then*

$$\begin{aligned}&R_d(D_1,D_2,D_S)\\&=\frac{1}{2}\log\frac{2\sigma_S^2\sigma_Z^2\sigma_{N_1}^2}{\sigma_{N_1}^2\sigma_{N_2}^2\Theta_1-2\sigma_Z^2\sigma_{N_2}^2D_1-\sigma_{N_1}^4\sigma_{N_2}^2}.\end{aligned}$$

7) *If i)* $D_2 < D_S + \sigma_{N_2}^2$, $(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_{N_1}^2} - \frac{1}{\sigma_{N_2}^2})^{-1} \leq D_S < \sigma_S^2 - \frac{\sigma_S^4 \Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}$, *and* $D_1 \geq \frac{\sigma_{N_1}^4}{\sigma_{N_2}^4} D_2 - \frac{\sigma_{N_1}^4}{\sigma_Z^2 \sigma_{N_2}^2} \Theta_2 + \frac{\sigma_{N_1}^4}{\sigma_Z^4} D_S$, *or ii)* $D_2 < \frac{(D_S^2 - \sigma_Z^4)\sigma_{N_2}^4}{4 D_S \sigma_Z^4}$, $D_S < \min\{(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_{N_1}^2} - \frac{1}{\sigma_{N_2}^2})^{-1}, \sigma_S^2 - \frac{\sigma_S^4 \Delta_2}{(\sigma_S^2+\sigma_{N_2}^2)^2}\}$, *and* $D_1 \geq \frac{\sigma_{N_1}^4}{\sigma_{N_2}^4} D_2 - \frac{\sigma_{N_1}^4}{\sigma_Z^2 \sigma_{N_2}^2} \Theta_2 + \frac{\sigma_{N_1}^4}{\sigma_Z^4} D_S$,

$$\begin{aligned} & R_d(D_1, D_2, D_S) \\ & = \frac{1}{2} \log \frac{2\sigma_S^2 \sigma_Z^2 \sigma_{N_2}^2}{\sigma_{N_1}^2 \sigma_{N_2}^2 \Theta_2 - 2\sigma_Z^2 \sigma_{N_1}^2 D_2 - \sigma_{N_1}^2 \sigma_{N_2}^4}. \end{aligned}$$

V. Conclusion

We have studied centralized Gaussian semantic source coding and its distributed counterpart in terms of their rate-distortion functions. There are several directions worthy of pursuing for future work. For example, it is of great interest to investigate more general correlation structures between the observable variables and the state variable. The i.i.d. assumption adopted in our work also appears to be overly restrictive. This can be remedied by considering the one-shot formulation, which is better justified from a practical perspective. One may further go beyond the quadratic Gaussian setting to deal with more realistic source models and loss functions. Here the notorious technical difficulties inherent in distributed source coding will likely become a roadblock. Nevertheless, it remains promising to make good progress within the log-loss framework [35] that is most relevant to machine learning applications.

References

[1] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[2] R. Dobrushin and B. Tsybakov, “Information transmission with additional noise," *IRE Trans. Inf. Theory*, vol. 8, no. 5, pp. 293–304, Sep. 1962.

[3] J. Wolf and J. Ziv, “Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. 16, no. 4, pp. 406–411, Jul. 1970.

[4] T. Berger, Rate Distortion Theory. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[5] H. Witsenhausen, “Indirect rate distortion problems," *IEEE Trans. Inf. Theory*, vol. 26, no. 5, pp. 518–521, Sep. 1980.

[6] J. Liu, S. Shao, W. Zhang, and H. V. Poor, “An indirect rate-distortion characterization for semantic sources: General model and the case of Gaussian observation," 2022, arXiv:2201.12477. [Online]. Available: https://arxiv.org/abs/2201.12477

[7] T. Guo, Y. Wang, J. Han, H. Wu, B. Bai, W. Han, “Semantic compression with side information: A rate-distortion perspective," 2022, arXiv:2208.06094. [Online]. Available: https://arxiv.org/abs/2208.06094

[8] L.-Y. Duan, V. Chandrasekhar, S. Wang, Y. Lou, J. Lin, Y. Bai, T. Huang, A. C. Kot, and W. Gao, “Compact descriptors for video analysis: The emerging MPEG standard," *IEEE MultiMedia*, vol. 26, no. 2, pp. 44–54, 1 Apr.-Jun. 2019.

[9] S. Ma, X. Zhang, S. Wang, X. Zhang, C. Jia, and S. Wang, “Joint feature and texture coding: Toward smart video representation via front-end intelligence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3095–3105, Oct. 2019.

[10] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, “Video coding for machines: A paradigm of collaborative compression and intelligent analytics.," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.

[11] S. Yang, Y. Hu, W. Yang, L. Duan, and J. Liu, “Towards coding for human and machine vision: Scalable face image coding,? *IEEE Trans. Multimedia*, vol. 23, pp. 2957–2971, 2021.

[12] T. Berger, “Multiterminal source coding," in *The Information Theory Approach to Communications* (CISM International Centre for Mechanical Sciences), vol. 229, G. Longo, Ed. New York, NY, USA: Springer-Verlag, 1978, pp. 171–231.

[13] S.-Y. Tung, “Multiterminal source coding," Ph.D. dissertation, School Electr. Eng., Cornell Univ., Ithaca, NY, USA, 1978.

[14] T. Berger, Z. Zhang, and H. Viswanathan, “The CEO problem," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996.

[15] H. Viswanathan and T. Berger, “The quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vo. 43, no. 5, pp. 1549–1559, Sep. 1997.

[16] Y. Oohama, “Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1912–1923, Nov. 1997.

[17] Y. Oohama, “The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.

[18] J. Chen, X. Zhang, T. Berger, and S. B. Wicker, “An upper bound on the sum-rate distortion function and its corresponding rate allocation schemes for the CEO problem," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 977–987, Aug. 2004.

[19] V. Prabhakaran, D. Tse, and K. Ramchandran, “Rate region of the quadratic Gaussian CEO problem," in *Proc. IEEE Int. Symp. Inf. Theory*, 2004, p. 117.

[20] Y. Oohama, “Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, pp. 2577–2593, Jul. 2005.

[21] A. B. Wagner, S. Tavildar, and P. Viswanath, “Rate region of the quadratic Gaussian two-encoder source-coding problem," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1938–1961, May 2008.

[22] J. Wang, J. Chen, and X. Wu, “On the sum rate of Gaussian multiterminal source coding: New proofs and results," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3946–3960, Aug. 2010.

[23] Y. Yang, Y. Zhang, and Z. Xiong, “A new sufficient condition for sum-rate tightness in quadratic Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 408–423, Jan. 2013.

[24] J. Wang and J. Chen, “Vector Gaussian two-terminal source coding," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3693–3708, Jun. 2013.

[25] J. Wang and J. Chen, “Vector Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5533–5552, Sep. 2014.

[26] Y. Oohama, “Indirect and direct Gaussian distributed source coding problems," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7506–7539, Dec. 2014.

[27] J. Chen, F. Etezadi, and A. Khisti, “Generalized Gaussian multiterminal source coding and probabilistic graphical models," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Aachen, Germany, Jun. 25 - 30, 2017, pp. 719–723.

[28] Y. Wang, L. Xie, S. Zhou, M. Wang, and J. Chen, “Asymptotic rate-distortion analysis of symmetric remote Gaussian source coding: Centralized encoding vs. eistributed encoding," *Entropy*, vol. 21(2), 213, pp. 1–14, Feb. 2019.

[29] Y. Wang, L. Xie, X. Zhang, and J. Chen, “Robust distributed compression of symmetrically correlated Gaussian sources," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2343–2354, Mar. 2019.

[30] J. Chen, L. Xie, Y. Chang, J. Wang, and Y. Wang, “Generalized Gaussian multiterminal source coding: The symmetric case," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2115–2128, Apr. 2020.

[31] L. Xie, X. Tu, S. Zhou, and J. Chen, “Generalized Gaussian multiterminal source coding in the high-resolution regime," IEEE Transactions on Communications,*IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3782–3791, Jun. 2020.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[33] J. Xiao and Z. Luo, “Compression of correlated Gaussian sources under individual distortion criteria," in *Proc. 43rd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sep. 28 - 30, 2005, pp. 438–447.

[34] A. Lapidoth and S. Tinguely, “Sending a bivariate Gaussian over a Gaussian MAC," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2714–2752, Jun. 2010.

[35] T. A. Courtade and T. Weissman, “Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.

Functional Representation Lemma: Algorithms and Applications

Yanina Shkel
École Polytechnique Fédérale de Lausanne (EPFL)
email: yanina.shkel@epfl.ch

***Abstract*—Functional Representation Lemma (FRL) is an information-theoretic technique that fixes a correlated 'reference' information source, and extracts a 'residual' information about the original source. Recently, there has been a lot of interest in FRL since variants of this technique appear across different problems in information theory, and data science more broadly.**

In this tutorial talk we overview the FRL problem. We highlight some of its applications: these include the problems of privacy and causal inference, as well as proof techniques for single-shot information-theoretic bounds. Finally, we review known algorithms for constructing functional representations. We particularly focus on the greedy algorithms previously proposed in literature.

I. Extended Abstract

A. Overview and Applications

We begin with the Simple Functional Representation Lemma (FRL) which can be found in [1] and was independently derived in [2]–[4], among others. Given two jointly distributed discrete random variable (X, Y), the lemma states that there exists a random variable Z such that

$$I(Y; Z) = 0, \tag{1}$$

$$H(X|Z, Y) = 0, \tag{2}$$

$$\text{and } |\mathcal{Z}| \le |\mathcal{Y}|(|\mathcal{X}| - 1) + 1. \tag{3}$$

That is, Y and Z are independent, X is a deterministic function of Y and Z, and the support of Z is bounded. This result could be shown with a construction that we here call the *Simple FRL algorithm*. See, for example, [1, Appendix B] and [4, Lemma 1] for a detailed exposition.

The simple FRL shows that a random variable Z that satisfies (1) and (2) exists. However, there are many more interesting questions that arise about properties of this random variable. One line of work focuses on minimizing $H(X|Z)$ (or maximizing $I(X; Z)$). The best known result, known as the Strong Functional Representation Lemma (SFRL), states that it is possible to construct Z, such that

$$H(X|Z) \le I(X; Y) + \log(I(X; Y) + 1) + O(1), \tag{4}$$

where $I(X; Y)$ is a trivial lower bound on $H(X|Z)$ [5], [6]. An extensions of SFRL, known as the Poisson Matching Lemma, has been proposed in [7]. These results find extensive applications in derivations of single-shot coding bounds [5]–[8], as well as for problems in information-theoretic privacy [4], [9].

Another line of work on minimizing the entropy $H(Z)$ finds applications in the problem of private compression [4], [15], causal inference [16]–[20], as well as a number of other problems in statistics [21]. Let $Q = \bigwedge_{y\in\mathcal{Y}} P_{X|Y=y}$ be the lower bound with respect to majorization [21] of the set of distributions $\{P_{X|Y=y}\}_{y\in\mathcal{Y}}$. It can be shown that

$$H(Q) \le H(Z) \le H(Q) + 2 - 2^{2-|\mathcal{Y}|}. \tag{5}$$

The lower bound in (5) was shown in [21]. It was also shown in [21] that the upper bound for $|\mathcal{Y}| = 2$ holds via a greedy algorithms that we refer to as the *majorization-based algoirhtm*. The general upperbound in (5) was shown in [22] using the technique of *geometric splitting*.

An improvement on (5) has been recently shown using the information spectrum of Z. Specifically [23], [24] show that

$$\mathbb{P}[\imath_Z(Z) > t] \ge \sup_{y\in\mathcal{Y}} \mathbb{P}[\imath_{X|Y}(X|Y) > t | Y = y] \tag{6}$$

where $\imath_Z(z) = \log \frac{1}{P_Z(z)}$ and $\imath_{X|Y}(x|y) = \log \frac{1}{P_{X|Y}(x|y)}$. Moreover, [23], [24] show that there exists a distribution Q^* such that

$$H(Q) \le H(Q^*) \le H(Z). \tag{7}$$

This Q^* could be found with a simple greedy algorithm from the information spectrum envelope on the right-hand-side of (6). Finally, [23], [25] show that an algorithm that we call the *natural greedy algorithm* is within $\frac{\log_2(e)}{e} \approx 0.53$ bits of the minimal achievable entropy, while, in general, the problem is known to be NP-hard.

B. On Greedy Algorithms

In this talk, we particularly focus on greedy algorithm for the problem of constructing Z. This includes the the majorization-based algorithm which attempts to best approximate the greatest lowebound Q in (5). The natural greedy algorithm, on the other hand, puts as much probability mass as possible into the likelier realizations of Z. Greedy algorithms do not just play a role with constructing the random variable Z. The also show up in the evaluations of lower bounds in (5) and (7).

References

[1] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[2] B. Hajek and M. Pursley, "Evaluation of an achievable rate region for the broadcast channel," *IEEE Transactions on Information Theory*, vol. 25, no. 1, pp. 36–46, January 1979.

[3] F. Willems and E. van der Meulen, "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Transactions on Information Theory*, vol. 31, no. 3, pp. 313–327, 1985.

[4] Y. Y. Shkel, R. S. Blum, and H. V. Poor, "Secrecy by design with applications to privacy and compression," *IEEE Transactions on Information Theory*, vol. 67, no. 2, pp. 824–843, 2021.

[5] C. T. Li and A. E. Gamal, "Strong functional representation lemma and applications to coding theorems," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 589–593.

[6] P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," *IEEE Trans. Inf. Theor.*, vol. 56, no. 1, pp. 438–449, jan 2010. [Online]. Available: https://doi.org/10.1109/TIT.2009.2034824

[7] C. T. Li and V. Anantharam, "A unified framework for one-shot achievability via the poisson matching lemma," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2624–2651, 2021.

[8] L. Theis and A. B. Wagner, "A coding theorem for the rate-distortion-perception function," 2021.

[9] A. Zamani, T. J. Oechtering, and M. Skoglund, "On the privacy-utility trade-off with and without direct access to the private data," *IEEE Transactions on Information Theory*, pp. 1–1, 2023.

[10] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Information Theory Workshop (ITW 2014)*, 2014, pp. 501–505.

[11] F. d. P. Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5011–5038, Aug 2017.

[12] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, 2016. [Online]. Available: https://www.mdpi.com/2078-2489/7/1/15

[13] B. Rassouli and D. Gündüz, "Information-theoretic privacy-preserving schemes based on perfect privacy," 2023. [Online]. Available: https://arxiv.org/abs/2301.11754

[14] A. Zamani, T. J. Oechtering, and M. Skoglund, "On the privacy-utility trade-off with and without direct access to the private data," 2022. [Online]. Available: https://arxiv.org/abs/2212.12475

[15] Y. Y. Shkel and H. V. Poor, "A compression perspective on secrecy measures," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 163–176, 2021.

[16] M. Kocaoglu, A. Dimakis, S. Vishwanath, and B. Hassibi, "Entropic causal inference," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10674

[17] S. Compton, K. Greenewald, D. A. Katz, and M. Kocaoglu, "Entropic causal inference: Graph identifiability," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 4311–4343. [Online]. Available: https://proceedings.mlr.press/v162/compton22a.html

[18] M. Kocaoglu, A. G. Dimakis, S. Vishwanath, and B. Hassibi, "Entropic causality and greedy minimum entropy coupling," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 1465–1469.

[19] A. Painsky, S. Rosset, and M. Feder, "Innovation representation of stochastic processes with application to causal inference," *IEEE Transactions on Information Theory*, vol. 66, no. 2, pp. 1136–1154, 2020.

[20] ——, "Memoryless representation of markov processes," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2294–298.

[21] F. Cicalese, L. Gargano, and U. Vaccaro, "Minimum-entropy couplings and their applications," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3436–3451, 2019.

[22] C. T. Li, "Efficient approximate minimum entropy coupling of multiple probability distributions," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5259–5268, 2021.

[23] S. Compton, "A tighter approximation guarantee for greedy minimum entropy coupling," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 168–173.

[24] Y. Y. Shkel and A. Kumar Yadav, "Information spectrum converse for minimum entropy couplings and functional representations," in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 66–71.

[25] S. Compton, D. Katz, B. Qi, K. Greenewald, and M. Kocaoglu, "Minimum-entropy coupling approximation guarantees beyond the majorization barrier," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, F. Ruiz, J. Dy, and J.-W. van de Meent, Eds., vol. 206. PMLR, 25–27 Apr 2023, pp. 10 445–10 469. [Online]. Available: https://proceedings.mlr.press/v206/compton23a.html

Pointwise Redundancy in One-Shot Lossy Compression via Poisson Functional Representation

Cheuk Ting Li
The Chinese University of Hong Kong
Hong Kong SAR, China
email: ctli@ie.cuhk.edu.hk

***Abstract*—We present a construction of one-shot variable-length lossy source coding schemes using the Poisson functional representation, and give bounds on its pointwise redundancy. This allows us to describe the distribution of the encoding length in a precise manner.**

I. Introduction

Variable-length lossy source coding has been considered, for example, in D-semifaithful codes [1], [2] where the distortion must be bounded almost surely. The redundancy of D-semifaithful codes, i.e., the difference between the encoding length and the rate distortion function, has been studied in [3]–[6].

For one-shot variable-length lossy source coding with the expected distortion constraint $\mathbb{E}[d(X,Y)] \le D$,[1] it was proved in [7] that there is a prefix-free code with expected length $\le R(D) + \log(R(D)+1) + 6$, showing that the optimal one-shot expected length is always within a logarithmic gap from the rate-distortion function $R(D)$. The proof utilizes the Poisson functional representation [7], [8], where the codebook is constructed as a Poisson process. Also see [9]–[11] for related results.

In this work, we utilize the Poisson functional representation to construct one-shot variable-length lossy source coding schemes, and give bounds on their pointwise redundancy. This allows us to describe the distribution of the encoding length in a more precise manner, compared to only bounding its expectation. The proofs and details of the results mentioned in this abstract, and the generalization to the lossy Gray-Wyner system [12], can be found in the preprint [13].

II. Main Results

A one-shot variable-length lossy compression scheme for the source $X \in \mathcal{X}$, $X \sim P_X$ with reconstruction space $\mathcal{Y}$ is a pair $(P_{M|X}, g)$, where $P_{M|X}$ is a stochastic encoder (a conditional distribution from $\mathcal{X}$ to $\{0,1\}^*$, where $\{0,1\}^*$ is the set of bit sequences of any length), and $g : \{0,1\}^* \to \mathcal{Y}$ is a decoding function. The encoder observes $X \sim P_X$ and outputs the description $M|X \sim P_{M|X}$. The decoder observes M and outputs the reconstruction $\tilde{Y} = g(M)$. We can choose whether to impose the prefix-free condition on M or not. We may impose an expected distortion constraint $\mathbb{E}[d(X,\tilde{Y})] \le D$, where $d : \mathcal{X} \times \mathcal{Y} \to [0,\infty)$ is a distortion function.

We can also replace the variable-length description M by a positive integer K, and assume that the encoder produces a positive integer description. Note that we can convert K into a variable-length description with $\lfloor \log K \rfloor$ bits without the prefix-free condition [14], or $\le \log K + 2\log(\log K + 1) + 1$ bits with the prefix-free condition using the Elias delta code [15].

The following theorem can be proved using the Poisson functional representation construction similar to [7, Theorem 2], with an analysis using techniques in [8]. Refer to [13] for the proof.

Theorem 1: Fix any P_X, $P_{Y|X}$ and Q_Y satisfying $P_{Y|X}(\cdot|x) \ll Q_Y$ for P_X-almost all x's. Fix any collection of functions $\psi_i : \mathcal{X} \times \mathcal{Y} \times \mathbb{Z}_{>0} \to \mathbb{R}$ that are nondecreasing in the third argument for $i = 1,\ldots,\ell$. Then there exists a lossy compression scheme with description $K \in \mathbb{Z}_{>0}$ and reconstruction $\tilde{Y}$ such that

$$\mathbb{E}\big[\psi_i(X,\tilde{Y},K)\big] \le \mathbb{E}\big[\psi_i(X,Y,\ell J)\big]$$

for $i = 1,\ldots,\ell$, where $(X,Y) \sim P_X P_{Y|X}$, and $J \in \mathbb{Z}_{>0}$ is distributed as

$$J|(X,Y) \sim \mathrm{Geom}\bigg(\Big(\frac{\mathrm{d}P_{Y|X}(\cdot|X)}{\mathrm{d}Q_Y}(Y) + 1\Big)^{-1}\bigg).$$

This theorem is quite general. For example, to bound the expected distortion, take $\psi_i(x,y,k) = d(x,y)$. To bound the excess distortion probability, take $\psi_i(x,y,k) = \mathbf{1}\{d(x,y) > D\}$. To bound the probability that K cannot be encoded into n bits (for a fixed-length code), take $\psi_i(x,y,k) = \mathbf{1}\{k > 2^n\}$. To bound the expected length with (resp. without) the prefix-free condition, we may take $\psi_i(x,y,k) = \log k$ (resp. $\psi_i(x,y,k) = \log k + 2\log(\log k + 1) + 1$).

We can also use Theorem 1 to bound the pointwise redundancy. We consider three different notions of pointwise redundancy: **Pointwise rate redundancy (PRR)**, studied in [5], [16], is given by

$$|M| - R(D),$$

i.e., the difference between the length $|M|$ of the description M and the rate-distortion function $R(D)$ where $D = \mathbb{E}[d(X,\tilde{Y})]$.

This work was partially supported by an ECS grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No.: CUHK 24205621].

[1] Note that the probability of excess distortion $\mathbb{P}(d(X,Y) > D) = \mathbb{E}[\mathbf{1}\{d(X,Y) > D\}]$ can also be written as an expected distortion.

Pointwise source-wise redundancy (PSR), studied in [5], is given by

$$|M| - \jmath(X, D),$$

where $\jmath(x, D)$ is the *d-tilted information* [5], [17], [18] $\jmath(x, D) := -\log \mathbb{E}[2^{-\lambda^*(d(x,Y^*)-D)}]$, where $Y^* \sim P_Y$ follows the Y-marginal of $P_X P_{Y|X}$ where $P_{Y|X}$ is the conditional distribution that attains the minimum in $R(D)$ (assume unique minimizer), and $\lambda^* := -R'(D)$. **Pointwise source-distortion-wise redundancy (PSDR)** is defined as

$$|M| - \jmath(X, D, d(X, \tilde{Y})),$$

where we write $\jmath(x, D, \delta) := -\log \mathbb{E}[2^{-\lambda^*(d(x,Y^*)-\delta)}] = \jmath(x, D) - \lambda^*(\delta - D)$, which can be interpreted as the amount of information needed to convey x within a distortion δ when the overall expected distortion is D. The expectations of these three redundancies must be nonnegative for prefix-free codes, but might be negative if we do not impose the prefix-free condition. We first state a corollary of Theorem 1 that can bound any of the three pointwise redundancies for the case without the prefix-free condition.

Corollary 2: Fix any P_X, $P_{Y|X}$, distortion function $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$, function $\eta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $\gamma \in \mathbb{R}$. Then there exists a lossy compression scheme without prefix-free condition such that $\mathbb{E}[d(X, \tilde{Y})] \le \mathbb{E}[d(X, Y)]$, and

$$\begin{aligned}&\mathbb{P}\left(|M| - \eta(X, \tilde{Y}) \ge \gamma\right)\\ &\le \mathbb{E}\left[\min\left\{2^{-\eta(X,Y)-\gamma+1}(2^{\iota_{X;Y}(X;Y)} + 1),\, 1\right\}\right],\end{aligned}$$

where $(X, Y) \sim P_X P_{Y|X}$.

The result for PSDR is especially simple.

Corollary 3: For $D > 0$, under the regularity conditions in [18],[2] there exists a lossy compression scheme without prefix-free condition, with $\mathbb{E}[d(X, \tilde{Y})] \le D$, and with PSDR satisfying

$$\mathbb{P}\left(|M| - \jmath(X, D, d(X, \tilde{Y})) \ge \gamma\right) \le 2^{-\gamma+2}$$

for every $\gamma \in \mathbb{R}$.

The results for prefix-free codes are slightly more complicated.

Corollary 4: Fix any P_X, $P_{Y|X}$, distortion function $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$, function $\eta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and $\gamma \in \mathbb{R}$. Then there exists a prefix-free lossy compression scheme such that $\mathbb{E}[d(X, \tilde{Y})] \le \mathbb{E}[d(X, Y)]$, and

$$\begin{aligned}&\mathbb{P}\left(|M| - \eta(X, \tilde{Y}) \ge \gamma\right)\\ &\le \mathbb{E}\Big[\min\left\{2^{-\eta(X,Y)-\gamma+2}([\eta(X, Y) + \gamma]_+ + 1)^2\right.\\ &\qquad \left.\cdot\, (2^{\iota_{X;Y}(X;Y)} + 1),\, 1\right\}\Big],\end{aligned}$$

where $(X, Y) \sim P_X P_{Y|X}$.

Corollary 5: For $D > 0$, $\gamma \in \mathbb{R}$, under the regularity conditions in [18] (see Corollary 3), there exists a prefix-free lossy compression scheme with $\mathbb{E}[d(X, \tilde{Y})] \le D$, and with PSDR satisfying

$$\begin{aligned}&\mathbb{P}\left(|M| - \jmath(X, D, d(X, \tilde{Y})) \ge \gamma\right)\\ &\le 2^{-\gamma+3}\mathbb{E}\big[([\iota_{X;Y}(X;Y) + \gamma]_+ + 1)^2\big].\end{aligned}$$

[2] The regularity conditions in [18] are: $R(\delta)$ is finite for some δ, there exists a finite set $\mathcal{E} \subseteq \mathcal{Y}$ such that $\mathbb{E}[\min_{y\in\mathcal{E}} d(X, y)] < \infty$, and the minimum in $R(D)$ is achieved by a unique $P_{Y|X}$.

References

[1] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *The Annals of Probability*, pp. 441–452, 1990.

[2] B. Yu and T. P. Speed, "A rate of convergence result for a universal d-semifaithful code," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 813–820, 1993.

[3] J. C. Kieffer, "Sample converses in source coding theory," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 263–268, 1991.

[4] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion. 1. known statistics," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 71–91, 1997.

[5] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 136–152, 2000.

[6] A. Dembo and I. Kontoyiannis, "Critical behavior in lossy source coding," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1230–1236, 2001.

[7] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6967–6978, Nov 2018.

[8] C. T. Li and V. Anantharam, "A unified framework for one-shot achievability via the Poisson matching lemma," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2624–2651, 2021.

[9] E. C. Posner and E. R. Rodemich, "Epsilon entropy and data compression," *The Annals of Mathematical Statistics*, pp. 2079–2125, 1971.

[10] P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 438–449, Jan 2010.

[11] M. Braverman and A. Garg, "Public vs private coin in bounded-round information," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2014, pp. 502–513.

[12] R. Gray and A. Wyner, "Source coding for a simple network," *Bell System Technical Journal*, vol. 53, no. 9, pp. 1681–1721, 1974.

[13] C. T. Li, "Pointwise redundancy in one-shot lossy compression via Poisson functional representation," *arXiv preprint*, 2024.

[14] W. Szpankowski and S. Verdú, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4017–4025, 2011.

[15] P. Elias, "Universal codeword sets and representations of the integers," *IEEE transactions on information theory*, vol. 21, no. 2, pp. 194–203, 1975.

[16] B. Oğuz and V. Anantharam, "Pointwise lossy source coding theorem for sources with memory," in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 363–367.

[17] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1, pp. 57–71, 1974.

[18] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3309–3338, 2012.

Exponential Strong Converse in Multi-user Problems

Shun Watanabe
Tokyo University of Agriculture and Technology
Tokyo, Japan
email: shunwata@cc.tuat.ac.jp

Abstract—**The exponential strong converse for a coding problem states that, if a coding rate (or a rate pair) is beyond the theoretical limit, the correct decoding probability converges to zero exponentially. The exponential strong converse theorem was initiated by Arimoto and by Dueck and Körner for the pont-to-point channel coding; even though tight exponents have been identified for single-user problems and simple multi-user problems, such as the Slepian-Wolf problem, tight exponents have been unsolved for multi-user problems. In this tutorial paper, we revisit the exponential strong converse theorems, and provide alternative proofs for single-user problems via manipulations of information quantities as in the weak converse argument (called "change-of-measure argument" in the literature). Then, we present the recently obtained result by Takeuchi and Watanabe providing the tight exponential strong converse for the source coding with coded side-information.**

I. Introduction

The strong converse for a coding theorem claims that the optimal asymptotic rate possible with vanishing probability cannot be improved by allowing a fixed error probability. The exponential strong converse further claims that, if a coding rate is beyond the asymptotic limit, the correct decoding probability converges to zero exponentially. Proving such a claim was initiated by Arimoto for the channel coding problem [2]; later, the strong converse exponent was studied by Dueck and Körner in [4]; see also [10] for the equivalence of the two exponents derived in [2] and [4]. Also, the strong converse exponent for the Slepian-Wolf problem was derived by Oohama and Han in [13].

Even though the tight strong converse exponent for point-to-point problems or simple multi-user problems, such as the Slepian-Wolf problem, have been identified, the strong converse exponent for multi-user problems have been unsolved until recently. A significant progress was made by Oohama in a series of paper including [11], [12]. More recently, the tight strong converse of the Wyner-Ahlswede-Körner (WAK) problem [1], [18] was derived in [14]; the converse part of [14] is based on a manipulations of information quantities as in the weak converse argument, called the "change-of-measure argument" in [15]. In this tutorial paper, we provide alternative proofs of the strong converse exponents for single-user problems by using the same methodology.

The change-of-measure argument was originally introduced by Gu and Effros in [6], [7] to prove strong converse for source coding problems where there exists a terminal that observes all the random variables involved; a particular example is the Gray-Wyner (GW) problem [5]. In the argument of [6], [7], we evaluate the performance of a given code not under the original source (or channel) but under another modified measure which depends on the code and under which the code is error free.[1] A type based modification of this argument was used in [17] to derive the second-order rate region of the GW problem. A difficulty of applying this argument to the so-called distributed coding problems, such as the WAK problem, is that the characterization of asymptotic limits involve auxiliary random variables and Markov chain constraints. This technical difficulty was circumvented in [16] for the WAK problem by relating the WAK problem to an extreme case of the GW problem. By using the idea of "soft Markov constraint" introduced by Oohama [11], the argument was further developed in [15] so that it can be applied to distributed coding problems; furthermore, the argument was also extended so that it can be applied to secrecy problems such as the secret key generation and the wiretap channel. More recently, a variation of the change-of-measure argument was further developed by Hamad, Wigger, and Sarkiss in [8] so that it can be applied to more involved multi-user networks in a concise manner; rather than adding a Markov constraint as a penalty term, they prove the Markov constraint in an asymptotic limit.

II. Preliminaries

We use the same notations as [3]. For instance, the entropy of random variable X is denoted as $H(X)$; the mutual information between X and Y is denoted as $I(X \wedge Y)$; and the KL-divergence between distributions P and Q is denoted as $D(P\|Q)$. The logarithm is base 2.

Let $X^n = (X_1, \ldots, X_n)$ be an independently identically distributed (i.i.d.) source on a finite alphabet $\mathcal{X}$. For a given set $\mathcal{C} \subset \mathcal{X}^n$, a key step of the change-of-measure argument is to construct a modified measure by conditioning:

$$P_{\tilde{X}^n}(x^n) := \frac{P_{X^n}(x^n)\mathbf{1}[x^n \in \mathcal{C}]}{P_{X^n}(\mathcal{C})},$$

where $\mathbf{1}[\cdot]$ is the indicator function. A key observation, which was used in Marton's proof of the blowing-up lemma [9], is that the modified measure is not too far from the original measure in the following sense:

$$D(P_{\tilde{X}^n}\|P_{X^n}) = \sum_{x^n \in \mathcal{C}} P_{\tilde{X}^n}(x^n) \log \frac{P_{\tilde{X}^n}(x^n)}{P_{X^n}(x^n)}$$

[1] In the original argument [6], [7], the modified measure is constructed by conditioning on typical sets in addition to the error free set; on the other hand, the argument in [15] only conditions on the error free set.

$$= \log \frac{1}{P_{X^n}(\mathcal{C})}.$$

The conditional measure $P_{\tilde{X}^n}$ is not i.i.d. in general. By using the sub-additivity and concavity of entropy, we can directly derive a single-letter upper bound on the joint entropy as

$$H(\tilde{X}^n) \le \sum_{j=1}^n H(\tilde{X}_j) \le nH(\tilde{X}_J),$$

where J is the random variable uniformly distributed on the index set $\{1,\dots,n\}$. It is not possible to derive a single-letter lower bound on the joint entropy $H(\tilde{X}^n)$ directly; instead, we manipulate it with the divergence term:

$$\begin{aligned} H(\tilde{X}^n) + D(P_{\tilde{X}^n}\|P_{X^n}) &= \sum_{x^n} P_{\tilde{X}^n}(x^n) \log \frac{1}{P_{X^n}(x^n)} \\ &= \sum_{j=1}^n \sum_{x^n} P_{\tilde{X}^n}(x^n) \log \frac{1}{P_X(x_j)} \\ &= \sum_{j=1}^n \sum_{x} P_{\tilde{X}_j}(x) \log \frac{1}{P_X(x)} \\ &= n \sum_{x} P_{\tilde{X}_J}(x) \log \frac{1}{P_X(x)} \\ &= n\big[H(\tilde{X}_J) + D(P_{\tilde{X}_J}\|P_X)\big]. \quad (1) \end{aligned}$$

By the convexity of the KL-divergence, we can also derive a single-letter lower bound on the KL-divergence:

$$\begin{aligned} D(P_{\tilde{X}^n}\|P_{X^n}) &= \sum_{j=1}^n D(P_{\tilde{X}_j|\tilde{X}^{j-1}}\|P_X|P_{\tilde{X}^{j-1}}) \\ &\ge \sum_{j=1}^n D(P_{\tilde{X}_j}\|P_X) \\ &\ge nD(P_{\tilde{X}_J}\|P_X). \end{aligned}$$

The derivation of the strong converse exponent proceed by a judicious use of the above single-letter bounding manipulations.

III. Lossy Source Coding

In this section, we consider the lossy source coding. For a finite alphabet $\mathcal{X}$, let $X^n = (X_1,\dots,X_n)$ be an independently identically distributed (i.i.d.) source with distribution $P_{X^n} = P_X^n$. For a finite reproduction alphabet $\mathcal{Y}$, we consider an encoder $\varphi : \mathcal{X}^n \to \mathcal{M}$ and a decoder $\psi : \mathcal{M} \to \mathcal{Y}^n$. For a distortion measure $d : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$, let $d_n(x^n,y^n) = \sum_{j=1}^n d(x_j,y_j)$. For a distortion level $\Delta \ge 0$, we shall evaluate non-excess distortion probability:

$$p_c := \Pr\big(d_n(X^n, \psi(\varphi(X^n))) \le n\Delta\big).$$

For fixed Δ, a rate R is defined to be achievable if, for every $0 < \varepsilon < 1$ and for sufficiently large n, there exists a code (φ,ψ) such that the non-excess distortion probability satisfies $p_c \ge 1-\varepsilon$ and the coding rate satisfies $\frac{1}{n}\log|\mathcal{M}| \le R$. Then, the rate-distortion function $R(P_X,\Delta)$ is defined as the infimum of achievable rates. It is well known that the rate-distortion function is characterized as

$$R(P_X,\Delta) = \min_{\substack{P_{Y|X}: \\ \mathbb{E}[d(X,Y)]\le\Delta}} I(X\wedge Y).$$

We provide an alternative proof for the following exponential strong converse of the lossy source coding.

Proposition 1 For any code (φ,ψ) such that $\frac{1}{n}\log|\mathcal{M}| \le R$, the non-excess distortion probability satisfies

$$\frac{1}{n}\log(1/p_c) \ge \min_{P_{\tilde{X}}} \big[D(P_{\tilde{X}}\|P_X) + |R(P_{\tilde{X}},\Delta) - R|^+\big],$$

where $|a|^+ := \max[a,0]$.

Note that the exponent is positive if and only if $R < R(P_X,\Delta)$. It is known that the strong converse exponent in Proposition 1 is tight [3, Ex. 9.6].

Proof. Let

$$\mathcal{C} := \big\{x^n \in \mathcal{X}^n : d_n(x^n, \psi(\varphi(x^n))) \le n\Delta\big\},$$

and let

$$P_{\tilde{X}^n}(x^n) := \frac{P_{X^n}(x^n)\mathbf{1}[x^n\in\mathcal{C}]}{P_{X^n}(\mathcal{C})}.$$

Then, we have

$$D(P_{\tilde{X}^n}\|P_{X^n}) = \log(1/p_c)$$

and

$$\begin{aligned} \log(1/p_c) &= D(P_{\tilde{X}^n}\|P_{X^n}) \\ &\ge nD(P_{\tilde{X}_J}\|P_X). \end{aligned} \quad (2)$$

Note that the rate R can be lower bounded as

$$\begin{aligned} nR &\ge \log|\mathcal{M}| \\ &\ge H(\tilde{Y}^n) \\ &= I(\tilde{X}^n \wedge \tilde{Y}^n), \end{aligned}$$

where $\tilde{Y}^n = \psi(\varphi(\tilde{X}^n))$. Thus, we have

$$\begin{aligned} \log(1/p_c) &= D(P_{\tilde{X}^n}\|P_{X^n}) \\ &\ge D(P_{\tilde{X}^n}\|P_{X^n}) + I(\tilde{X}^n\wedge\tilde{Y}^n) - nR. \end{aligned}$$

Furthermore, by (1), we have

$$\begin{aligned} &D(P_{\tilde{X}^n}\|P_{X^n}) + I(\tilde{X}^n\wedge\tilde{Y}^n) \\ &= D(P_{\tilde{X}^n}\|P_{X^n}) + H(\tilde{X}^n) - H(\tilde{X}^n|\tilde{Y}^n) \\ &= nD(P_{\tilde{X}_J}\|P_X) + nH(\tilde{X}_J) - \sum_{j=1}^n H(\tilde{X}_j|\tilde{Y}^n, \tilde{X}_j^-) \\ &\ge nD(P_{\tilde{X}_J}\|P_X) + nH(\tilde{X}_J) - \sum_{j=1}^n H(\tilde{X}_j|\tilde{Y}_j) \\ &= nD(P_{\tilde{X}_J}\|P_X) + nH(\tilde{X}_J) - nH(\tilde{X}_J|\tilde{Y}_J, J) \\ &\ge nD(P_{\tilde{X}_J}\|P_X) + nH(\tilde{X}_J) - nH(\tilde{X}_J|\tilde{Y}_J) \\ &= nD(P_{\tilde{X}_J}\|P_X) + nI(\tilde{X}_J\wedge\tilde{Y}_J), \end{aligned}$$

where $\tilde{X}_j^- = (\tilde{X}_1, \ldots, \tilde{X}_{j-1})$. Also, since the support of the changed measure $P_{\tilde{X}^n}$ is $\mathcal{C}$, note that

$$\Delta \geq \mathbb{E}\left[\frac{1}{n} d_n(\tilde{X}^n, \tilde{Y}^n)\right] = \mathbb{E}[d(\tilde{X}_J, \tilde{Y}_J)].$$

Thus, we have

$$\begin{aligned}\frac{1}{n}\log(1/p_c) &\geq D(P_{\tilde{X}_J}\|P_X) + \big(I(\tilde{X}_J \wedge \tilde{Y}_J) - R\big) \\ &\geq D(P_{\tilde{X}_J}\|P_X) + \big(R(P_{\tilde{X}_J}, \Delta) - R\big). \qquad (3)\end{aligned}$$

By combining (2) and (3), we have

$$\frac{1}{n}\log(1/p_c) \geq D(P_{\tilde{X}_J}\|P_X) + |R(P_{\tilde{X}_J}, \Delta) - R|^+.$$

Finally, by replacing $P_{\tilde{X}_J}$ with the minimum over $P_{\tilde{X}}$, we have the claim of the proposition. ■

IV. Channel Coding

In this section, we consider the channel coding. Let W^n be a discrete memoryless channel (DMC) from a finite input alphabet $\mathcal{X}$ to a finite output alphabet $\mathcal{Y}$. For a message set $\mathcal{M}$, a channel code consists of an encoder $\varphi : \mathcal{M} \to \mathcal{X}^n$ and a decoder $\psi : \mathcal{Y}^n \to \mathcal{M}$. Let

$$p_c := \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{M}|} W^n(\psi^{-1}(m)|\varphi(m))$$

be the average correct decoding probability. A rate R is defined to be achievable if, for every $0 < \varepsilon < 1$ and for sufficiently large n, there exists a code (φ, ψ) such that the average correct decoding probability satisfies $p_c \geq 1 - \varepsilon$ and the coding rate satisfies $\frac{1}{n}\log|\mathcal{M}| \geq R$. Then, the channel capacity $C(W)$ is defined as the supremum of achievable rates. It is well known that the channel capacity is characterized as

$$C(W) = \max_{P_X} I(X \wedge Y),$$

where the mutual information is evaluated with respect to (X, Y) induced by the input distribution P_X and the channel W.

We provide an alternative proof for the following exponential strong converse of the channel coding.

Proposition 2 For any code (φ, ψ) such that $\frac{1}{n}\log|\mathcal{M}| \geq R$, the average correct decoding probability satisfies

$$\frac{1}{n}\log(1/p_c) \geq \min_{P_{\tilde{X}\tilde{Y}}} \big[D(P_{\tilde{Y}|\tilde{X}}\|W|P_{\tilde{X}}) + |R - I(\tilde{X} \wedge \tilde{Y})|^+\big].$$

Note that the exponent is positive if and only if $R > C(W)$. It is known that the strong converse exponent in Proposition 2 is tight [4]. Furthermore, it also coincides with the strong converse exponent by Arimoto [2]; see [10] for the equivalence.

Proof. Let

$$\mathcal{C} := \big\{(m, x^n, y^n) : \psi(y^n) = m\big\}.$$

For

$$P_{MX^nY^n}(m, x^n, y^n) = \frac{1}{|\mathcal{M}|}\mathbf{1}[x^n = \varphi(m)]W^n(y^n|x^n),$$

let

$$\begin{aligned}&P_{\tilde{M}\tilde{X}^n\tilde{Y}^n}(m, x^n, y^n) \\ &\quad := \frac{P_{MX^nY^n}(m, x^n, y^n)\mathbf{1}[(m, x^n, y^n) \in \mathcal{C}]}{P_{MX^nY^n}(\mathcal{C})}.\end{aligned}$$

Then, we have

$$D(P_{\tilde{M}\tilde{X}^n\tilde{Y}^n}\|P_{MX^nY^n}) = \log(1/p_c).$$

By noting that $P_{Y^n|MX^n} = W^n$, we have

$$\begin{aligned}D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) &= D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|P_{Y^n|MX^n}|P_{\tilde{M}\tilde{X}^n}) \\ &\leq D(P_{\tilde{M}\tilde{X}^n\tilde{Y}^n}\|P_{MX^nY^n}) \\ &= \log(1/p_c).\end{aligned}$$

By the convexity of the KL-divergence, we also have

$$\begin{aligned}D(P_{\tilde{Y}^n|\tilde{X}^n}\|W^n|P_{\tilde{X}^n}) &\leq D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) \\ &\leq \log(1/p_c). \qquad (4)\end{aligned}$$

Furthermore, by the monotonicity of the KL-divergence, we also have

$$\begin{aligned}&D(P_{\tilde{M}}\|P_M) + D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) \\ &\leq D(P_{\tilde{M}\tilde{X}^n}\|P_{MX^n}) + D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) \\ &= D(P_{\tilde{M}\tilde{X}^n\tilde{Y}^n}\|P_{MX^nY^n}) \\ &= \log(1/p_c). \qquad (5)\end{aligned}$$

Now, by noting that P_M is uniform distribution on $\mathcal{M}$, we have

$$\begin{aligned}&nR \\ &\leq \log|\mathcal{M}| \\ &= H(\tilde{M}) + D(P_{\tilde{M}}\|P_M) \\ &= I(\tilde{M} \wedge \tilde{Y}^n) + D(P_{\tilde{M}}\|P_M) \\ &\leq I(\tilde{M} \wedge \tilde{Y}^n) + D(P_{\tilde{M}}\|P_M) \\ &\quad + \big[\log(1/p_c) - D(P_{\tilde{M}}\|P_M) - D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n})\big] \\ &= I(\tilde{M} \wedge \tilde{Y}^n) - D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) + \log(1/p_c) \\ &= I(\tilde{M}, \tilde{X}^n \wedge \tilde{Y}^n) - D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) + \log(1/p_c) \\ &= I(\tilde{X}^n \wedge \tilde{Y}^n) + I(\tilde{M} \wedge \tilde{Y}^n|\tilde{X}^n) \\ &\quad - D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) + \log(1/p_c) \\ &= I(\tilde{X}^n \wedge \tilde{Y}^n) + D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|P_{\tilde{Y}^n|\tilde{X}^n}|P_{\tilde{M}\tilde{X}^n}) \\ &\quad - D(P_{\tilde{Y}^n|\tilde{M}\tilde{X}^n}\|W^n|P_{\tilde{M}\tilde{X}^n}) + \log(1/p_c) \\ &= I(\tilde{X}^n \wedge \tilde{Y}^n) - D(P_{\tilde{Y}^n|\tilde{X}^n}\|W^n|P_{\tilde{X}^n}) + \log(1/p_c), \qquad (6)\end{aligned}$$

where the second equality follows since $\tilde{M}$ can be decoded from $\tilde{Y}^n$ with 0 error probability, the second inequality follows from (5), and the forth equality follows since $\tilde{X}^n$ is a function of $\tilde{M}$.[2] Now, we conduct the single-letter procedure as follows:

$$\begin{aligned}
& I(\tilde{X}^n \wedge \tilde{Y}^n) - D(P_{\tilde{Y}^n|\tilde{X}^n} \| W^n | P_{\tilde{X}^n}) \\
&= H(\tilde{Y}^n) - H(\tilde{Y}^n|\tilde{X}^n) - D(P_{\tilde{Y}^n|\tilde{X}^n} \| W^n | P_{\tilde{X}^n}) \\
&= H(\tilde{Y}^n) - \sum_{x^n,y^n} P_{\tilde{X}^n\tilde{Y}^n}(x^n,y^n) \log \frac{1}{W^n(y^n|x^n)} \\
&= H(\tilde{Y}^n) - n\sum_{x,y} P_{\tilde{X}_J\tilde{Y}_J}(x,y) \log \frac{1}{W(y|x)} \\
&= H(\tilde{Y}^n) - nH(\tilde{Y}_J|\tilde{X}_J) - nD(P_{\tilde{Y}_J|\tilde{X}_J} \| W | P_{\tilde{X}_J}) \\
&\le nH(\tilde{Y}_J) - nH(\tilde{Y}_J|\tilde{X}_J) - nD(P_{\tilde{Y}_J|\tilde{X}_J} \| W | P_{\tilde{X}_J}) \\
&= nI(\tilde{X}_J \wedge \tilde{Y}_J) - nD(P_{\tilde{Y}_J|\tilde{X}_J} \| W | P_{\tilde{X}_J}). \qquad (7)
\end{aligned}$$

Thus, by combining (6) and (7), we have

$$\frac{1}{n}\log(1/p_c) \ge D(P_{\tilde{Y}_J|\tilde{X}_J} \| W | P_{\tilde{X}_J}) + \big(R - I(\tilde{X}_J \wedge \tilde{Y}_J)\big). \qquad (8)$$

Note also that

$$\begin{aligned}
D(P_{\tilde{Y}^n|\tilde{X}^n} \| W^n | P_{\tilde{X}^n}) &= \sum_{j=1}^{n} D(P_{\tilde{Y}_j|\tilde{X}^n\tilde{Y}_j^-} \| W | P_{\tilde{X}^n\tilde{Y}_j^-}) \\
&\ge \sum_{j=1}^{n} D(P_{\tilde{Y}_j|\tilde{X}_j} \| W | P_{\tilde{X}_j}) \\
&= nD(P_{\tilde{Y}_J|\tilde{X}_J J} \| W | P_{\tilde{X}_J J}) \\
&\ge nD(P_{\tilde{Y}_J|\tilde{X}_J} \| W | P_{\tilde{X}_J}). \qquad (9)
\end{aligned}$$

Thus, by combining (4) and (9), we have

$$\frac{1}{n}\log(1/p_c) \ge D(P_{\tilde{Y}_J|\tilde{X}_J} \| W | P_{\tilde{X}_J}). \qquad (10)$$

Thus, by combining (8) and (10), we have

$$\frac{1}{n}\log(1/p_c) \ge D(P_{\tilde{Y}_J|\tilde{X}_J} \| W | P_{\tilde{X}_J}) + |R - I(\tilde{X}_J \wedge \tilde{Y}_J)|^+.$$

Finally, by replacing $P_{\tilde{X}_J\tilde{Y}_J}$ with the minimum over $P_{\tilde{X}\tilde{Y}}$, we have the claim of the proposition. ■

V. Source Coding with Coded Side-information

In this section, we consider the source coding with coded side-information, also known as the Wyner-Ahlswede-Körner (WAK) problem [1], [18]. For finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, let (X^n, Y^n) be i.i.d. correlated source with distribution $P_{X^nY^n}$. A code consists of two encoders $\varphi_1 : \mathcal{X}^n \to \mathcal{M}_1$ and $\varphi_2 : \mathcal{Y}^n \to \mathcal{M}_2$, and a decoder $\psi : \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{X}^n$. We shall evaluate the correct decoding probability:

$$p_c := \Pr\big(\psi(\varphi_1(X^n), \varphi_2(Y^n)) = X^n\big).$$

A rate pair (R_1, R_2) is defined to be achievable if, for every $0 < \varepsilon < 1$ and for sufficiently large n, there exists a code $(\varphi_1, \varphi_2, \psi)$ such that the correct decoding probability satisfies $p_c \ge 1 - \varepsilon$ and rate pair satisfies $\frac{1}{n}\log|\mathcal{M}_1| \le R_1$ and $\frac{1}{n}\log|\mathcal{M}_2| \le R_2$, respectively. Then, the achievable region $\mathcal{R}_{\mathtt{WAK}}(P_{XY})$ is defined as the closure of all achievable rate pairs. It is well known that the achievable region is characterized as

$$\begin{aligned}
\mathcal{R}_{\mathtt{WAK}}(P_{XY}) = \big\{(R_1, R_2) : \exists P_{U|Y} \in \mathcal{P}(\mathcal{U}|\mathcal{Y}) \text{ s.t.} \\
R_1 \ge H(X|U), R_2 \ge I(U \wedge Y)\big\}
\end{aligned}$$

where $\mathcal{P}(\mathcal{U}|\mathcal{Y})$ is the set of all channels from $\mathcal{Y}$ to an auxiliary alphabet $\mathcal{U}$ satisfying $|\mathcal{U}| \le |\mathcal{Y}| + 1$.

Note that the characterization of the achievable region involves an auxiliary random variable U that does not appear in the problem setting. Furthermore, U is generated only from Y via channel $P_{U|Y}$; in other words, U, Y, and X must satisfy the Markov chain. In many cases, difficulty of analyzing multi-user problems stem from the existence of auxiliary random variables and Markov chain constraints, and the WAK problem is the most basic problem involving such difficulties.

The following exponential strong converse of the WAK problem was obtained in [14].

Proposition 3 For any code $(\varphi_1, \varphi_2, \psi)$, the correct decoding probability satisfies

$$\begin{aligned}
& \frac{1}{n}\log(1/p_c) \\
& \ge \min_{P_{\tilde{U}\tilde{X}\tilde{Y}}} \big\{D(P_{\tilde{U}\tilde{X}\tilde{Y}} \| P_{\tilde{U}|\tilde{Y}}P_{XY}) + |I(\tilde{U} \wedge \tilde{Y}) - R_2|^+ : \\
& \qquad\qquad R_1 \ge H(\tilde{X}|\tilde{U})\big\},
\end{aligned}$$

where the minimization is taken over joint distributions on $\mathcal{U}\times\mathcal{X}\times\mathcal{Y}$ for an auxiliary alphabet satisfying $|\mathcal{U}| \le |\mathcal{X}||\mathcal{Y}|+2$.

For the proof, see [14]; furthermore, it can be proved that the bound in Proposition 3 is asymptotically tight.

In contrast to the characterization of the achievable region, the exponent in Proposition 3 does not involve the Markov chain constraint. In fact, we can decompose the divergence term as

$$D(P_{\tilde{U}\tilde{X}\tilde{Y}} \| P_{\tilde{U}|\tilde{Y}}P_{XY}) = D(P_{\tilde{X}\tilde{Y}} \| P_{XY}) + I(\tilde{U} \wedge \tilde{X}|\tilde{Y}).$$

Thus, in the analysis of the strong converse exponent, the Markov chain constraint is imposed as a (potentially non-zero) penalty term. The idea of introducing this kind of penalty term rather than the exact Markov chain constraint was proposed by Oohama in [11], which culminated in the tight strong exponent of the WAK problem in [14].

References

[1] R. Ahlswede and J. Körner, "Source coding with side information and a converse for the degraded broadcast channel," *IEEE Trans. Inform. Theory*, vol. 21, no. 6, pp. 629–637, November 1975.

[2] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 19, no. 3, pp. 357–359, May 1973.

[3] I. Csiszár and J. Körner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[2] Note that $\tilde{M}$ may not be a function of $\tilde{X}^n$ when the encoder is not one-to-one, and $I(\tilde{M} \wedge \tilde{Y}^n|\tilde{X}^n)$ may not be 0.

[4] G. Dueck and J. Körner, “Reliability function of a discrete memoryless channel at rates above capacity,” *IEEE Trans. Inform. Theory*, vol. 25, no. 1, pp. 82–85, January 1979.

[5] R. M. Gray and A. D. Wyner, “Source coding for a simple network,” *Bell Labs. Technical Journal*, vol. 53, no. 9, pp. 1681–1721, November 1974.

[6] W. Gu and M. Effros, “A strong converse for a collection of network source coding problems,” in *Proc. IEEE Int. Symp. Inf. Theory 2009*, 2009, pp. 2316–2320.

[7] ——, “A strong converse in source coding for super-source networks,” in *Proc. IEEE Int. Symp. Inf. Theory 2011*, 2011, pp. 395–399.

[8] M. Hamad, M. A. Wigger, and M. Sarkiss, “Strong converse using typical changes of measures and asymptotic markov chains,” 2023, arXiv:2301.06289.

[9] K. Marton, “A simple proof of the blowing-up lemma,” *IEEE Trans. Inform. Theory*, vol. 32, no. 3, pp. 445–446, May 1986.

[10] Y. Oohama, “On two strong converse theorems for discrete memoryless channels,” *IEICE Trans. Fundamentals*, vol. E98-A, no. 12, pp. 2471–2475, December 2015.

[11] ——, “Exponential strong converse for source coding with side information at the decoder,” *Entropy*, vol. 20, no. 5, p. 352, April 2018.

[12] ——, “Exponential strong converse for one helper source coding problem,” *Entropy*, vol. 21, no. 6, p. 567, June 2019.

[13] Y. Oohama and T. S. Han, “Universal coding for the Slepian-Wolf data compression system and the strong converse theorem,” *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1908–1919, November 1994.

[14] D. Takeuchi and S. Watanabe, “Tight exponential strong converse for source coding problem with encoded side information,” in *Proc. IEEE Int. Symp. Inf. Theory 2023*, Taipei, Taiwan, 2023, pp. 1366–1371.

[15] H. Tyagi and S. Watanabe, “Strong converse using change of measure arguments,” *IEEE Trans. Inform. Theory*, vol. 66, no. 2, pp. 689–703, February 2020.

[16] S. Watanabe, “A converse bound on Wyner-Ahlswede-Körner network via Gray-Wyner network,” in *Proc. 2017 IEEE Information Theory Workshop (ITW)*, Kaohsiung, Taiwan, 2017.

[17] ——, “Second-order region for Gray-Wyner network,” *IEEE Trans. Inform. Theory*, vol. 63, no. 2, pp. 1006–1018, February 2017.

[18] A. D. Wyner, “On source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.

Improved Capacity Outer Bound for Private Quadratic Monomial Computation

Karen M. Dæhli, Sarah A. Obead, Hsuan-Yin Lin, and Eirik Rosnes
Simula UiB, N–5006 Bergen, Norway
Emails: kamadaehli@gmail.com, {sarah, lin, eirikrosnes}@simula.no

***Abstract*—In private computation, a user wishes to retrieve a function evaluation of messages stored on a set of databases without revealing the function's identity to the databases. Obead *et al.* introduced a capacity outer bound for private nonlinear computation, dependent on the order of the candidate functions. Focusing on private *quadratic monomial* computation, we propose three methods for ordering candidate functions: a graph edge-coloring method, a graph-distance method, and an entropy-based greedy method. We confirm, via an exhaustive search, that all three methods yield an optimal ordering for $f < 6$ messages. For $6 \leq f \leq 12$ messages, we numerically evaluate the performance of the proposed methods compared with a directed random search. For almost all scenarios considered, the entropy-based greedy method gives the smallest gap to the best-found ordering.**

I. Introduction

Private computation (PC) [1] is a generalization of the renowned private information retrieval (PIR) problem that aims at addressing privacy concerns in distributed computing services. For example, in distributed machine learning, many common classification, dimensionality reduction, and linear regression algorithms operate on the inner products of the data samples rather than the individual data samples. In PC, the user wants to privately download a function evaluation of the messages stored across a set of databases, i.e., without leaking any information to the databases (in an information-theoretic manner) on the identity of the desired function evaluation.

To measure the efficiency of a PC protocol, the PC rate, defined as the ratio between the (smallest) size of the function evaluation and the number of downloaded symbols, is typically considered. The maximum PC rate is referred to as the PC capacity, and it is known for the case of linear function evaluations, referred to as private linear computation (PLC), from noncolluding replicated and coded databases [1]–[3]. Private polynomial computation (PPC) was first considered by Karpuk in [4], and later in [5]–[8]. The capacity of PPC is still unknown, and there is generally a substantial gap between the best achievable rate and the best-known capacity outer bound. Private inner product retrieval from noncolluding replicated databases was considered in [9], while the general case of PC for nonlinear function evaluations was considered in [10] where an outer bound on the capacity was first introduced. It was noted in [10] that the value of the PC capacity outer bound depends on how the candidate functions are ordered.[1] To the best of our knowledge, an optimal order for the outer bound has not yet been considered in the open literature.

In this work, inspired by private inner product retrieval [9], we focus on the case of private quadratic nonparallel monomial computation (PQNMC) and propose three methods for finding a *good* ordering of the μ candidate functions. In PQNMC, the set of candidate functions is the set of all quadratic *nonparallel* monomials of f messages where each message symbol is chosen from a size-q finite field $\mathbb{F}_q$, thus, $\mu = {}^{f(f-1)}/_{2}$. Given that graphs present a strong framework for illustrating the interdependence among random variables, offering insights into the dependency structure of the candidate function set, we propose two graph-based methods. The first is an edge-coloring method we name *(enhanced) edge-coloring* ((E-)EC) and the second is a graph-distance method we name *longest-distance first* (LDF). Then, we compare the resulting PC capacity outer bound with the one found by our third method: an *entropy-based greedy* (EBG) algorithm.

For $f < 6$ messages chosen from $\mathbb{F}_2$, we verify through an exhaustive search that the proposed methods output *optimal* orders. However, we note that the orders are not unique and are finite field-dependent, which illustrates the difficulty of finding a general ordering of quadratic nonparallel monomials that will optimize the outer bound of the PQNMC capacity. Moreover, for larger numbers of messages, an exhaustive search quickly becomes infeasible even over $\mathbb{F}_2$. As a result, we opt for a *directed* random search to numerically analyze the performance of the proposed methods. Accordingly, we note that for $6 \leq f \leq 10$ and $f = 12$ messages, the EBG algorithm outperforms the proposed graph-based methods with the smallest gap to the best-found ordering. Nevertheless, the significance of the graph-based methods arises as a relatively low-complexity alternative to the EBG method as the complexity of computing the entropies needed for the EBG algorithm grows exponentially with the number of candidate monomials μ with base equal to the size q of the underlying finite field $\mathbb{F}_q$. Finally, although we consider PQNMC in this work, we note that the results may also have independent interest beyond PC.

II. Preliminaries

A. Notation

We denote by $\mathbb{N}$ the set of all positive integers and $[a] \triangleq \{1, 2, \ldots, a\}$ for $a \in \mathbb{N}$. A random variable is denoted by a capital Roman letter, e.g., X, while its realization is denoted by the corresponding small Roman letter, e.g., x.

[1] Due to the inherent relation between PC and PIR, a similar observation was made in [11] for PIR with dependent messages, i.e., dependent PIR (DPIR).

Vectors are boldfaced, e.g., $\boldsymbol{X}$ denotes a random vector, and $\boldsymbol{x}$ denotes a deterministic vector. Sets are denoted by calligraphic uppercase letters, e.g., $\mathcal{X}$. Concatenation of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_a$ is represented by $(\boldsymbol{x}_1 \mid \cdots \mid \boldsymbol{x}_a)$. Furthermore, some constants and functions are depicted by Greek letters or a special font, e.g., X. The entropy of X is represented by $\mathsf{H}(X)$. A degree g monomial $\boldsymbol{z}^{\boldsymbol{i}}$ in f variables $z_1, \ldots, z_f$ over a finite field $\mathbb{F}_q$ is written as $\boldsymbol{z}^{\boldsymbol{i}} = z_1^{i_1} \cdots z_f^{i_f}$, where $\boldsymbol{i} \triangleq (i_1, \ldots, i_f) \in (\{0\} \cup \mathbb{N})^f$ is the exponent vector with $\sum_{j=1}^{f} i_j = g$ and $i_j \leq q-1$ for all $j \in [f]$.[2] A simple undirected graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$ is denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

B. Problem Statement

We consider the PQNMC problem, which is formally described as follows.[3] Consider a distributed storage system (DSS) consisting of n noncolluding databases, each storing a replica of f independent messages. The messages are denoted by $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(f)}$ and each message $\boldsymbol{W}^{(m)} = \bigl(W_1^{(m)}, \ldots, W_{\beta\mathsf{L}}^{(m)}\bigr)$, $m \in [f]$, is a length-$\beta\mathsf{L}$ vector with independent and identically distributed symbols that are chosen uniformly at random from the field $\mathbb{F}_q$ for some $\beta, \mathsf{L} \in \mathbb{N}$.[4] Hence, we have

$$\mathsf{H}\bigl(\boldsymbol{W}^{(m)}\bigr) = \beta\mathsf{L}, \forall\, m \in [f],$$
$$\mathsf{H}\bigl(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(f)}\bigr) = f\beta\mathsf{L} \quad \text{(in q-ary units).}$$

In PQNMC, a user wishes to privately compute exactly one quadratic nonparallel monomial $X_i^{(k,\ell)} \triangleq W_i^{(k)} W_i^{(\ell)}$, $\forall\, i \in [\beta\mathsf{L}]$, for some $k, \ell \in [f]$, $k < \ell$, out of $\mu \triangleq \binom{f}{2}$ *candidate* quadratic nonparallel monomials. For convenience, we denote by $\mathcal{T} \triangleq \{(k,\ell) \colon k, \ell \in [f], k < \ell\}$ the set of all ordered 2-tuples, where $|\mathcal{T}| = \mu$. In q-ary units, we have

$$\mathsf{H}\bigl(\boldsymbol{X}^{(k,\ell)}\bigr) = \beta\mathsf{L}\,\mathsf{H}\bigl(X^{(k,\ell)}\bigr), \forall\,(k,\ell) \in \mathcal{T},$$
$$\mathsf{H}\bigl(\boldsymbol{X}^{(1,2)}, \ldots, \boldsymbol{X}^{(f-1,f)}\bigr) = \beta\mathsf{L}\,\mathsf{H}\bigl(X^{(1,2)}, \ldots, X^{(f-1,f)}\bigr),$$

for *prototype* random variables $X^{(k,\ell)}$.

The user privately selects an index (k,ℓ) and wishes to compute the (k,ℓ)-th quadratic nonparallel monomial while keeping the requested index (k,ℓ) private from each database. In order to retrieve the desired function $\boldsymbol{X}^{(k,\ell)}$, $(k,\ell) \in \mathcal{T}$, from the DSS, the user sends a random query to the j-th database for all $j \in [n]$. The user generates the queries without any prior knowledge of the realizations of the stored messages, and they are independent of the candidate quadratic nonparallel monomials. In response to the received query, the j-th database sends an answer back to the user.

To measure the efficiency of a PC protocol, we consider the required number of downloaded q-ary symbols for retrieving the $\beta\mathsf{L}$ q-ary symbols of the desired function evaluation.

Definition 1 (PQNMC Rate and Capacity): The rate of a PQNMC protocol, denoted by R, is defined as the ratio between the *smallest* desired monomial size $\beta\mathsf{L}\,\mathsf{H}_{\mathsf{min}}$ and the total required download cost D, i.e.,

$$\mathsf{R} \triangleq \frac{\beta\mathsf{L}\,\mathsf{H}_{\mathsf{min}}}{\mathsf{D}},$$

where $\mathsf{H}_{\mathsf{min}} \triangleq \min_{(k,\ell)\in\mathcal{T}} \mathsf{H}\bigl(X^{(k,\ell)}\bigr)$. The PQNMC *capacity*, denoted by $\mathsf{C}_{\mathsf{PQNMC}}$, is the maximum achievable PQNMC rate over all possible PQNMC protocols.

Note that for quadratic nonparallel monomials, we have $\mathsf{H}_{\mathsf{max}} \triangleq \max_{(k,\ell)\in\mathcal{T}} \mathsf{H}\bigl(X^{(k,\ell)}\bigr) = \mathsf{H}_{\mathsf{min}}$. Accordingly, every quadratic nonparallel monomial carries the same amount of information and following the terminology of DPIR [11] we denote the PQNMC problem as *balanced*.

Let $\mathcal{P}(\mathcal{T})$ be the set of all permutations on the set $\mathcal{T}$, and denote an ordered set $\mathcal{S} \triangleq (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_\mu) \in \mathcal{P}(\mathcal{T})$. Using the same approach as [10], it can be shown that the PQNMC capacity is bounded from above by $\mathsf{C}_{\mathsf{PQNMC}} \leq \overline{\mathsf{C}}(\mathcal{S})$, where

$$\overline{\mathsf{C}}(\mathcal{S}) \triangleq \frac{n^{\mu}\,\mathsf{H}_{\mathsf{min}}}{\sum_{v=1}^{\mu} n^{\mu-v+1}\,\mathsf{H}\bigl(X^{(\boldsymbol{s}_v)} \,\big|\, X^{(\boldsymbol{s}_1)}, \ldots, X^{(\boldsymbol{s}_{v-1})}\bigr)}. \tag{1}$$

The goal of this work is to determine the best (lowest) outer bound to the PQNMC capacity $\mathsf{C}_{\mathsf{PQNMC}}$, among all the possible orders of the quadratic nonparallel monomials for a given number of messages f, i.e., we are interested in obtaining the best-ordered set that achieves $\min_{\mathcal{S}\in\mathcal{P}(\mathcal{T})} \overline{\mathsf{C}}(\mathcal{S})$.

Remark 1: We have observed by exhaustive search for $f \leq 5$ that the set of optimal orderings (the ones that minimize the capacity outer bound in (1)) is independent of $n \geq 2$, which suggests that one can choose $n = 2$ for finding an optimal order for the capacity outer bound $\overline{\mathsf{C}}(\mathcal{S})$.

C. Edge-Coloring and Matching

Quadratic nonparallel monomials in f variables can be represented by a simple undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = [f]$ where the vertices k and ℓ correspond to the messages $W^{(k)}$ and $W^{(\ell)}$ (for *prototype* random variables $W^{(k)}$ and $W^{(\ell)}$), respectively, and the edge $(k,\ell) \in \mathcal{E}$ represents the nonparallel monomial $X^{(k,\ell)}$. Thus, we have $\mu = |\mathcal{E}|$ and the set of all quadratic nonparallel monomials in f variables are represented by a *complete* graph $\mathcal{K}_f$.

Definition 2 (Distance): Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the distance between two vertices $u, v \in \mathcal{V}$, denoted by $d(u,v)$, is the length of the shortest path connecting them, measured in number of edges.

Definition 3 (Matching): Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a matching $\mathcal{M}$ in $\mathcal{G}$ is a set of pairwise nonadjacent edges, i.e., a set of edges where no two edges share common vertices. When $|\mathcal{V}|$ is an even number, a *perfect matching* is a matching that includes all vertices of the graph, and when $|\mathcal{V}|$ is an odd number, a *near-perfect matching* is a matching that includes $|\mathcal{V}| - 1$ vertices of the graph.

Definition 4 (Edge-Coloring [13, Ch. 17]): A *proper* edge-coloring of a graph is an assignment of colors to the edges

[2] For nonvanishing polynomials, following the Combinatorial Nullstellensatz theorem [12, Thm. 1.2], the degree of every variable in a multivariate polynomial must be strictly smaller than the finite field size.

[3] A monomial $m(\boldsymbol{z})$ is said to be *parallel* if it can be raised by another monomial to a positive integer power, i.e., $m(\boldsymbol{z}) = (\boldsymbol{z}^{\boldsymbol{i}})^d$ for some $d \in \mathbb{N}$ and $\boldsymbol{i} \in (\{0\} \cup \mathbb{N})^f$.

[4] For consistency, we use the notation required for the achievable rate in [10, Thm. 2] where asymptotically $\mathsf{L} \to \infty$ but β is fixed.

such that the edges incident to a vertex have distinct colors. A graph $\mathcal{G}$ is said to be κ-edge colorable if $\mathcal{G}$ has a proper edge-coloring with κ colors. The *chromatic index* of a graph $\mathcal{G}$, denoted by $\chi'(\mathcal{G})$, is the minimum number of colors required to properly edge-color $\mathcal{G}$.

The definition of proper edge-coloring of a graph $\mathcal{G}$ implies that the κ-edge coloring of a graph partitions the graph edge set $\mathcal{E}$ into κ (near) perfect matchings $\mathcal{M}_1, \ldots, \mathcal{M}_\kappa$ such that $\mathcal{E} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_\kappa$, and the sets $\mathcal{M}_1, \ldots, \mathcal{M}_\kappa$ are known as the color sets of edges. For complete graphs of f vertices, denoted by $\mathcal{K}_f$, it is known [14, Thm. 1] that

$$\chi'(\mathcal{K}_f) = \begin{cases} f-1 & \text{if } f \text{ is even,} \\ f & \text{if } f \text{ is odd.} \end{cases}$$

Remark 2: Let $\mathscr{M}$ be the set of all (near) perfect matchings of a complete graph $\mathcal{K}_f$. Let $\mathcal{S}[1:\eta] = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_\eta)$ be the first η elements of the ordered set $\mathcal{S}$, where $\eta \triangleq {}^{f(f-1)}\!/_{2\chi'(\mathcal{K}_f)}$ is the number of edges within a complete graph matching. For $\mathcal{S}$ to be an *optimal* order of quadratic nonparallel monomials of $f > 3$ variables, we conjecture that $\mathcal{S}[1:\eta]$ must constitute a (near) perfect matching of $\mathcal{K}_f$, i.e., $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_\eta\} \in \mathscr{M}$.

III. Algorithms to Determine a Good Order

In this section, we propose three methods for finding a *good* order, one based on edge-coloring, one based on graph distance, and one entropy-based greedy algorithm.

Remark 3: Computing the capacity bound of (1) for a given order entails the computation of μ conditional entropies, resulting in a computational complexity of order $\mathcal{O}(q^\mu)$. Taking that into account, it can be seen that performing an exhaustive search over all possible orders to optimize the PQNMC capacity outer bound would intuitively require a complexity of order $\mathcal{O}(\mu! \times q^\mu)$.

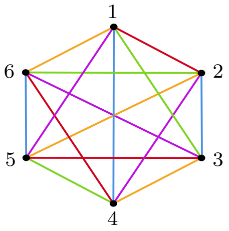
A. Edge-Coloring

The key idea is to first find a proper edge-coloring of the complete graph $\mathcal{K}_f$. Then, build an order based on grouping the edges according to their color, i.e., first take the edges corresponding to one of the colors, then the edges corresponding to another color, etc., until all colors have been considered. The time complexity of finding an order based on simple edge-coloring follows from the time complexity of finding a proper edge-coloring of $\mathcal{K}_f$. We follow the procedure in [15, App. A] which runs in polynomial time, i.e., of order $\mathcal{O}(\mu)$.

We first give an example to illustrate how edge-coloring can give us a good order $\mathcal{S}_{\text{EC}}$ for the capacity outer bound.

Example 1: For $f = 6$ messages, a proper 5-edge coloring of $\mathcal{K}_6$ is as follows:

Purple	$(1,5)(2,4)(3,6)$
Yellow	$(1,6)(2,5)(3,4)$
Red	$(1,2)(3,5)(4,6)$
Green	$(1,3)(2,6)(4,5)$
Blue	$(1,4)(2,3)(5,6)$



The resulting order is

$$\mathcal{S}_{\text{EC}} = \big((1,5),(2,4),(3,6),(1,6),(2,5),(3,4),(1,2),\\ (3,5),(4,6),(1,3),(2,6),(4,5),(1,4),(2,3),(5,6)\big),$$

and for $n = 2$ and $q = 2$, the capacity outer bound is $\overline{\mathrm{C}}(\mathcal{S}_{\text{EC}}) = 0.5198943946817$.

The permutation of the colors affects the value of $\overline{\mathrm{C}}$, which indicates that edge-coloring by itself does not guarantee finding an optimal order. Thus, a search over color permutations for the capacity bound, i.e., over $(\chi'(\mathcal{K}_f) - 1)!$ permutations (the first η edges corresponding to a single color can be fixed; see Remark 2), can potentially improve it in exchange for added complexity. We refer to this improved method as enhanced edge-coloring (E-EC), and its complexity is of order $\mathcal{O}(\mu + (\chi'(\mathcal{K}_f) - 1)! \times q^\mu)$. For example, if we reorder the colors in Example 1 as (purple, yellow, blue, red, green), we obtain the order

$$\mathcal{S}_{\text{E-EC}} = \big((1,5),(2,4),(3,6),(1,6),(2,5),(3,4),(1,4),\\ (2,3),(5,6),(1,2),(3,5),(4,6),(1,3),(2,6),(4,5)\big),$$

which results, for $n = 2$ and $q = 2$, in the improved capacity outer bound $\overline{\mathrm{C}}(\mathcal{S}_{\text{E-EC}}) = 0.5198121367672$.

We have observed that even within a set of edges of a given color, the permutation of the edges also affects the value of $\overline{\mathrm{C}}$. For instance, considering the blue color in Example 1, the values of $\overline{\mathrm{C}}$ between the orders $\{(1,4),(2,3),(5,6)\}$ and $\{(2,3),(1,4),(5,6)\}$ are different. However, when searching for the best order within the edges of every color, the computational complexity becomes $\mathcal{O}(\mu + ((\eta!)^{\chi'(\mathcal{K}_f)}) \times q^\mu)$, which renders finding the best order quickly infeasible. Here, we are presenting the (E-)EC solution as a low-complexity solution for finding a *good* order. Thus, we opt out of optimizing the (E-)EC solution any further, and we simply order the edges (k,ℓ) within each color set according to the lexicographical order on $[f] \times [f]$.

The E-EC method is briefly summarized as Algorithm 1.

B. Longest-Distance First

The goal of LDF is to minimize dependency among the first selected monomials within an order. Thus, the intuition behind the LDF method also follows from graph matching. However, unlike in the (E-)EC method, we do not restrict ourselves to (near) perfect matchings of the complete graph $\mathcal{K}_f$. Here, we follow the convention that if two vertices belong to different connected components, then the distance is defined as infinite [13], i.e., there is no path connecting the two vertices. The LDF method is summarized with the following sequential steps (further details and a pseudo-code can be found in [15, App. B]).

Start with the null graph $\mathcal{G} = \mathcal{N}_f$, i.e., a graph with $\mathcal{V} = [f]$ and $\mathcal{E} = \emptyset$. Then, adhere to the following steps, adding an edge (u,v) to $\mathcal{G}$ and partial order $\mathcal{S}_{\text{LDF}}$ with each step repetition.

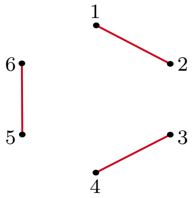
1) Repeatedly add an edge not adjacent to any other edge.
2) Repeatedly add an edge that connects two vertices with the longest distance and lowest degree in the graph, until a length-f cycle is formed.

```
Algorithm 1: Searching for a good order for $\overline{\mathrm{C}}$ based on edge-coloring (E-EC)
Input  : $f$, $q$, $n$
Output: A good order of edges $\mathcal{S}_{\text{E-EC}}$
$\mathcal{E}_1, \dots \mathcal{E}_{\chi'(\mathcal{K}_f)}$ ← color sets of edges with edges ordered
    lexicographically on $[f] \times [f]$
$\mathcal{E}_{\text{E-EC}} \leftarrow \mathcal{E}_1$, $i \leftarrow 1$
$\mathcal{E}_{\text{c}} \leftarrow$ a permutation of the remaining color sets of edges
$\mathcal{S}_{\text{E-EC}} \leftarrow (\mathcal{E}_{\text{E-EC}} \mid \mathcal{E}_{\text{c}})$
Compute $\mathrm{C}_{\text{E-EC–best}} = \overline{\mathrm{C}}(\mathcal{S}_{\text{E-EC}})$
while $i \le (\chi'(\mathcal{K}_f) - 1)!$ do
    $i \leftarrow i + 1$
    $\mathcal{E}_{\text{c}} \leftarrow$ next permutation of the remaining color sets of
      edges
    if $\overline{\mathrm{C}}(\mathcal{E}_{\text{E-EC}} \mid \mathcal{E}_{\text{c}}) < \mathrm{C}_{\text{E-EC–best}}$ then
        $\mathcal{S}_{\text{E-EC}} \leftarrow (\mathcal{E}_{\text{E-EC}} \mid \mathcal{E}_{\text{c}})$, $\mathrm{C}_{\text{E-EC-best}} \leftarrow \overline{\mathrm{C}}(\mathcal{S}_{\text{E-EC}})$
    end
end
return $\mathcal{S}_{\text{E-EC}}$
```

3) Repeatedly add an edge that connects two vertices with the lexicographically smallest numbers of induced length-l cycles for $3 \le l \le f$, until $\mathcal{G}$ is complete, i.e., $\mathcal{G} = \mathcal{K}_f$.

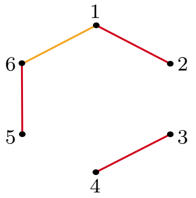
As for the (E-)EC method, we elaborate on the LDF algorithm with an illustrative example.

Example 2: For $f = 6$, first, add an edge that is not adjacent to any other edge. For example $(1, 2)$, then $(3, 4)$. As a result, one remaining edge can be added following Step 1), which is $(5, 6)$. Note that $\mathcal{E} = \{(1, 2), (3, 4), (5, 6)\}$ constitute a perfect matching of $\mathcal{K}_6$, adhering to Remark 2. The corresponding graph $\mathcal{G}$ and graph distances are illustrated in the following figure, where $v \in \mathcal{V}$ denotes the vertex label.



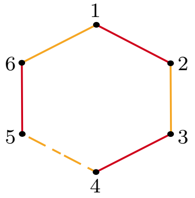
2	1				
3	∞	∞			
4	∞	∞	1		
5	∞	∞	∞	∞	
6	∞	∞	∞	∞	1
v	1	2	3	4	5

Next, to add an edge that connects two vertices with the longest distance and lowest degree in the graph, we select, for example, the edge $(1, 6)$ with $d(1, 6) = \infty$ and both vertices of degree 1. As a result, we have $d(2, 5) = 3$ and $d(2, 6) = d(1, 5) = 2$, as depicted in the following figure:



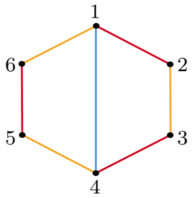
2	1				
3	∞	∞			
4	∞	∞	1		
5	2	3	∞	∞	
6	1	2	∞	∞	1
v	1	2	3	4	5

Next, to repeat Step 2), we can select from any available edge (u, v) with vertices of degree 1 and $d(u, v) = \infty$, i.e., $u, v \in \{2, 3, 4, 5\}$. For example, select the edge $(2, 3)$, then $(4, 5)$ forming a length-f cycle as illustrated in the left-hand side (l.h.s.) of the figure below where a dashed line indicates the order of adding. Now, for Step 3), we count the cycles induced from adding each of the remaining edges as seen in the following table, where $\boldsymbol{o} = (o_1, \dots, o_i, \dots, o_{f-2})$ and o_i is the number of cycles of length $i + 2$.



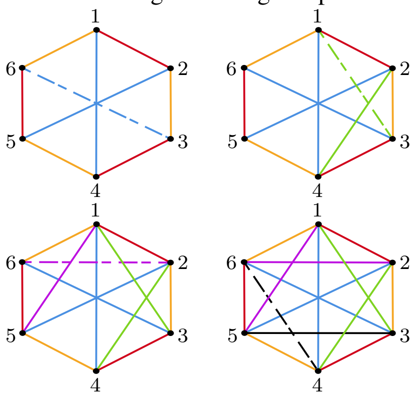
(u, v)	$\boldsymbol{o}$
$(1, 4), (2, 5), (3, 6)$	$(0, 2, 0, 1)$
$(1, 3), (1, 5), (2, 4)$ $(2, 6), (3, 5), (4, 6)$	$(1, 0, 1, 1)$

Each of the edges $(1, 4)$, $(2, 5)$, and $(3, 6)$ induces the smallest number of cycles, lexicographically, thus we select one of these edges. Let that edge be $(1, 4)$. The corresponding graph $\mathcal{G}$ is illustrated in the l.h.s of the following figure:



(u, v)	$\boldsymbol{o}$
$(2, 5), (3, 6)$	$(0, 5, 0, 2)$
$(2, 6), (3, 5)$	$(1, 2, 3, 1)$
$(1, 3), (1, 5)$ $(2, 4), (4, 6)$	$(2, 2, 1, 1)$

By repeating Step 3), the choices for the following edges remain the same. As can be seen from the above table, edges $(2, 5)$ and $(3, 6)$ induce the same number of cycles in $\mathcal{G}$. Thus, we select for example $(2, 5)$. The remaining edges follow from repeating Step 3) and are added to the order and $\mathcal{G}$ as illustrated in the following left-to-right top-to-bottom order:



At the end of the LDF procedure, we obtain the order

$$\mathcal{S}_{\text{LDF}} = \big((1, 2), (3, 4), (5, 6), (1, 6), (2, 3), (4, 5), (1, 4), (2, 5), (3, 6), (2, 4), (1, 3), (1, 5), (2, 6), (3, 5), (4, 6)\big),$$

and for $n = 2$ and $q = 2$, the capacity outer bound is $\overline{\mathrm{C}}(\mathcal{S}_{\text{LDF}}) = 0.5197824997350$, which is strictly better than for the E-EC method. Interestingly, $\mathcal{S}_{\text{LDF}}$ corresponds to a proper edge-coloring, i.e., no two adjacent edges share the same color, but it does not correspond to an optimal edge-coloring.

C. Entropy-Based Greedy Method

The EBG method starts with an empty graph and sequentially adds edges in a greedy manner, i.e., the edge that minimizes the (partial) bound in (1), computed based on the new edge and the already added edges, is added at each step in the algorithm. In case of ties, one of the candidate edges is

TABLE I
COMPARISON BETWEEN PQNMC CAPACITY OUTER BOUNDS OBTAINED WITH THE (E-)EC ($\overline{\mathsf{C}}(\mathcal{S}_{\text{(E-)EC}})$), LDF ($\overline{\mathsf{C}}(\mathcal{S}_{\text{LDF}})$), AND EBG ($\overline{\mathsf{C}}(\mathcal{S}_{\text{EBG}})$) METHODS, AS WELL AS WITH THE BEST BOUND FOUND BY EXHAUSTIVE/DIRECTED RANDOM SEARCH ($\overline{\mathsf{C}}(\mathcal{S}_{\text{ES/RS}})$), FOR $n = 2$ DATABASES, AND FOR A FIELD SIZE OF $q = 2$. THE BEST BOUND FOR EACH NUMBER OF MESSAGES f IS MARKED IN BOLD. THE BEST-KNOWN ACHIEVABLE RATE R FROM [10, THM. 2] IS GIVEN AS WELL TO SHOW THE CAP TO THE CAPACITY OUTER BOUND.

f	5	6	7	8	9	10	11	12
$\overline{\mathsf{C}}(\mathcal{S}_{\text{EC}})$	0.5382035621102	0.5198943946817	0.5158988408975	0.5088200966114	0.5071434701312	0.5041602427037	0.5033789063480	0.5020207578041
$\overline{\mathsf{C}}(\mathcal{S}_{\text{E-EC}})^5$	**0.5321513151313**	0.5198121367672	0.5130098344723	0.5085684044374	0.5058885273733	0.5039972538181	0.5028028499055	0.5019311781396
$\overline{\mathsf{C}}(\mathcal{S}_{\text{LDF}})$	**0.5321513151313**	**0.5197824997350**	**0.5129571653366**	0.5085546467521	0.5058724664437	0.5039960955809	**0.5027529132784**	0.5019069907637
$\overline{\mathsf{C}}(\mathcal{S}_{\text{EBG}})$	**0.5321513151313**	**0.5197824997350**	**0.5129571653366**	0.5085546463038	0.5058724626997	**0.5039958945996**	0.5027582097217	**0.5019068074415**
$\overline{\mathsf{C}}(\mathcal{S}_{\text{ES/RS}})$	**0.5321513151313**	**0.5197824997350**	**0.5129571653366**	**0.5085546398430**	**0.5058724411573**	0.5039961304091	0.5027529200313	0.5019070293099
R	0.5026676304668	0.5001371033940	0.5000032431709	0.5000000359051	0.5000000001891	0.5000000000005	0.5000000000000	0.5000000000000

selected at random. In particular, in the first step, an arbitrary edge is added to an initially empty graph. Then, in the second step the bound in (1) is computed with $\mu = 2$ based on the previously added edge and a new edge selected among the possible remaining edges. The new edge that minimizes the computed partial bound is then selected and added to the graph. In this manner, an edge is added to the graph in each step of the algorithm and a monomial order is constructed. The complexity of the EBG method is of order $\mathcal{O}(\mu(\mu+1)/2 \times q^{\mu})$. Finally, note that the order returned by the EBG method is independent of the value of $n \in \mathbb{N}$, including $n = 1$, while Remark 1 requires $n \geq 2$. Hence, $n = 1$ can be used for better numerical stability when conducting the EBG search.

IV. DISCUSSION AND RESULTS

In Table I, we compare the results from the proposed (E-)EC, LDF, and EBG methods for $5 \leq f \leq 12$ messages, for $n = 2$ databases, and for a field size of $q = 2$ with those of an exhaustive search (for $f = 5$) and a directed random search (for $6 \leq f \leq 12$). The directed random search is done by first fixing at least the first f edges according to edge-coloring (corresponding to two or three colors) and then conducting a random search among the remaining orders. As can be seen from the table, the LDF and EBG methods and the exhaustive/directed random search yield the same capacity outer bound for $f \leq 7$ messages, while for $f = 8$ and $f = 9$ messages a directed random search gives slightly better results (in the 8-th digit). (E-)EC gives the same bound as LDF for $f = 4$ messages, while for $f \geq 6$, E-EC performs worse compared to the LDF and EBG methods. The EC method performs in general slightly worse compared to the E-EC method, but has the lowest computational complexity. Interestingly, for $f = 11$, the LDF method outperforms all other methods. The best-known achievable rate R from [10, Thm. 2] is given in the last row of Table I to show the cap to the capacity outer bound. As a final remark, we note that for larger q (results not included here), the gap between the bounds produced by the LDF and EBG methods increases, which can be attributed to the fact that the proposed simple undirected graph model captures less of the dependencies for larger q.

V. CONCLUSION

We proposed two graph-based methods and one EBG algorithm to optimize the order of quadratic monomials in an outer bound for the PQNMC capacity. For $f < 6$ messages, all three methods minimize the bound, while for $6 \leq f \leq 12$ the results were compared with those of a directed random search. For almost all examined cases, the EBG algorithm yields the smallest gap to the best-found monomial ordering.

REFERENCES

[1] H. Sun and S. A. Jafar, “The capacity of private computation,” *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3880–3897, Jun. 2019.

[2] S. A. Obead and J. Kliewer, “Achievable rate of private function retrieval from MDS coded databases,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 17–22, 2018, pp. 2117–2121.

[3] S. A. Obead, H.-Y. Lin, E. Rosnes, and J. Kliewer, “Private linear computation for noncolluding coded databases,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 847–861, Mar. 2022.

[4] D. Karpuk, “Private computation of systematically encoded data with colluding servers,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 17–22, 2018, pp. 2112–2116.

[5] N. Raviv and D. A. Karpuk, “Private polynomial computation from Lagrange encoding,” *IEEE Trans. Inf. Forens. Secur.*, vol. 15, pp. 553–563, 2020.

[6] S. A. Obead, H.-Y. Lin, E. Rosnes, and J. Kliewer, “Private polynomial function computation for noncolluding coded databases,” *IEEE Trans. Inf. Forens. Secur.*, vol. 17, pp. 1800–1813, 2022.

[7] Y. Yakimenka, H.-Y. Lin, and E. Rosnes, “On the capacity of private monomial computation,” in *Proc. Int. Zurich Sem. Inf. Commun. (IZS)*, Zurich, Switzerland, Feb. 26–28, 2020, pp. 31–35.

[8] J. Zhu, Q. Yan, X. Tang, and S. Li, “Symmetric private polynomial computation from Lagrange encoding,” *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2704–2718, Apr. 2022.

[9] M. H. Mousavi, M. A. Maddah-Ali, and M. Mirmohseni, “Private inner product retrieval for distributed machine learning,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 355–359.

[10] S. A. Obead, H.-Y. Lin, E. Rosnes, and J. Kliewer, “On the capacity of private nonlinear computation for replicated databases,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 25–28, 2019.

[11] Z. Chen, Z. Wang, and S. A. Jafar, “The asymptotic capacity of private search,” *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4709–4721, Aug. 2020.

[12] N. Alon, “Combinatorial nullstellensatz,” *Combinatorics, Probability Comput.*, vol. 8, no. 1–2, pp. 7–29, Jan. 1999.

[13] J. A. Bondy and U. S. R. Murty, *Graph Theory*. London, U.K.: Springer, 2008.

[14] M. Behzad, G. Chartrand, and J. K. Cooper, Jr., “The colour numbers of complete graphs,” *J. London Math. Soc.*, vol. s1-42, no. 1, pp. 226–228, 1967.

[15] K. M. Dæhli, S. A. Obead, H.-Y. Lin, and E. Rosnes, “Improved capacity outer bound for private quadratic monomial computation,” Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.06125

[5] Due to high computational complexity, we fix two color sets of edges and then search over the remaining color permutations for $f = 11$ and $f = 12$.

Weakly-Private Information Retrieval From MDS-Coded Distributed Storage

Asbjørn O. Orvedal, Hsuan-Yin Lin, and Eirik Rosnes
Simula UiB, N–5006 Bergen, Norway
Emails: asbjorn.orvedal@gmail.com, {lin, eirikrosnes}@simula.no

***Abstract*—We consider the problem of weakly-private information retrieval (WPIR) when data is encoded by a maximum distance separable code and stored across multiple servers. In WPIR, a user wishes to retrieve a piece of data from a set of servers without leaking too much information about which piece of data she is interested in. We study and provide the first WPIR protocols for this scenario and present results on their optimal trade-off between download rate and information leakage using the maximal leakage privacy metric.**

I. Introduction

Private information retrieval (PIR), introduced in a seminal paper by Chor *et al.* [1], [2], has been extensively studied for more than two decades in both the computer science and information theory communities, see, e.g., [3]–[8] and references therein. In PIR, the objective is to download a piece of data stored on a set of servers without leaking any information about which piece of data is being requested to the servers storing the data, while minimizing the overall communication cost. As the upload cost is typically much lower than the download cost, the download rate, defined as the ratio between the amount of requested information and the amount of downloaded information, is used as a measure to compare different PIR protocols. When data is replicated across several servers, the maximum achievable download rate, referred to as the PIR capacity, was derived in [9], while the capacity for the case where the data is encoded by a maximum distance separable (MDS) code and stored across a set of servers was settled in [10]. Arbitrary linear storage codes were considered in [11], [12].

Weakly-private information retrieval (WPIR), introduced independently by Lin *et al.* [13] and Samy *et al.* [14], is a relaxed version of PIR that allows for reducing the download cost at the expense of some information leakage on the identity of the requested piece of data to the servers storing it. So far, only the case of replicated data (across servers) and the single server case have been considered in the literature [15]–[21], while in this work we consider for the first time the case where the data is encoded by an MDS code and stored across multiple servers. WPIR protocols allow for a trade-off between download rate and privacy leakage, and the optimal trade-off curve for the case of multiple servers is still an open problem. As in previous works, we consider the maximal leakage (MaxL) privacy metric [22]–[24]. Our main contributions are as follows.

- We adapt the PIR protocols in [25], [26] for MDS-coded databases to allow for information leakage. The adapted protocols from [25], [26], referred to as the ZYQT and ZTSL MDS-WPIR schemes, respectively, yield a trade-off between download rate and information leakage, and we show that for the MaxL privacy metric the optimal trade-off is the solution of a convex optimization problem (see Theorem 1). The optimized ZYQT MDS-WPIR scheme yields the best trade-off but also has the largest query space.
- We propose a *new* WPIR protocol, referred to as the OLR MDS-WPIR scheme, with a much smaller query space than the ZYQT scheme while providing an equally good or better trade-off between download rate and information leakage. As for the ZYQT and ZTSL MDS-WPIR schemes, the optimal trade-off is the solution of a convex optimization problem (see Theorem 1).

II. Preliminaries and System Model

A. Notation

We denote by $\mathbb{N}$ the set of all positive integers, and $[a:b] \triangleq \{a, a+1, \ldots, b\}$ for $a, b \in \{0\} \cup \mathbb{N}$, $a \leq b$. Vectors (normally row-wise) are denoted by bold letters, random variables (RVs) (either scalar or vector) by uppercase letters, and sets by calligraphic uppercase letters, e.g., $\boldsymbol{x}$, X, and $\mathcal{X}$, respectively. Matrices are denoted by sans serif letters, while random matrices are represented by bold sans serif capital letters, e.g., $\boldsymbol{\mathsf{X}}$, and x represents its realization. The all-one (all-zero) row vector is denoted by $\mathbf{1}$ ($\mathbf{0}$), and its length will be clear from the context. When a set of indices $\mathcal{S}$ is given, $\boldsymbol{X}_{\mathcal{S}}$ denotes $\{\boldsymbol{X}_s \colon s \in \mathcal{S}\}$. $\mathsf{E}_X[\cdot]$ denotes expectation with respect to the RV X. $X \sim P_X$ denotes an RV distributed according to a probability mass function (PMF) $P_X(x)$, $x \in \mathcal{X}$, and $X \sim \mathrm{U}(\mathcal{S})$ a uniformly-distributed RV over a set $\mathcal{S}$. $\mathsf{H}(\cdot)$ denotes the entropy function, $(\cdot)^\intercal$ the transpose of a matrix, and $\gcd(a, b)$ the greatest common divisor of two positive integers a and b.

B. System Model

We consider an MDS-coded distributed storage system (DSS) with N noncolluding servers that store M independent files $\boldsymbol{\mathsf{W}}^{(1)}, \ldots, \boldsymbol{\mathsf{W}}^{(\mathsf{M})}$, where each file is represented as a random matrix $\boldsymbol{\mathsf{W}}^{(m)} = \bigl(W_{i,j}^{(m)}\bigr)$ of size $\lambda \times \mathsf{K}$, $\lambda, \mathsf{K} \in \mathbb{N}$. Each file $\boldsymbol{\mathsf{W}}^{(m)}$ is encoded row-wise using an $[\mathsf{N}, \mathsf{K}]$ MDS code $\mathcal{C}$ over some finite field $\mathbb{F}_q$ of size $q \geq \mathsf{N}$ resulting in the codewords $\bigl(X_{i,1}^{(m)},\ldots,X_{i,\mathsf{N}}^{(m)}\bigr)=(W_{i,1}^{(m)},\ldots,W_{i,\mathsf{K}}^{(m)})\mathsf{G}^{\mathcal{C}}$, $i\in[0:\lambda-1]$, where $\mathsf{G}^{\mathcal{C}}$ denotes a generator matrix for $\mathcal{C}$. Denote by $\boldsymbol{X}_j^{(m)}\triangleq\bigl(X_{0,j}^{(m)},\ldots,X_{\lambda-1,j}^{(m)}\bigr)^{\mathsf{T}}$ a vector consisting of λ code symbols generated by the code $\mathcal{C}$. Then, the j-th server stores $\boldsymbol{X}_j\triangleq\bigl((\boldsymbol{X}_j^{(1)})^{\mathsf{T}}|\cdots|(\boldsymbol{X}_j^{(\mathsf{M})})^{\mathsf{T}}\bigr)^{\mathsf{T}}$, $j\in[1:\mathsf{N}]$.

To retrieve a file $\mathbf{W}^{(M)}$, from the MDS-coded DSS, the user sends a query $\mathbf{Q}_j$ to the j-th server for all $j\in[1:\mathsf{N}]$. Here, $M\sim\mathrm{U}([1:\mathsf{M}])$ is an RV representing the desired file index. In response to the received query, server j returns the answer $\mathbf{A}_j$, which is a function of $\mathbf{Q}_j$ and the code symbols $\boldsymbol{X}_j$ stored in the server, back to the user. We formally describe an MDS-coded $(\mathsf{M},\mathsf{N},\mathsf{K})$ WPIR scheme as follows.

Definition 1 (MDS-WPIR Scheme)**.** *An* $(\mathsf{M},\mathsf{N},\mathsf{K})$ *MDS-WPIR scheme for an* $[\mathsf{N},\mathsf{K}]$ *MDS-coded DSS with* N *non-colluding servers consists of:*

- M *independent files* $\mathbf{W}^{(m)}$ *of size* $\lambda\times\mathsf{K}$*, for some* $\lambda\in\mathbb{N}$*,* $m\in[1:\mathsf{M}]$*.*
- *A global random strategy* $\mathbf{S}$*, whose alphabet is* $\mathcal{S}$*. In general, the realization of* $\mathbf{S}$ *is a matrix.*
- *An* (N,K) *MDS storage code* $\mathcal{C}$ *that encodes the file* $\mathbf{W}^{(m)}$ *into the matrix* $\mathbf{X}^{(m)}=\bigl(\boldsymbol{X}_1^{(m)}|\cdots|\boldsymbol{X}_{\mathsf{N}}^{(m)}\bigr)$ *as described above,* $m\in[1:\mathsf{M}]$*.*
- N *queries* $\mathbf{Q}_j=\phi_j(M,\mathbf{S})$ *with alphabet* $\mathcal{Q}_j$*,* $j\in[1:\mathsf{N}]$*, that are generated by the query-encoding functions* ϕ_j*. Query* $\mathbf{Q}_j$ *is sent to the* j*-th server.*
- N *answers* $\mathbf{A}_j=\psi_j(\mathbf{Q}_j,\boldsymbol{X}_j)$ *with alphabet* $\mathcal{A}=\mathbb{F}_q$*,* $j\in[1:\mathsf{N}]$*, that are constructed by the answer functions* ψ_j*. All answers* $\mathbf{A}_j$ *are sent back to the user.*
- N *answer lengths* $\ell_j(\mathbf{Q}_j)\in\{0\}\cup\mathbb{N}$*,* $j\in[1:\mathsf{N}]$*, each being a function of the corresponding query* $\mathbf{Q}_j$*.*

In addition, the scheme should satisfy the following condition of perfect retrievability:

$$\mathsf{H}\bigl(\mathbf{W}^{(M)}\bigm|\mathbf{A}_{[1:\mathsf{N}]},\mathbf{Q}_{[1:\mathsf{N}]},M\bigr)=0.$$

C. Maximal Leakage Metric

From Definition 1, one can notice that at the j-th server, the requested file index M can be inferred by observing the query distribution $P_{\mathbf{Q}_j}$, which results in an information leakage on M to the servers. In this work, we adopt a meaningful information-theoretic privacy metric from the computer science literature, the MaxL metric, to measure information leakage. Formally, given the query distributions $P_{M,\mathbf{Q}_j}$, $j\in[1:\mathsf{N}]$, of a given $(\mathsf{M},\mathsf{N},\mathsf{K})$ WPIR scheme $\mathscr{C}$, the overall MaxL about M of $\mathscr{C}$ is defined as

$$\rho^{(\mathrm{MaxL})}(\mathscr{C})\triangleq\max_{j\in[1:\mathsf{N}]}\mathsf{MaxL}(M;\mathbf{Q}_j),$$

where

$$\mathsf{MaxL}(M;\mathbf{Q})\triangleq\log_2\left(\sum_{\mathsf{q}\in\mathcal{Q}}\max_{m\in[\mathsf{M}]}P_{\mathbf{Q}|M}(\mathsf{q}|m)\right).$$

Note that an $[\mathsf{N},\mathsf{K}]$ MDS-coded PIR scheme is an $(\mathsf{M},\mathsf{N},\mathsf{K})$ WPIR scheme $\mathscr{C}$ that satisfies $\rho^{(\mathrm{MaxL})}(\mathscr{C})=0$, such a condition is refereed to as the *perfect privacy* constraint.

D. WPIR Download Cost and Rate

The overall download cost (in number of symbols over $\mathbb{F}_q$) and rate of a WPIR scheme $\mathscr{C}$, denoted by $\mathsf{D}(\mathscr{C})$ and $\mathsf{R}(\mathscr{C})$, respectively, are given by

$$\mathsf{D}(\mathscr{C})=\sum_{j=1}^{\mathsf{N}}\mathsf{E}_{\mathbf{Q}_j}[\ell_j(\mathbf{Q}_j)]\text{ and }\mathsf{R}(\mathscr{C})\triangleq\frac{\lambda\mathsf{K}}{\mathsf{D}(\mathscr{C})}.$$

III. General MDS-WPIR Schemes

In this section, we give a general description of the $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR schemes we consider in this work. We start by reviewing two MDS-PIR capacity-achieving schemes for small file sizes, namely the ZYQT scheme [25] and the ZTSL scheme [26].[1]

A. The ZYQT Scheme and the ZTSL Scheme

1) Storage Data Structure: The following *effective code parameters* are universally defined for an MDS-coded DSS:

$$n\triangleq\frac{\mathsf{N}}{\gcd(\mathsf{N},\mathsf{K})},\quad k\triangleq\frac{\mathsf{K}}{\gcd(\mathsf{N},\mathsf{K})},\quad r\triangleq n-k.$$

Moreover, the subpacketization size for each file is given by $\lambda=n-k$. For ease of exposition, we further append k dummy variables $X_{i,j}^{(m)}\equiv 0$ for $i\in[n-k:n-1]$, $j\in[1:\mathsf{N}]$, such that for all $m\in[1:\mathsf{M}]$,

$$\mathbf{X}^{(m)}=\begin{pmatrix}X_{0,1}^{(m)} & X_{0,2}^{(m)}\cdots\cdots & X_{0,\mathsf{N}}^{(m)}\\ \vdots & \vdots\quad\ddots & \vdots\\ X_{n-k-1,1}^{(m)} & X_{n-k-1,2}^{(m)}\cdots & X_{n-k-1,\mathsf{N}}^{(m)}\\ 0 & 0\cdots\cdots\cdots & 0\\ \vdots & \vdots\quad\ddots & \vdots\\ 0 & 0\cdots\cdots\cdots & 0\end{pmatrix}\left.\begin{matrix}\\ \\ \\ \\ \\ \\ \end{matrix}\right\}k\text{ rows}. \tag{1}$$

2) Query Generation: The query generation is the main difference among the $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR schemes. In our context, we will make use of the set

$$\mathcal{P}_k^n\triangleq\bigl\{\boldsymbol{s}^{\mathsf{T}}=(s_1,\ldots,s_k)^{\mathsf{T}}\colon s_i,s_{i'}\in[0:n-1],\\ s_i\neq s_{i'},\,\forall\,i,i'\in[1:k],i\neq i'\bigr\}$$

of column vectors. The global random strategy alphabet for the ZYQT and ZTSL schemes are, respectively, given by

$$\mathcal{S}_{\mathrm{ZYQT}}\triangleq\{\mathsf{s}=(\boldsymbol{s}_1^{\mathsf{T}},\ldots,\boldsymbol{s}_{\mathsf{M}}^{\mathsf{T}})\colon\boldsymbol{s}_{m'}^{\mathsf{T}}\in\mathcal{P}_k^n,\;m'\in[1:\mathsf{M}]\},$$

$$\mathcal{S}_{\mathrm{ZTSL}}\triangleq\left\{\boldsymbol{s}\in[0:n-1]^{\mathsf{M}}\colon\left(\sum_{m'=1}^{\mathsf{M}}s_{m'}\right)\bmod n=0\right\}.$$

Note that $|\mathcal{S}_{\mathrm{ZYQT}}|=\bigl(\binom{n}{k}k!\bigr)^{\mathsf{M}}$ and $|\mathcal{S}_{\mathrm{ZTSL}}|=n^{\mathsf{M}-1}$. Since the cost of uploading the queries for an MDS-PIR scheme depends on the cardinality of the global random strategy alphabet, it is apparent that the ZTSL scheme has a lower upload cost than the ZYQT scheme. It is also worth mentioning that MDS-PIR schemes are generally constructed using an $\mathbf{S}$ that is uniformly distributed over the set $\mathcal{S}$.

[1] Precisely, the ZTSL scheme we consider here is the so-called Construction-A MDS-PIR code that is referred in [26, Sec. III].

We next present the original query generation for the ZYQT and ZTSL MDS-PIR schemes for retrieving the m-th file $\mathbf{X}^{(m)}$, $m \in [1:\mathsf{M}]$. Notice that we do not adopt the uniformly-distributed S here. Thus, the leakage $\rho^{(\mathrm{MaxL})}$ is not necessarily equal to 0. We refer to the corresponding proposed schemes as the ZYQT MDS-WPIR and ZTSL MDS-WPIR schemes and denote them by $\mathscr{C}_{\mathrm{ZYQT}}$ and $\mathscr{C}_{\mathrm{ZTSL}}$, respectively.

$\mathscr{C}_{\mathrm{ZYQT}}$: The query $\mathsf{q}_j \in \mathcal{Q}_j$, $j \in [1:\mathsf{N}]$, generated from the query-encoding function ϕ_j is defined as

$$\begin{aligned}\mathsf{q}_j = (&\boldsymbol{s}_1^{\mathsf{T}}, \ldots, \boldsymbol{s}_{m-1}^{\mathsf{T}}, \big(\boldsymbol{s}_m^{\mathsf{T}} + (j-1)\mathbf{1}^{\mathsf{T}}\big) \bmod n,\\ &\boldsymbol{s}_{m+1}^{\mathsf{T}}, \ldots, \boldsymbol{s}_{\mathsf{M}}^{\mathsf{T}}), \quad \boldsymbol{s}_m^{\mathsf{T}} \in \mathcal{P}_k^n,\, m \in [1:\mathsf{M}].\end{aligned}$$

$\mathscr{C}_{\mathrm{ZTSL}}$: The query $\mathsf{q}_j \in \mathcal{Q}_j$, $j \in [1:\mathsf{N}]$, is generated by

$$\begin{aligned}\mathsf{q}_j = \Bigg[&\left.\begin{pmatrix} s_1 \cdots s_{m-1} & (s_m + (j-1)) & s_{m+1} \cdots s_{\mathsf{M}} \\ \vdots \ddots \vdots & \vdots & \vdots \ddots \vdots \\ s_1 \cdots s_{m-1} & (s_m + (j-1)) & s_{m+1} \cdots s_{\mathsf{M}} \end{pmatrix}\right\} k \text{ rows} \\ &+ \underbrace{\begin{pmatrix} 0 & 0 \cdots\cdots 0 \\ 1 & 1 \cdots\cdots 1 \\ \vdots & \vdots \ \ddots \ \vdots \\ k-1 & k-1 \cdots k-1 \end{pmatrix}}_{\mathsf{M} \text{ columns}}\Bigg] \bmod n,\end{aligned}$$

where $\boldsymbol{s} \in \mathcal{S}_{\mathrm{ZTSL}}$.

3) Answer Construction: Upon receiving a query (matrix)

$$\mathsf{q}_j = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,\mathsf{M}} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k,1} & q_{k,2} & \cdots & q_{k,\mathsf{M}} \end{pmatrix},$$

the j-th server uses the answer function ψ_j to construct the answer

$$\mathbf{A}_j = \psi_j(\mathsf{q}_j, \boldsymbol{X}_j) = \left(\sum_{m'=1}^{\mathsf{M}} X^{(m')}_{q_{1,m'},j}, \cdots, \sum_{m'=1}^{\mathsf{M}} X^{(m')}_{q_{k,m'},j}\right)^{\mathsf{T}}$$

consisting of k sub-responses. With the storage data defined in (1), the length of the answer is given by the number of nonzero components in $\mathbf{A}_j$, which is equal to

$$\ell_j(\mathsf{q}_j) = \sum_{i=1}^{k} \mathbb{1}\left\{\min_{m' \in [1:\mathsf{M}]} q_{i,m'} \leq n-k-1\right\},$$

where $\mathbb{1}\{\text{statement}\}$ is the indicator function whose value is 1 if the statement is true and 0 otherwise.

Finally, we remark that according to the query constructions for both the ZYQT and ZTSL MDS-WPIR schemes, the file $\mathbf{W}^{(m)}$ can always be reconstructed by the MDS property of the storage code $\mathcal{C}$ (the so-called K-out-of-N property).

B. Time-Sharing MDS-WPIR Scheme

Clearly, selecting a different global random strategy S leads to a different WPIR rate and privacy leakage of an MDS-WPIR scheme. This work aims to achieve the best trade-off between download rate and privacy leakage by using the best S for an MDS-WPIR scheme. However, the minimization problem of the information leakage for a given WPIR rate over the global random strategy for an MDS-WPIR scheme is generally not convex. Hence, in order to easily tackle the optimization problem, we make use of a time-sharing principle to *convexify* the optimization problem for determining the best rate-leakage trade-off [16, Sec. VII].

Definition 2 (Time-Sharing MDS-WPIR Scheme)**.** *Consider an MDS-WPIR scheme $\mathring{\mathscr{C}}$ with query-encoding functions $\mathring{\phi}_j$, answer functions $\mathring{\psi}_j$, and a global random strategy $\mathring{\mathsf{S}}$. The time-sharing MDS-WPIR scheme of $\mathring{\mathscr{C}}$ is made by the query-encoding functions $\phi_j = \mathring{\phi}_{\sigma^{T-1}(j)}(M, \mathsf{S})$ and the answer functions $\psi_j = \mathring{\psi}_{\sigma^{T-1}(j)}\big(\mathring{\phi}_{\sigma^{T-1}(j)}(M, \mathsf{S}), \boldsymbol{X}_j\big)$, $j \in [1:\mathsf{N}]$, for a given requested file index M, where $T \sim \mathrm{U}([1:\mathsf{N}])$, and $\sigma(\cdot)$ denotes a left circular shift, while l left circular shifts are obtained through function composition and denoted by $\sigma^l(\cdot)$. Such an MDS-WPIR scheme $\mathscr{C}$ is called the time-sharing scheme of $\mathring{\mathscr{C}}$.*

Remark 1.

- *A time-sharing MDS-WPIR scheme always has equal information leakage at each server [16, Th. 1].*
- *In the following, unless specified otherwise, all the MDS-WPIR schemes we discuss are assumed to be already post-processed by applying the time-sharing principle, and the minimization of MaxL is also done for the time-sharing scheme of an MDS-WPIR scheme.*

C. Minimization of MaxL for MDS-WPIR Schemes

Denote by $z_{\mathsf{s}} \triangleq P_{\mathsf{S}}(\mathsf{s})$ the PMF of the random strategy S. It can be shown that both the MaxL $\rho^{(\mathrm{MaxL})}(\mathscr{C})$ and the WPIR download cost $\mathsf{D}(\mathscr{C})$ of a given MDS-WPIR scheme $\mathscr{C}$ can be expressed in terms of z_{s}, $\mathsf{s} \in \mathcal{S}$. Thus, the minimization of $\rho^{(\mathrm{MaxL})}(\mathscr{C})$ under a download cost constraint $\mathsf{D}(\mathscr{C}) \leq \mathsf{D}$ can be re-written in terms of the variables $\{z_{\mathsf{s}}\}_{\mathsf{s} \in \mathcal{S}}$ as the optimization problem

$$\begin{aligned} \text{minimize} \quad & \rho^{(\mathrm{MaxL})}(\{z_{\mathsf{s}}\}_{\mathsf{s} \in \mathcal{S}}) && \text{(2a)}\\ \text{subject to} \quad & \mathsf{D}(\{z_{\mathsf{s}}\}_{\mathsf{s} \in \mathcal{S}}) \leq \mathsf{D}, && \text{(2b)}\\ & \sum_{\mathsf{s} \in \mathcal{S}} z_{\mathsf{s}} = 1. && \text{(2c)} \end{aligned}$$

The following theorem can be proved using a similar argument as in [16, Sec. VII].

Theorem 1. *The optimization problem* (2) *is convex.*

All the rate-leakage trade-off curves of the MDS-WPIR schemes we study in this work are based on solving the convex optimization problem above.

IV. New Proposed MDS-WPIR Scheme

This section presents a new MDS-WPIR scheme, referred to as the OLR MDS-WPIR scheme. We first present an example illustrating the motivation for studying the new MDS-WPIR scheme in Section IV-A. In particular, we will show that the ZTSL MDS-WPIR scheme is naturally not a good scheme as it is not functional in the high-rate region when there is leakage.

A. *Motivating Example:* $(\mathsf{M},\mathsf{N},\mathsf{K})=(2,3,2)$

For $(\mathsf{N},\mathsf{K})=(3,2)$, we have the effective code parameters

$$n=\frac{\mathsf{N}}{\gcd(\mathsf{N},\mathsf{K})}=3,\ k=\frac{\mathsf{K}}{\gcd(\mathsf{N},\mathsf{K})}=2,\ r=n-k=1,$$

and the subpacketization size for each file is $\lambda=n-k=1$.

For the $(2,3,2)$ ZTSL MDS-WPIR scheme, we have $\mathcal{S}_{\mathrm{ZTSL}}=\{(0,0),(1,2),(2,1)\}$, and the corresponding conditional query PMF $P_{\mathbf{Q}_j|M}(\mathsf{q}_j\mid m)$ and answer lengths are as follows:

$$\begin{array}{r}P_{\mathbf{Q}_j|M}(\mathsf{q}_j\mid m)\\ m\left\{\begin{array}{l}1\\2\end{array}\right.\\ P_{\mathbf{Q}_j}(\mathsf{q}_j)\\ \ell_j(\mathsf{q}_j)\end{array}
\begin{array}{c}\begin{array}{ccccccccc}\left(\begin{smallmatrix}0&0\\1&1\end{smallmatrix}\right)&\left(\begin{smallmatrix}1&2\\2&0\end{smallmatrix}\right)&\left(\begin{smallmatrix}2&1\\0&2\end{smallmatrix}\right)&\left(\begin{smallmatrix}1&0\\2&1\end{smallmatrix}\right)&\left(\begin{smallmatrix}2&2\\0&0\end{smallmatrix}\right)&\left(\begin{smallmatrix}0&1\\1&2\end{smallmatrix}\right)&\left(\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right)&\left(\begin{smallmatrix}0&2\\1&0\end{smallmatrix}\right)&\left(\begin{smallmatrix}1&1\\2&2\end{smallmatrix}\right)\end{array}\\
\left(\begin{array}{ccccccccc}\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}\\ \frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_2}{3}\\ \hline \frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1+z_2}{6}&\frac{z_2+z_3}{6}&\frac{z_1+z_3}{6}&\frac{z_1+z_3}{6}&\frac{z_1+z_2}{6}&\frac{z_2+z_3}{6}\\ \hdashline 1&1&1&1&1&1&2&2&0\end{array}\right),\end{array}\tag{3}$$

where $z_j\triangleq\Pr(\boldsymbol{s}_j)$ for $\boldsymbol{s}_j=(j-1,(n-j+1)\bmod n)\in\mathcal{S}_{\mathrm{ZTSL}}$, $j\in[1:n]$. A simple calculation gives

$$\mathsf{D}(\mathscr{C}_{\mathrm{ZTSL}})=3+z_1,\quad 0\leq z_1\leq 1,$$

which indicates that $\mathsf{D}(\mathscr{C}_{\mathrm{ZTSL}})$ can only range between 3 and 4, and never reaches $\mathsf{R}={}^{\lambda\mathsf{K}}/_{\mathsf{D}}={}^{2}/_{\mathsf{D}}=1$. Thus, the ZTSL MDS-WPIR scheme can not operate in the high-rate region.

B. *New* $(\mathsf{M},\mathsf{N},\mathsf{K})$ *MDS-WPIR Scheme*

We now describe the new proposed $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR scheme, referred to as the OLR MDS-WPIR scheme and denoted by $\mathscr{C}_{\mathrm{OLR}}$. Here, only the query generation is presented, as its answer construction is the same as Section III-A3.

1) Query Generation: The strategy set for our new MDS-WPIR scheme is defined as

$$\mathcal{S}_{\mathrm{OLR}}\triangleq\Biggl\{\mathsf{s}=(\boldsymbol{s}_1^{\mathsf{T}},\ldots,\boldsymbol{s}_{\mathsf{M}-1}^{\mathsf{T}})\colon\boldsymbol{s}_{m'}^{\mathsf{T}}\in\mathcal{P}_k^n,\ m'\in[1:\mathsf{M}],$$
$$\left(\sum_{m'=1}^{\mathsf{M}}\boldsymbol{s}_{m'}^{\mathsf{T}}\right)\bmod n=\mathbf{0}^{\mathsf{T}}\Biggr\}.$$

By definition, $|\mathcal{S}_{\mathrm{OLR}}|\leq\left(\binom{n}{k}k!\right)^{\mathsf{M}-1}<|\mathcal{S}_{\mathrm{ZYQT}}|=\left(\binom{n}{k}k!\right)^{\mathsf{M}}$, as we do not include all the possible vectors $\boldsymbol{s}_{m'}^{\mathsf{T}}\in\mathcal{P}_k^n$.

The query $\mathsf{q}_j\in\mathcal{Q}_j$, $j\in[1:\mathsf{N}]$, for retrieving the m-th file, $m\in[1:\mathsf{M}]$, is defined as

$$\mathsf{q}_j=(\boldsymbol{s}_1^{\mathsf{T}},\ldots,\boldsymbol{s}_{m-1}^{\mathsf{T}},\boldsymbol{q}_m^{\mathsf{T}},\boldsymbol{s}_m^{\mathsf{T}},\ldots,\boldsymbol{s}_{\mathsf{M}-1}^{\mathsf{T}}),\tag{4}$$

where $(\boldsymbol{s}_1^{\mathsf{T}},\ldots,\boldsymbol{s}_{\mathsf{M}-1}^{\mathsf{T}})=\mathsf{s}\in\mathcal{S}_{\mathrm{OLR}}$ and

$$\boldsymbol{q}_m^{\mathsf{T}}\triangleq\left((j-1)\mathbf{1}^{\mathsf{T}}-\sum_{m'\in[1:\mathsf{M}-1]}\boldsymbol{s}_{m'}^{\mathsf{T}}\right)\bmod n.$$

Example 1. *Consider the same code parameters* $(\mathsf{M},\mathsf{N},\mathsf{K})=(2,3,2)$ *as in Section IV-A. We consider the strategy set*

$$\mathcal{S}_{\mathrm{OLR}}=\{\underbrace{\left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)}_{z_1},\underbrace{\left(\begin{smallmatrix}0\\2\end{smallmatrix}\right)}_{z_2},\underbrace{\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)}_{z_3},\underbrace{\left(\begin{smallmatrix}1\\2\end{smallmatrix}\right)}_{z_4},\underbrace{\left(\begin{smallmatrix}2\\0\end{smallmatrix}\right)}_{z_5},\underbrace{\left(\begin{smallmatrix}2\\1\end{smallmatrix}\right)}_{z_6}\}.$$

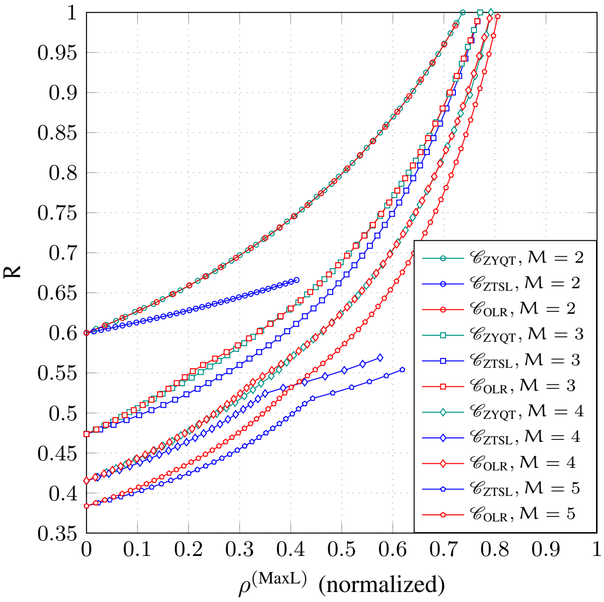


Fig. 1. Rate-leakage trade-off curve for the proposed MDS-WPIR protocols from $(3,2)$ MDS-coded storage with $\mathsf{M}=2$ (circle markers), $\mathsf{M}=3$ (square markers), $\mathsf{M}=4$ (diamond markers), and $\mathsf{M}=5$ (pentagon markers).

Similar to (3)*, we illustrate* 9 *out of the* 18 *query matrices based on* (4) *and the corresponding query distributions and answer lengths of the OLR MDS-WPIR scheme below:*

$$\begin{array}{r}P_{\mathbf{Q}_j|M}(\mathsf{q}_j\mid m)\\ m\left\{\begin{array}{l}1\\2\end{array}\right.\\ P_{\mathbf{Q}_j}(\mathsf{q}_j)\\ \ell_j(\mathsf{q}_j)\end{array}
\begin{array}{c}\begin{array}{ccccccccc}\left(\begin{smallmatrix}0&0\\2&1\end{smallmatrix}\right)&\left(\begin{smallmatrix}0&0\\1&2\end{smallmatrix}\right)&\left(\begin{smallmatrix}2&1\\0&0\end{smallmatrix}\right)&\left(\begin{smallmatrix}2&1\\1&2\end{smallmatrix}\right)&\left(\begin{smallmatrix}1&2\\0&0\end{smallmatrix}\right)&\left(\begin{smallmatrix}1&2\\2&1\end{smallmatrix}\right)&\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)&\left(\begin{smallmatrix}1&0\\2&2\end{smallmatrix}\right)&\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)\end{array}\\
\left(\begin{array}{ccccccccc}\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_4}{3}&\frac{z_5}{3}&\frac{z_6}{3}&\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}\\ \frac{z_2}{3}&\frac{z_1}{3}&\frac{z_5}{3}&\frac{z_6}{3}&\frac{z_3}{3}&\frac{z_4}{3}&\frac{z_3}{3}&\frac{z_4}{3}&\frac{z_1}{3}\\ \hline \frac{z_1+z_2}{3}&\frac{z_1+z_2}{6}&\frac{z_3+z_5}{3}&\frac{z_4+z_6}{6}&\frac{z_3+z_5}{6}&\frac{z_4+z_6}{6}&\frac{z_1+z_3}{6}&\frac{z_2+z_4}{6}&\frac{z_1+z_3}{6}\\ \hdashline 1&1&1&0&1&0&2&1&2\end{array}\right).\end{array}$$

As a result, one can compute the download cost $\mathsf{D}(\mathscr{C}_{\mathrm{OLR}})$ *and obtain*

$$\mathsf{D}(\mathscr{C}_{\mathrm{OLR}})=2+2(z_1+z_2+z_3+z_5)\geq 2,$$

which shows that $\mathsf{R}(\mathscr{C}_{\mathrm{OLR}})$ *can reach* ${}^{(n-k)\mathsf{K}}/_{2}=1$*, demonstrating a complete rate-leakage trade-off for the new MDS-WPIR scheme.*

V. Numerical Results

Here, we compare the optimal rate-leakage trade-off curves for our three proposed MDS-WPIR schemes $\mathscr{C}_{\mathrm{ZYQT}}$, $\mathscr{C}_{\mathrm{ZTSL}}$, and $\mathscr{C}_{\mathrm{OLR}}$. The optimal trade-off curve is obtained by solving the corresponding convex optimization problems as outlined in (2). For the sake of presentation, the leakage is normalized by $\log_2\mathsf{M}$ bits so that its range is from 0 to 1.

In Fig. 1, we consider the case of $\mathsf{N}=3$ servers and $\mathsf{K}=2$, and with different number of files M. As can be seen from the figure by comparing the green and the blue curves, $\mathscr{C}_{\mathrm{ZYQT}}$ gives a better rate-leakage trade-off curve than $\mathscr{C}_{\mathrm{ZTSL}}$ for all considered values of M. Moreover, the ZTSL scheme cannot be extended to a high information leakage. On the other, the OLR scheme performs equally well as the ZYQT scheme

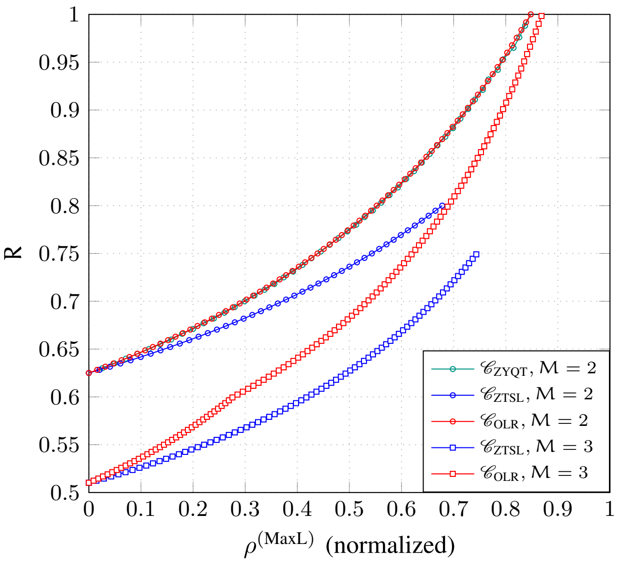


Fig. 2. Rate-leakage trade-off curve for the proposed MDS-WPIR protocols from $(5,3)$ MDS-coded storage with $\mathsf{M} = 2$ (circle markers) and $\mathsf{M} = 3$ (square markers).

for $\mathsf{M} = 2$ files and slightly better for a certain range of information leakage for $\mathsf{M} = 3$ and $\mathsf{M} = 4$ files, while at the same time allowing for a much smaller query space.

The corresponding rate-leakage trade-off curves for $\mathsf{N} = 5$ servers with $\mathsf{K} = 3$ are provided in Fig. 2. The same observations as in Fig. 1 can be made, i.e., the ZYQT scheme outperforms the ZTSL scheme, while the proposed OLR scheme yields an equal trade-off curve as the ZYQT scheme for $\mathsf{M} = 2$ files. As the query space is significant for the ZYQT scheme for $\mathsf{M} = 3$ files, we were not able to solve the corresponding convex optimization problem as outlined in (2) and therefore no curve for $\mathsf{M} > 2$ is presented. However, as mentioned previously, a nice feature of the OLR scheme is its smaller query space, and hence the corresponding optimization problem in (2) can be readily solved even for $\mathsf{M} = 3$. In particular, we have $|\mathcal{S}_{\mathrm{ZYQT}}| = 216000 > |\mathcal{S}_{\mathrm{OLR}}| = 1500$ for $\mathsf{M} = 3$.

VI. Conclusion

This work is the first to consider WPIR for coded storage. In particular, we proposed and compared three WPIR protocols for the case where the data is encoded by an MDS code and stored across multiple servers. Allowing for some leakage on the identity of the requested file index allows for a higher download rate, and we showed that the optimal trade-off of download rate and information leakage using the MaxL privacy metric is the solution to a convex optimization problem for all three proposed protocols.

References

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, “Private information retrieval,” in *Proc. 36th Annu. IEEE Symp. Found. Comp. Sci. (FOCS)*, Milwaukee, WI, USA, Oct. 23–25, 1995, pp. 41–50.

[2] ——, “Private information retrieval,” *J. ACM*, vol. 45, no. 6, pp. 965–982, Nov. 1998.

[3] A. Beimel, Y. Ishai, E. Kushilevitz, and J.-F. Raymond, “Breaking the $O(n^{1/(2k-1)})$ barrier for information-theoretic private information retrieval,” in *Proc. 43rd Annu. IEEE Symp. Found. Comp. Sci. (FOCS)*, Vancouver, BC, Canada, Nov. 16–19, 2002, pp. 261–270.

[4] S. Yekhanin, “Private information retrieval,” *Commun. ACM*, vol. 53, no. 4, pp. 68–73, Apr. 2010.

[5] H. Corrigan-Gibbs and D. Kogan, “Private information retrieval with sublinear online time,” in *Proc. 39th Annu. Int. Conf. Theory Appl. Crypto. Techn. (EUROCRYPT)*, Zagreb, Croatia, May 10–14, 2020, pp. 44–75.

[6] T. H. Chan, S.-W. Ho, and H. Yamamoto, “Private information retrieval for coded storage,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 14–19, 2015, pp. 2842–2846.

[7] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, “Private information retrieval from coded databases with colluding servers,” *SIAM J. Appl. Algebra Geom.*, vol. 1, no. 1, pp. 647–664, Nov. 2017.

[8] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, “Private information retrieval from MDS coded data in distributed storage systems,” *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[9] H. Sun and S. A. Jafar, “The capacity of private information retrieval,” *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.

[10] K. Banawan and S. Ulukus, “The capacity of private information retrieval from coded databases,” *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

[11] S. Kumar, H.-Y. Lin, E. Rosnes, and A. Graell i Amat, “Achieving maximum distance separable private information retrieval capacity with linear codes,” *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4243–4273, Jul. 2019.

[12] H.-Y. Lin, S. Kumar, E. Rosnes, and A. Graell i Amat, “Asymmetry helps: Improved private information retrieval protocols for distributed storage,” in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, Nov. 25–29, 2018.

[13] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, “Weakly-private information retrieval,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1257–1261.

[14] I. Samy, R. Tandon, and L. Lazos, “On the capacity of leaky private information retrieval,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1262–1266.

[15] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, “The capacity of single-server weakly-private information retrieval,” *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 415–427, Mar. 2021.

[16] ——, “Multi-server weakly-private information retrieval,” *IEEE Trans. Inf. Theory*, vol. 68, no. 2, pp. 1197–1219, Feb. 2022.

[17] C. Qian, R. Zhou, C. Tian, and T. Liu, “Improved weakly private information retrieval codes,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Espoo, Finland, Jun. 26–Jul. 1, 2022, pp. 2840–2845.

[18] I. Samy, M. Attia, R. Tandon, and L. Lazos, “Asymmetric leaky private information retrieval,” *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 5352–5369, Aug. 2021.

[19] R. Zhou, T. Guo, and C. Tian, “Weakly private information retrieval under the maximal leakage metric,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, CA, USA, Jun. 21–26, 2020, pp. 1089–1094.

[20] Y. Yakimenka, H.-Y. Lin, E. Rosnes, and J. Kliewer, “Optimal rate-distortion-leakage tradeoff for single-server information retrieval,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 832–846, Mar. 2022.

[21] C.-W. Weng, Y. Yakimenka, H.-Y. Lin, E. Rosnes, and J. Kliewer, “Generative adversarial user privacy in lossy single-server information retrieval,” *IEEE Trans. Inf. Forens. Secur.*, vol. 17, pp. 3495–3510, 2022.

[22] G. Smith, “On the foundations of quantitative information flow,” in *Proc. 12th Int. Conf. Found. Softw. Sci. Comput. Struct. (FoSSaCS)*, York, U.K., Mar. 22–29, 2009, pp. 288–302.

[23] G. Barthe and B. Köpf, “Information-theoretic bounds for differentially private mechanisms,” in *Proc. 24th IEEE Comput. Secur. Found. Symp. (CSF)*, Cernay-la-Ville, France, Jun. 27–29, 2011, pp. 191–204.

[24] I. Issa, A. B. Wagner, and S. Kamath, “An operational approach to information leakage,” *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[25] J. Zhu, Q. Yan, C. Qi, and X. Tang, “A new capacity-achieving private information retrieval scheme with (almost) optimal file length for coded servers,” *IEEE Trans. Inf. Forens. Secur.*, vol. 15, pp. 1248–1260, 2020.

[26] R. Zhou, C. Tian, H. Sun, and T. Liu, “Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size,” *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4904–4916, Aug. 2020.

Strong Converse for Bi-Static ISAC with Two Detection-Error Exponents

Mehrasa Ahmadipour*, Michèle Wigger†, Shlomo Shamai‡
*UMPA, ENS de Lyon mehrasa.ahmadipour@ens-lyon.fr
†LTCI Telecom Paris, IP Paris, France, Email: michele.wigger@telecom-paris.fr
‡Technion—Israel Institue of Technology, Israel, Email: sshlomo@ee.technion.ac.il

***Abstract*—The paper considers an information-theoretic bi-static integrated sensing and communication (ISAC) model and provides new results on the fundamental limits of this model. Specifically, we show a strong converse result for memoryless ISAC where the channel transition law depends on an underlying binary hypothesis, which the radar receiver wishes to determine with largest exponential decay rates of the probabilities of detection error under the two hypotheses. In this sense, the channel is a compound channel, and we establish a strong converse under maximum probability of error criteria. We further prove that the fundamental limits of our ISAC system remain unchanged under average probability of error criteria as long as the admissible channel decoding errors are bounded by $1/2$.**

***Index Terms*—Integrated sensing and communication, bi-static radar, detection error exponent.**

I. Introduction

Huge technological efforts are being made to integrate radar systems with communication systems, in particular in view of the future 6G mobile communication standard [1], [2] and its deployment for autonomous navigation or smart manufacturing sites. In *integrated sensing and communication systems (ISAC)*, the idea is to use a common waveform for both tasks: the emitted signals are modulated so as to achieve reliable data communication while the backscatters of these signals are used to sense the environment, detect hazardous events, or infer properties of other terminals (e.g., velocities or directions of other cars).

ISAC has already inspired a plethora of works in the signal processing and communications communities, see for example [3]–[10] and references therein, as well as (to a lesser extent) in the information-theoretic community [11]–[21]. The results in [17]–[20] and the present manuscript all focus on the system model in Figure 1 consisting of a transmitter (Tx) sending a message to a receiver (Rx) over a state-dependent discrete memoryless channel (SDMC). A bi-static radar receiver close to the Tx receives the backscattered signal, and due to the proximity to the Tx, this radar receiver also knows the Tx's channel inputs and compares them to the backscatterers.

We follow the model in [17]–[20], where the channel is memoryless and stationary with a transition law that depends on a binary hypothesis. The goal of the radar receiver is to detect the underlying hypothesis. In a real-world application, the hypothesis can correspond to the presence or absence of an obstacle, which the radar receiver wishes to determine. As in [18], we measure sensing performance in terms of the two exponential decay-rates of the detection error probabilities under the two hypotheses. The difference between [18] and the present work is that we allow for positive probabilities of decoding error in the communication and determine the fundamental limits of ISAC under this assumption. In particular, we prove a strong converse by showing that even when one allows for positive decoding error probabilities neither the achievable rate nor the achievable sensing performance can be improved. Our results are based on maximum error-probability criteria. In fact, since the communication channel is a compound channel (on a set of two channels) it is known that a geneal strong converse fails under average-probability of error criteria even for only the channel coding part [22], [23].

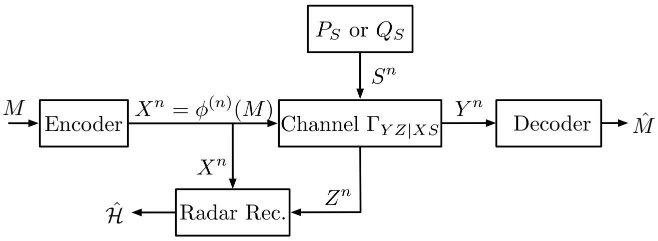


Fig. 1: Bi-static Radar ISAC Model

The converse proof in this paper is an extension of the channel coding strong converse proof in [24] to incorporate also the sensing bounds, see also [20] for a strong converse for bi-static ISAC when sensing performance is measured in terms of Stein's exponent or distortion. Strong converse proofs based on change-of-measure arguments go back to Gu and Effros [25], [26] and can be also found in various other works, e.g., [27]. The proof method was formalized and first applied to channel coding by Tyagi and Watanabe [28].

Notation: Upper-case letters are used for random quantities and lower-case letters for deterministic realizations. Calligraphic font is used for sets. All random variables are assumed finite and discrete. We abbreviate the n-tuples $(X_1, \ldots, X_n)$ and $(x_1, \ldots, x_n)$ as X^n and x^n. We further abbreviate *independent and identically distributed* as *i.i.d.* and *probability mass function* as *pmf*. Pmfs of i.i.d. random tuples are denoted by $P^{\otimes}$.

Entropy, conditional entropy, and mutual information functionals are written as $H(\cdot)$, $H(\cdot|\cdot)$, and $I(\cdot;\cdot)$, potentially with pmfs in the subscripts. The Kullback-Leibler divergence is denoted by $D(\cdot\|\cdot)$. We denote by $\boldsymbol{\pi}_{x^n}$ the type of a sequence x^n. We also use Landau notation, where $o(1)$ denotes a function that tends to 0 as $n \to \infty$.

II. Setup and Main Results

Consider the bi-static radar receiver model over a memoryless channel in Fig. 1. A transmitter (Tx) that wishes to communicate a random message M to a receiver (Rx) over a state-dependent channel. The message M is uniformly distributed over the set $\mathcal{M} = \{1, \ldots, 2^{nR}\}$ with $R > 0$ and $n > 0$ denoting the rate and blocklength of communication, respectively. The channel from the Tx to the Rx depends on a state-sequence $S^n = (S_1, \ldots, S_n)$ which depends on a binary hypothesis $\mathcal{H} \in \{0, 1\}$. Under the null hypothesis $\mathcal{H} = 0$ it is i.i.d. according to the pmf P_S and under the alternative hypothesis $\mathcal{H} = 1$ it is i.i.d. according to the pmf Q_S. For a given blocklength n, the Tx thus produces the n-length sequence of channel inputs

$$X^n = \phi^{(n)}(M) \tag{1}$$

for some choice of the encoding function $\phi^{(n)} \colon \{1, \ldots, 2^{nR}\} \to \mathcal{X}^n$.

Based on X^n and S^n the channel produces the sequences Y^n observed at the Rx and the backscattered signal Z^n. The channel is assumed memoryless and described by the stationary transition law $\Gamma_{YZ|SX}$ implying that the pair (Y_t, Z_t) is produced according to the channel law $\Gamma_{YZ|SX}$ based on the time-t symbols (X_t, S_t).

The Rx attempts to guess message M based on the sequence of channel outputs Y^n:

$$\hat{M} = g^{(n)}(Y^n) \tag{2}$$

using a decoding function of the form $g^{(n)} \colon \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$.

Performance of communication is measured in terms of maximum error probability

$$p^{(n)}(\text{error}) := \max_{\mathsf{H} \in \{0,1\}} \max_{m \in \{1,\ldots,2^{nR}\}} \Pr\left[\hat{M} \neq M | \mathcal{H} = \mathsf{H}, M = m\right]. \tag{3}$$

We assume a radar receiver close to the Tx, that wishes to guess the underlying hypothesis based on the inputs and backscattered signals. I.e., it produces a guess of the form

$$\hat{\mathcal{H}} = h^{(n)}(X^n, Z^n) \in \{0, 1\}. \tag{4}$$

Radar sensing performance is measured in terms of error-exponent pairs. That means, it is required that the type-I and type-II error probabilities

$$\alpha_n := \max_{m \in \mathcal{M}} \Pr\left[\hat{\mathcal{H}} = 1 \middle| \mathcal{H} = 0, M = m\right] \tag{5}$$

and

$$\beta_n := \max_{m \in \mathcal{M}} \Pr\left[\hat{\mathcal{H}} = 0 \middle| \mathcal{H} = 1, M = m\right] \tag{6}$$

decay exponentially fast to 0 with largest possible exponents.

Definition 1: A rate-exponent pair (R, θ) is (ϵ, r)-achievable over the state-dependent DMC $(\mathcal{X}, \mathcal{Y}, \Gamma_{YZ|XS})$ with state-distributions P_S and Q_S, if there exists a sequence of encoding, decoding, and estimation functions $\{(\phi^{(n)}, g^{(n)}, h^{(n)})\}$ such that for each blocklength n the maximum probability of error (over the two hypotheses) satisfies

$$\overline{\lim_{n\to\infty}}\, p^{(n)}(\text{error}) \leq \epsilon, \tag{7}$$

while the detection error probabilities satisfy:

$$-\underline{\lim_{n\to\infty}}\, \frac{1}{n} \log \alpha_n \geq r, \tag{8}$$

$$-\underline{\lim_{n\to\infty}}\, \frac{1}{n} \log \beta_n \geq \theta. \tag{9}$$

We use the abbreviations:

$$P_{YZ|X}(y, z|x) := \sum_{s \in \mathcal{S}} P_S(s) \Gamma_{YZ|SX}(y, z|s, x) \tag{10}$$

$$Q_{YZ|X}(y, z|x) := \sum_{s \in \mathcal{S}} Q_S(s) \Gamma_{YZ|SX}(y, z|s, x). \tag{11}$$

Let also $P_{Y|X}, P_{Z|X}$ and $Q_{Z|X}, Q_{Y|X}$ denote the respective conditional marginals.

Theorem 1: For any $\epsilon \in [0, 1)$ and $r > 0$, a rate-exponent triple (R, θ, r) is (ϵ, r)-achievable, if and only if, there exists a pmf P_X satisfying

$$R \leq \min\{I_{P_X P_{Y|X}}(X; Y), I_{P_X Q_{Y|X}}(X; Y)\}, \tag{12}$$

and

$$\theta \leq \min_{\substack{\bar{P}_{Z|X}: \\ \mathbb{E}_{P_X}[D(\bar{P}_{Z|X} \| P_{Z|X})] \leq r}} \mathbb{E}_{P_X}\left[D(\bar{P}_{Z|X} \| Q_{Z|X})\right]. \tag{13}$$

Proof: Achievability follows by standard random coding arguments for a compound channel, where the transmitter uniformly picks the codewords over the set of n-length sequences of a fixed type P_X and the decoder uses a universal decoding rule such as a maximum mutual information (MMI) decoder. The radar receiver checks the conditional type of the received sequence z^n given the transmitted codeword x^n and decides on $\hat{\mathcal{H}} = 0$ if the conditional type $\boldsymbol{\pi}_{z^n|x^n}$ satisfies

$$\mathbb{E}_{P_X}[D(\boldsymbol{\pi}_{z^n|x^n} \| P_{Z|X})] \leq r. \tag{14}$$

The converse, which is the main contribution of this paper, is proved in Section III. ■

Remark 1: Above Theorem 1 applies to a setup with *maximum* probabilities of error, see (3), (5), and (6). The theorem however applies unchanged also when the definitions (3), (5), and (6) are replaced by the following *average* probabilities of error:

$$p^{(n)}(\text{error}) := \max_{\mathsf{H} \in \{0,1\}} \Pr\left[\hat{M} \neq M | \mathcal{H} = \mathsf{H}\right]. \tag{15a}$$

$$\alpha_n := \Pr\left[\hat{\mathcal{H}} = 1 \middle| \mathcal{H} = 0\right] \tag{15b}$$

$$\beta_n := \Pr\left[\hat{\mathcal{H}} = 0 \middle| \mathcal{H} = 1\right] \tag{15c}$$

under the condition that $\epsilon \in [0, 1/2)$. The converse proof under these average probability of error criteria is obtained by first applying expurgation arguments and then following similar steps as in Section III.

III. Strong Converse Proof

Fix a sequence of encoding, decoding, and estimation functions $\{(\phi^{(n)}, g^{(n)}, h^{(n)})\}_{n=1}^{\infty}$. Assume that (7) and (8) are satisfied. For readability, we will also write $x^n(\cdot)$ for the function $\phi^{(n)}(\cdot)$. Choose a sequence of small positive numbers $\{\mu_n\}_{n=1}^{\infty}$ satisfying

$$\lim_{n\to\infty} \mu_n \;=\; 0 \tag{16}$$

$$\lim_{n\to\infty} n\cdot\mu_n^2 \;=\; \infty. \tag{17}$$

Let T be uniform over $\{1,\ldots,n\}$ independent of all other quantities and consider an increasing subsequence of blocklengths $\{n_i\}$ so that the expected type $\mathbb{E}_M[\boldsymbol{\pi}_{x^n(M)}(x)]$ converges and denote the convergence point by $P_X(x)$:

$$P_X(x) := \lim_{i\to\infty} \mathbb{E}_M[\boldsymbol{\pi}_{x^{n_i}(M)}(x)], \qquad x\in\mathcal{X}. \tag{18}$$

In the remainder of this proof, we restrict attention to this subsequence of blocklengths $\{n_i\}_{i=1}^{\infty}$.

Proof of Channel Coding Bound: We first prove the converse bound for channel coding. We start by considering the case $\mathcal{H}=0$ and channel transition law $P_{Y|X}$.

Fix a blocklength $n \in \{n_i\}$. Based on the two conditions

$$g^{(n)}\left(y^n\right) = m \tag{19a}$$

$$\left|\pi_{x^n(m)y^n}(a,b) - \pi_{x^n(m)}(a)P_{Y|X}(b|a)\right| \le \mu_n, \tag{19b}$$

define for each message $m\in\mathcal{M}$ the set

$$\mathcal{D}_{\mathcal{C},m} := \{y^n\colon \quad \text{(19a) and (19b)}\}. \tag{20}$$

Introduce the new random tuple $Y_{\mathcal{C}}^n$ with the following conditional pmf given the message M:

$$P_{Y_{\mathcal{C}}^n|M}(y^n|m) = \frac{P_{Y|X}^{\otimes n}(y^n|x^n(m))}{\Delta_{\mathcal{C},m}}\cdot \mathbb{1}\{y^n\in\mathcal{D}_{\mathcal{C},m}\}, \tag{21}$$

for

$$\Delta_{\mathcal{C},m} := \sum_{y^n} P_{Y|X}^{\otimes n}(y^n|x^n(m))\cdot\mathbb{1}\{y^n\in\mathcal{D}_{\mathcal{C},m}\}. \tag{22}$$

By the union bound, by (7), and by [29, Remark to Lemma 2.12], we have:

$$\Delta_{\mathcal{C},m} \ge 1-\epsilon-\frac{|\mathcal{X}||\mathcal{Y}|}{4\mu_n^2 n}, \quad \forall m\in\mathcal{M}. \tag{23}$$

Continue to notice that:

$$\begin{aligned} R &\overset{(a)}{=} \frac{1}{n} I(M;Y_{\mathcal{C}}^n) && (24)\\ &\le \frac{1}{n}\sum_{i=1}^{n} H(Y_{\mathcal{C},i}) - \frac{1}{n}H\left(Y_{\mathcal{C}}^n\middle|M\right) && (25)\\ &= H(Y_{\mathcal{C},T}|T) - \frac{1}{n}H\left(Y_{\mathcal{C}}^n\middle|M\right) && (26)\\ &\le H\left(Y_{\mathcal{C},T}\right) - \frac{1}{n}H\left(Y_{\mathcal{C}}^n\middle|M\right), && (27) \end{aligned}$$

where we defined the random variable T to be uniform over $\{1,\ldots,n\}$ independent of the other random variables. Here, (a) holds because $M=g(Y_{\mathcal{C}}^n)$ by Condition (19a).

Define next $X_T = x_T(M)$ (the T-th symbol of codeword $x^n(M)$), and notice that

$$\begin{aligned} P_{X_T Y_{\mathcal{C},T}}(x,y) &= \frac{1}{n}\sum_{t=1}^{n} P_{X_t Y_{\mathcal{C},t}}(x,y) && (28)\\ &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\mathbb{1}\left\{(X_t,Y_{\mathcal{C},t})=(x,y)\right\}\right] && (29)\\ &= \mathbb{E}\left[\pi_{x^n(M)Y_{\mathcal{C}}^n}(x,y)\right] && (30)\\ &= \mathbb{E}_M\left[\pi_{x^n(M)}(x)\right]\cdot P_{Y|X}(y|x) + o(1), && (31) \end{aligned}$$

where the last equality holds by Condition (19b). By continuity of the entropy functional and the definition in (18):

$$\lim_{n_i\to\infty} H\left(Y_{\mathcal{C},T}\right) = H_{P_X P_{Y|X}}(Y). \tag{32}$$

Next, by definition and by (21):

$$\begin{aligned} &\frac{1}{n}H(Y_{\mathcal{C}}^n|M=m)\\ &= -\frac{1}{n}\sum_{y^n\in\mathcal{D}_{\mathcal{C},m}} P_{Y_{\mathcal{C}}^n|M=m}(y^n)\log P_{Y_{\mathcal{C}}^n|M=m}(y^n) && (33)\\ &\ge -\frac{1}{n}\sum_{y^n\in\mathcal{D}_{\mathcal{C},m}} P_{Y_{\mathcal{C}}^n|M=m}(y^n)\log\frac{P_{Y|X}^{\otimes n}(y^n|x^n(m))}{\Delta_{\mathcal{C},m}}\\ & && (34)\\ &= -\frac{1}{n}\sum_{t=1}^{n}\sum_{y_t\in\mathcal{Y}} P_{Y_{\mathcal{C},t}|M=m}(y_t)\log P_{Y|X}(y_t|x_t(m))\\ &\quad+\frac{1}{n}\log\Delta_{\mathcal{C},m}, && (35)\\ &= -\frac{1}{n}\sum_{t=1}^{n}\sum_{y\in\mathcal{Y}}\mathbb{E}\left[\mathbb{1}\{Y_{\mathcal{C},t}=y\}\middle|M=m\right]\log P_{Y|X}(y|x_t(m))\\ &\quad+\frac{1}{n}\log\Delta_{\mathcal{C},m}, && (36)\\ &= -\frac{1}{n}\sum_{t=1}^{n}\sum_{y\in\mathcal{Y}}\mathbb{E}\left[\sum_{x\in\mathcal{X}}\mathbb{1}\{x_t(m)=x, Y_{\mathcal{C},t}=y\}\middle|M=m\right]\\ &\qquad\cdot\log P_{Y|X}(y|x)\\ &\quad+\frac{1}{n}\log\Delta_{\mathcal{C},m}, && (37)\\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathbb{E}\left[\pi_{x^n(m)Y_{\mathcal{C}}^n}(x,y)\middle|M=m\right]\cdot\log P_{Y|X}(y|x)\\ &\quad+\frac{1}{n}\log\Delta_{\mathcal{C},m} && (38)\\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\pi_{x^n(m)}(x)P_{Y|X}(y|x)\cdot\log P_{Y|X}(y|x)+o(1),\\ & && (39) \end{aligned}$$

where the last equality holds by Condition (19b) and because the two Inequalities (23) and $\epsilon\in[0,1)$ imply that $\frac{1}{n}\log\Delta_{\mathcal{C},m}$

vanishes as $n \to \infty$. Taking the average over all messages yields:

$$\frac{1}{n} H(Y_{\mathcal{C}}^n|M) \geq -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathbb{E}\left[\pi_{x^n(M)}(x)\right] \log P_{Y|X}(y|x) \\ +o(1). \tag{40}$$

Using the definition of P_X, we obtain:

$$\begin{aligned} &\lim_{i\to\infty} \frac{1}{n_i} H(Y_{\mathcal{C}}^n|M) \\ &\quad = -\sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) \log P_{Y|X}(y|x) \\ &\quad = H_{P_X P_{Y|X}}(Y|X). \end{aligned} \tag{41}$$

Combining (27) with (32) and (41), we can conclude that

$$R \leq H_{P_X P_{Y|X}}(Y) - H_{P_X P_{Y|X}}(Y|X) \tag{42}$$

$$= I_{P_X P_{Y|X}}(X;Y), \tag{43}$$

for P_X as defined in (18).

Following the same steps (19)–(43) but with the channel law $Q_{Y|X}$ instead of $P_{Y|X}$, one can show that

$$R \leq \min\left\{I_{P_X P_{Y|X}}(X;Y),\ I_{P_X Q_{Y|X}}(X;Y)\right\}. \tag{44}$$

Proof of the Error Exponents: Fix a small value of $\delta > 0$ and consider any conditional type $\bar{P}_{Z|X}$ so that

$$\mathbb{E}_{P_X}\left[D\left(\bar{P}_{Z|X} \,\|\, P_{Z|X}\right)\right] < r - \delta. \tag{45}$$

Fix a sufficiently large blocklength $n \in \{n_i\}$ and a message m, for which the following two inequalities hold:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\pi}_{x^n(m)}}\left[D\left(\bar{P}_{Z|X} \,\|\, P_{Z|X}\right)\right] \\ &\quad \leq \mathbb{E}_M\left[\mathbb{E}_{\boldsymbol{\pi}_{x^n(M)}}\left[D\left(\bar{P}_{Z|X} \,\|\, P_{Z|X}\right)\right]\right] \qquad (46) \\ &\quad < r - \delta/2. \qquad (47) \end{aligned}$$

Then, based on the conditions

$$h^{(n)}\left(x^n(m), z^n\right) = 0 \tag{48a}$$

$$\left|\pi_{x^n(m)z^n}(a,b) - \pi_{x^n(m)}(a) P_{Z|X}(b|a)\right| \leq \mu_n, \tag{48b}$$

define the set

$$\mathcal{D}_{\mathcal{S},m} \triangleq \{z^n\colon \quad \text{(48a) and (48b)}\}. \tag{49}$$

Define also the new random variable $Z_{\mathcal{S}}^n$ of conditional law

$$P_{Z_{\mathcal{S}}^n}(z^n) = \frac{P_{Z|X}^{\otimes n}(z^n|x^n(m))}{\Delta_{\mathcal{S},m}} \cdot \mathbb{1}\left\{z^n \in \mathcal{D}_{\mathcal{S},m}\right\}, \tag{50}$$

for

$$\Delta_{\mathcal{S},m} := \sum_{z^n} P_{Z|X}^{\otimes n}(z^n|x^n(m)) \cdot \mathbb{1}\left\{z^n \in \mathcal{D}_{\mathcal{S},m}\right\}. \tag{51}$$

Defining $D_b(a\|b) := a \log_2 \frac{a}{b} + (1-a)\log_2 \frac{1-a}{1-b}$, we notice the sequence of (in)equalities:

$$\begin{aligned} &-\frac{1}{n} \log \Pr\left[\hat{\mathcal{H}} = 0 \middle| \mathcal{H} = 1, M = m\right] \\ &\leq -\frac{1}{n} \log \sum_{z^n\in\mathcal{D}_{\mathcal{S},m}} Q_{Z|X}^{\otimes n}(z^n|x^n(m)) \qquad (52) \\ &= \frac{1}{n} D_b\left(\sum_{z^n\in\mathcal{D}_{\mathcal{S},m}} P_{Z_{\mathcal{S}}^n}(z^n) \,\|\, \sum_{z^n\in\mathcal{D}_{\mathcal{S},m}} Q_{Z|X}^{\otimes n}(z^n|x^n(m))\right) \\ &\qquad (53) \\ &\leq \frac{1}{n} D\left(P_{Z_{\mathcal{S}}^n} \middle\| \, Q_{Z|X}^{\otimes n}(\cdot|x^n(m))\right) \qquad (54) \\ &\leq \frac{1}{n} \sum_{z^n\in\mathcal{Z}^n} P_{Z_{\mathcal{S}}^n}(z^n) \log \frac{P_{Z|X}^{\otimes n}(z^n|x^n(m))}{Q_{Z|X}^{\otimes n}(z^n|x^n)} \\ &\quad -\frac{1}{n} \log \Delta_{\mathcal{S},m} \qquad (55) \\ &= \frac{1}{n} \sum_{x\in\mathcal{X}} \sum_{\substack{t\in\{1,\ldots,n\}\colon\\ x_t(m)=x}} \sum_z P_{\tilde{Z}_{\mathcal{S},t}}(z) \log \frac{P_{Z|X}(z|x)}{Q_{Z|X}(z|x)} \\ &\quad -\frac{1}{n} \log \Delta_{\mathcal{S},m}, \qquad (56) \end{aligned}$$

where Inequality (53) holds because $\sum_{z^n\in\mathcal{D}_{\mathcal{S},m}} P_{Z_{\mathcal{S}}^n}(z^n) = 1$.

To bound the term $\Delta_{Z,m}$, let $\mathcal{T}_{x^n(m)}(\bar{P}_{Z|X})$ be the set of z^n sequences satisfying (48a) (and thus $\mathcal{D}_{Z,m} \subseteq \mathcal{T}_{x^n(m)}(\bar{P}_{Z|X})$) and notice that

$$\begin{aligned} \alpha_n &\geq \Pr\left[\hat{\mathcal{H}} = 1 \middle| \mathcal{H} = 0, M = m\right] \qquad (57) \\ &\geq \sum_{z^n\in\mathcal{T}_{x^n(m)}(\bar{P}_{Z|X})\setminus\mathcal{D}_{\mathcal{S},m}} P_{Z|X}^{\otimes n}(z^n|x^n(m)) \qquad (58) \\ &= \sum_{z^n\in\mathcal{T}_{x^n(m)}(\bar{P}_{Z|X})} P_{Z|X}^{\otimes n}(z^n|x^n(m)) - \Delta_{\mathcal{S},m} \qquad (59) \\ &= 2^{-n\left(\mathbb{E}_{\boldsymbol{\pi}_{x^n(m)}(x)}\left[D(\bar{P}_{Z|X}\|P_{Z|X})\right]+o(1)\right)} - \Delta_{\mathcal{S},m}, \qquad (60) \end{aligned}$$

where the last equality holds by a conditional version of Sanov's theorem. Therefore,

$$\Delta_{\mathcal{S},m} \geq 2^{-n\left(\mathbb{E}_{\boldsymbol{\pi}_{x^n(m)}}\left[D(\bar{P}_{Z|X}\|P_{Z|X})\right]+o(1)\right)} - \alpha_n, \tag{61}$$

and by the condition on the type-I error probability (8) and our assumption (46)–(47):

$$-\frac{1}{n} \log \Delta_{\mathcal{S},m} \leq \mathbb{E}_{\boldsymbol{\pi}_{x^n(m)}}\left[D(\bar{P}_{Z|X}\|P_{Z|X})\right] + o(1). \tag{62}$$

We next observe that

$$\begin{aligned} \frac{1}{n} \sum_{\substack{t\in\{1,\ldots,n\}\colon\\ x_t(\tilde{M})=x}} P_{\tilde{Z}_t}(z) &= \frac{1}{n} \sum_{\substack{t\in\{1,\ldots,n\}\colon\\ x_t(\tilde{M})=x}} \mathbb{1}\{\tilde{Z}_t = z\} \qquad (63) \\ &= \frac{1}{n} \sum_{t\in\{1,\ldots,n\}} \mathbb{1}\{x_t(\tilde{M}) = x, \tilde{Z}_t = z\} \\ &\qquad (64) \\ &= \boldsymbol{\pi}_{x^n(m)\tilde{Z}^n}(x,z) \qquad (65) \\ &= \boldsymbol{\pi}_{x^n(m)}(x)\bar{P}_{Z|X}(z|x) + o(1), \qquad (66) \end{aligned}$$

where the last equation holds by the type-condition (48a).

From the definition of β_n and by combining (56), (62), and (66), we obtain:

$$-\frac{1}{n} \log \beta_n$$

$$\le -\frac{1}{n}\log \Pr\left[\hat{\mathcal{H}}=0\middle|\mathcal{H}=1, M=m\right] \tag{67}$$

$$\begin{aligned}\le \sum_x \boldsymbol{\pi}_{x^n(m)}(x)\sum_z \bar{P}_{Z|X}(z|x)\frac{\bar{P}_{Z|X}(z|x)}{Q_{Z|X}(z|x)}\\ +\mathbb{E}_{\boldsymbol{\pi}_{x^n(m)}}\left[D(\bar{P}_{Z|X}\|P_{Z|X})\right]+o(1)\end{aligned} \tag{68}$$

$$\le \sum_x \boldsymbol{\pi}_{x^n(m)}(x)D(\bar{P}_{Z|X}(\cdot|x)\|Q_{Z|X}(\cdot|x))+o(1) \tag{69}$$

$$\le \sum_x \mathbb{E}_M[\boldsymbol{\pi}_{x^n(M)}(x)]D(\bar{P}_{Z|X}(\cdot|x)\|Q_{Z|X}(\cdot|x))+o(1), \tag{70}$$

where the last step holds by (47).

The desired converse is then immediately established by considering the accumulation point of the increasing blocklengths $\{n_i\}$, and by using definition (18).

IV. Conclusion and Future Directions

We established the strong converse for a memoryless ISAC problem with bi-static radar when the sensing performance is measured in terms of the largest exponential decay rates of the detection error probabilities under the two hypotheses. Notice that our model also includes as special case the setups where the receiver has perfect or imperfect channel-state information by including this state-information as part of the output.

Interesting future research directions include extensions to mono-static radar systems where the transmitter can apply closed-loop encodings depending also on past generalized feedback systems or systems with memory. Analyzing other sensing criteria is also of interest, such as the estimation error when the distribution of the state-sequence depends on a single continuous-valued parameter as for example the Doppler shift in a radar application.

Acknowledgments

This work was supported by the German Research Foundation (DFG) via the German-Israeli Project Cooperation (DIP), under Project SH 19371-1.

References

[1] O. Li, J. He, K. Zeng, Z. Yu, X. Du, Z. Zhou, Y. Liang, G. Wang, Y. Chen, P. Zhu, W. Tong, D. Lister, and L. Ibbetson, "Integrated sensing and communication in 6G: a prototype of high resolution multichannel THz sensing on portable device," *EURASIP Journal on Wireless Communications and Networking*, vol. 2022, no. 1, p. 106, 2022.

[2] M. Chafii, L. Bariah, S. Muhaidat, and M. Debbah, "Twelve scientific challenges for 6G: Rethinking the foundations of communications theory," 2023. [Online]. Available: https://arxiv.org/abs/2207.01843

[3] S. H. Dokhanchi, M. B. Shankar, M. Alaee-Kerahroodi, T. Stifter, and B. Ottersten, "Adaptive waveform design for automotive joint radar-communications system," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4280–4284.

[4] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[5] C. Sturm and W. Wiesbeck, "Waveform design and signal processing aspects for fusion of wireless communications and radar sensing," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1236–1259, 2011.

[6] M. P. Roberton and E. Brown, "Integrated radar and communications based on chirped spread-spectrum techniques," *IEEE MTT-S International Microwave Symposium Digest, 2003*, vol. 1, pp. 611–614, 2003.

[7] G. N. Saddik, R. S. Singh, and E. R. Brown, "Ultra-wideband multifunctional communications/radar system," *IEEE Transactions on Microwave Theory and Techniques*, vol. 55, no. 7, pp. 1431–1437, 2007.

[8] A. Liu, Z. Huang, M. Li, Y. Wan, W. Li, T. X. Han, C. Liu, R. Du, D. K. P. Tan, J. Lu, Y. Shen, F. Colone, and K. Chetty, "A survey on fundamental limits of integrated sensing and communication," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 994–1034, 2022.

[9] F. Liu, L. Zheng, Y. Cui, C. Masouros, A. P. Petropulu, H. Griffiths, and Y. C. Eldar, "Seventy years of radar and communications: The road from separation to integration," *IEEE Signal Processing Magazine*, vol. 40, no. 5, pp. 106–121, 2023.

[10] C. Ouyang, Y. Liu, H. Yang, and N. Al-Dhahir, "Integrated sensing and communications: A mutual information-based framework," *IEEE Communications Magazine*, vol. 61, no. 5, pp. 26–32, 2023.

[11] A. Liu, M. Li, M. Kobayashi, and G. Caire, *Fundamental Limits for ISAC: Information and Communication Theoretic Perspective*. Singapore: Springer Nature Singapore, 2023, pp. 23–52.

[12] M. Kobayashi, G. Caire, and G. Kramer, "Joint state sensing and communication: Optimal tradeoff for a memoryless case," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 111–115.

[13] M. Kobayashi, H. Hamad, G. Kramer, and G. Caire, "Joint state sensing and communication over memoryless multiple access channels," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 270–274.

[14] M. Ahmadipour, M. Kobayashi, M. Wigger, and G. Caire, "An information-theoretic approach to joint sensing and communication," *IEEE Transactions on Information Theory*, pp. 1–1, 2022.

[15] M. Ahmadipour, M. Wigger, and M. Kobayashi, "Coding for sensing: An improved scheme for integrated sensing and communication over MACs," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 3025–3030.

[16] O. Günlü, M. R. Bloch, R. F. Schaefer, and A. Yener, "Secure integrated sensing and communication," *IEEE Journal on Selected Areas in Information Theory*, vol. 4, pp. 40–53, 2023.

[17] H. Joudeh and F. M. J. Willems, "Joint communication and binary state detection," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 1, pp. 113–124, 2022.

[18] H. Wu and H. Joudeh, "On joint communication and channel discrimination," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 3321–3326.

[19] M.-C. Chang, S.-Y. Wang, T. Erdoğan, and M. R. Bloch, "Rate and detection-error exponent tradeoff for joint communication and sensing of fixed channel states," *IEEE Journal on Selected Areas in Information Theory*, vol. 4, pp. 245–259, 2023.

[20] M. Ahmadipour, M. Wigger, and S. Shamai, "Strong converses for memoryless bi-static ISAC," in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 1818–1823.

[21] G. Choi and N. Lee, "Information-theoretical approach to integrated pulse-doppler radar and communication systems," 2023. [Online]. Available: https://arxiv.org/abs/2302.05046

[22] R. Ahlswede, "Certain results in coding theory for compound channels," in *Proc. Coll. Information Theory*, vol. 1, 1967, pp. 35–60.

[23] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2148–2177, 1998.

[24] M. Hamad, M. Wigger, and M. Sarkiss, "Strong converses using typical changes of measures and asymptotic markov chains," *IEEE Transactions on Information Theory*, pp. 1–1, 2023.

[25] W. Gu and M. Effros, "A strong converse for a collection of network source coding problems," in *2009 IEEE International Symposium on Information Theory*. IEEE, 2009, pp. 2316–2320.

[26] ——, "A strong converse in source coding for super-source networks," in *2011 IEEE International Symposium on Information Theory*. IEEE, 2011, pp. 395–399.

[27] T. Han and K. Kobayashi, "Exponential-type error probabilities for multiterminal hypothesis testing," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 2–14, 1989.

[28] H. Tyagi and S. Watanabe, "Strong converse using change of measure arguments," *IEEE Transactions on Information Theory*, vol. 66, no. 2, pp. 689–703, 2019.

[29] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

Universal Neyman–Pearson Classification with a Known Hypothesis

Parham Boroumand
University of Cambridge
pb702@cam.ac.uk

Albert Guillén i Fàbregas
University of Cambridge
Universitat Pompeu Fabra
guillen@ieee.org

***Abstract*—We propose a universal classifier for binary Neyman–Pearson classification where null distribution is known while only a training sequence is available for the alternative distribution. The proposed classifier interpolates between Hoeffding's classifier and the likelihood ratio test and attains the same error probability prefactor as the likelihood ratio test, i.e., the same prefactor as if both distributions were known. Similarly to Hoeffding's universal hypothesis test, the proposed classifier is shown to attain the optimal error exponent tradeoff attained by the likelihood ratio test whenever the ratio of training to observation samples exceeds a certain value.**

I. Prelimenaries

Consider the following binary classification problem where an observation $\boldsymbol{x} = (x_1, \ldots, x_n)$ is generated in an i.i.d. fashion from either of two possible distributions P_0 or P_1 defined on the probability simplex $\mathcal{P}(\mathcal{X})$ with alphabet size $|\mathcal{X}| < \infty$. We assume that the distribution P_0 is known while only a sequence of training samples $\boldsymbol{z} = (z_1, \ldots, z_k) \sim P_1^k$ generated in an i.i.d. fashion from P_1 is available; training and test sequences are sampled independently from each other. We also assume that both $P_0(x) > 0, P_1(x) > 0$ and $\frac{P_0(x)}{P_1(x)} \le c$ for each $x \in \mathcal{X}$ for some positive c. Also we let k, the length of the training, be such that $k = \alpha n$ for some positive α.

The type of an n-length sequence $\boldsymbol{y}$ is defined as $\hat{T}_{\boldsymbol{y}}(a) = \frac{N(a|\boldsymbol{y})}{n}$, where $N(a|\boldsymbol{y})$ is the number of occurrences of symbol $a \in \mathcal{X}$ in sequence $\boldsymbol{y}$. The types of the observation and training sequences $\boldsymbol{x}, \boldsymbol{z}$ are denoted by $\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}$ respectively. The set of all sequences of length n with type P, denoted by $\mathcal{T}_P^n$, is called the type class. The set of types formed with length n sequences on the simplex $\mathcal{P}(\mathcal{X})$ is denoted as $\mathcal{P}_n(\mathcal{X})$.

Let $\phi(\boldsymbol{z}, \boldsymbol{x}) : \mathcal{X}^k \times \mathcal{X}^n \to \{0, 1\}$ be a classifier that decides the distribution that generated the observation $\boldsymbol{x}$ upon processing the training sequence $\boldsymbol{z}$. We consider deterministic classifiers ϕ that decide in favor of P_0 if $\boldsymbol{x} \in \mathcal{A}_0(P_0, \boldsymbol{z})$, where $\mathcal{A}_0(P_0, \boldsymbol{z}) \subset \mathcal{X}^n$ is the decision region for the first hypothesis and is a function of P_0 and the training samples $\boldsymbol{z}$. We define $\mathcal{A}_1(P_0, \boldsymbol{z}) = \mathcal{X}^n \setminus \mathcal{A}_0$ to be the decision region for the second hypothesis. If we assume no prior knowledge on either distribution, the two possible pairwise error probabilities determine the performance of the classifier. Specifically, the type-I and type-II error probabilities are defined as

$$\epsilon_0(\phi) = \sum_{\boldsymbol{z} \in \mathcal{X}^k} P_1(\boldsymbol{z}) \sum_{\boldsymbol{x} \in \mathcal{A}_1(P_0, \boldsymbol{z})} P_0(\boldsymbol{x}), \tag{1}$$

$$\epsilon_1(\phi) = \sum_{\boldsymbol{z} \in \mathcal{X}^k} P_1(\boldsymbol{z}) \sum_{\boldsymbol{x} \in \mathcal{A}_0(P_0, \boldsymbol{z})} P_1(\boldsymbol{x}). \tag{2}$$

In the case where both distributions are known, the training sequence is not needed and the classifier becomes a hypothesis test. In this case, the classifier is said to be optimal whenever it achieves the optimal error probability tradeoff given by

$$\min_{\phi : \epsilon_0(\phi) \le \xi} \epsilon_1(\phi), \tag{3}$$

where $\xi \in [0, 1]$. It is well known that likelihood ratio test

$$\phi^{\mathrm{lrt}}(\boldsymbol{x}) = \mathbb{1}\left\{ \frac{P_1^n(\boldsymbol{x})}{P_0^n(\boldsymbol{x})} \ge e^{n\gamma} \right\}, \tag{4}$$

attains the optimal tradeoff (3) for every γ. This is the well-known Neyman–Pearson lemma [1]. The likelihood ratio test can also be expressed as a function of the type of the observation $\hat{T}_{\boldsymbol{x}}$ as e.g.[2], [3]

$$\phi^{\mathrm{lrt}}(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{ D(\hat{T}_{\boldsymbol{x}} \| P_0) - D(\hat{T}_{\boldsymbol{x}} \| P_1) \ge \gamma \big\} \tag{5}$$

where $D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$ is the relative entropy between distributions P and Q. The optimal error exponent tradeoff (E_0, E_1) is defined as

$$\begin{aligned} E_1^*(E_0) \triangleq \sup \big\{ & E_1 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t.} \\ & \forall n > n_0, \epsilon_0(\phi) \le e^{-nE_0} \quad \text{and} \quad \epsilon_1(\phi) \le e^{-nE_1} \big\}. \end{aligned} \tag{6}$$

By using Sanov's Theorem [2], [4], the optimal error exponent tradeoff (E_1, E_0), attained by the likelihood ratio is given by

$$E_0(\phi^{\mathrm{lrt}}) = \min_{Q \in \mathcal{Q}_0(\gamma)} D(Q \| P_0), \tag{7}$$

$$E_1(\phi^{\mathrm{lrt}}) = \min_{Q \in \mathcal{Q}_1(\gamma)} D(Q \| P_1), \tag{8}$$

where

$$\mathcal{Q}_0(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) - D(Q\|P_1) \ge \gamma \big\}, \tag{9}$$

$$\mathcal{Q}_1(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_0) - D(Q\|P_1) \le \gamma \big\}. \tag{10}$$

By varying the threshold γ in the range $-D(P_0\|P_1) \le \gamma \le D(P_1\|P_0)$, Eqs. (7) and (8) fully characterize the error exponent tradeoff in (6).

This work has been funded in part by the European Research Council under ERC grant agreement 725411 and by the Spanish Ministry of Economy and Competitiveness under grant PID2020-116683GB-C22.

The classification problem described above with known P_0 and a training sequence from P_1, can also be viewed as the composite binary hypothesis problem where additional training sequence samples are given for the second hypotheses. In a composite hypothesis testing setting with given P_0 and the other hypothesis is unrestricted to $\mathcal{P}(\mathcal{X})$, Hoeffding proposed the generalized likelihood-ratio test given by [5]

$$\phi^{\text{glrt}}(\boldsymbol{x}) = \mathbb{1}\big\{D(\hat{T}_{\boldsymbol{x}}\|P_0) > E_0\big\}, \tag{11}$$

By Sanov's theorem, the error exponent of Hoeffding's test is given by

$$E_0(\phi^{\text{glrt}}) = E_0, \tag{12}$$

$$E_1(\phi^{\text{glrt}}) = \min_{\substack{Q\in\mathcal{P}(\mathcal{X}),\\ D(Q\|P_0)\le E_0}} D(Q\|P_1). \tag{13}$$

By varying the threshold E_0 in the range $0 \le E_0 \le D(P_1\|P_0)$, (12) and (13) fully characterize the optimal error exponent tradeoff in (6). Using a large deviations refinement [6], [7], the type-I error probability of the likelihood ratio test can be expressed as

$$\epsilon_0(\phi^{\text{lrt}}) = \frac{1}{\sqrt{n}}e^{-nE_0}\big(c + o(1)\big), \tag{14}$$

while, for Hoeffding's test it can be expressed as [8], [6]

$$\epsilon_0(\phi^{\text{glrt}}) = n^{\frac{|\mathcal{X}|-3}{2}}e^{-nE_0}\big(c' + o(1)\big) \tag{15}$$

where c, c' are constants that only depend on P_0, P_1 and the corresponding test thresholds. Since the likelihood ratio and Hoeffding's tests attain the optimal error exponent tradeoff (6), for any fixed E_0, then $E_1(\phi^{\text{glrt}}) = E_1(\phi^{\text{lrt}})$. As a result, when the number of observations is large, Hoeffding's test, although attaining the optimal error exponent tradeoff, suffers in exponential prefactor when compared to the likelihood ratio's $\frac{1}{\sqrt{n}}$ for observation alphabets such that $|\mathcal{X}| > 2$. For $|\mathcal{X}| = 2$, the decision regions for the likelihood ratio and Hoeffding's tests coincide and thus, (15) is the same as (14).

II. Fixed Sample Sized Universal Classifier

We propose a classifier that interpolates between the likelihood ratio and Hoeffding's tests that attains a prefactor that is independent of the alphabet size and is equal to $\frac{1}{\sqrt{n}}$. In addition, we show that if the ratio of training samples to the number of test samples α exceeds a certain threshold, the proposed test also achieves the optimal error exponent tradeoff.

Hoeffding's test can favor the second hypothesis for test sequences with types close to P_0 while far from P_1. Suppose we have a training sequence type $\hat{T}_{\boldsymbol{z}}$, we can relax the Hoeffding's test from a ball centered at P_0 to a hyperplane tangent to the Hoeffding's test ball, directed towards the type of the training sequence – this is precisely what enables the improvement in the prefactor of the type-I probability of error. We propose the following classifier

$$\phi_\beta(\hat{T}_{\boldsymbol{x}}, \hat{T}_{\boldsymbol{z}}) = \mathbb{1}\big\{\beta D(\hat{T}_{\boldsymbol{x}}\|\hat{T}'_{\boldsymbol{z}}) - D(\hat{T}_{\boldsymbol{x}}\|P_0) \ \le \gamma(E_0, \hat{T}'_{\boldsymbol{z}})\big\}, \tag{16}$$

where $0 \le \beta \le 1$, the threshold $\gamma(E_0, Q_1)$ is given by

$$\gamma(E_0, Q_1) = \beta \min_{\substack{Q\in\mathcal{P}(\mathcal{X}),\\ D(Q\|P_0)\le E_0}} D(Q\|Q_1) - E_0, \tag{17}$$

and the perturbed training type $\hat{T}'_{\boldsymbol{z}}(a)$ is

$$\hat{T}'_{\boldsymbol{z}}(a) = \big(1-\delta_n\big)\hat{T}_{\boldsymbol{z}}(a) + \frac{\delta_n}{|\mathcal{X}|}, \tag{18}$$

where, δ_n can be chosen as any function of the order $o(n^{-1})$. We add this small perturbation of the training type to avoid the cases where some of the alphabet symbols have not been observed in the training sequence. We define the decision regions of the proposed classifier by

$$\mathcal{A}_0(\hat{T}_{\boldsymbol{z}}, \beta) = \{Q : Q \in \mathcal{P}(\mathcal{X}), \phi_\beta(Q, \hat{T}_{\boldsymbol{z}}) = 0\}, \tag{19}$$

$$\mathcal{A}_1(\hat{T}_{\boldsymbol{z}}, \beta) = \{Q : Q \in \mathcal{P}(\mathcal{X}), \phi_\beta(Q, \hat{T}_{\boldsymbol{z}}) = 1\}. \tag{20}$$

Since parameter β controls how much the training weights in the decision, we have that when $\beta = 0$ we recover Hoeffding's test while for $\beta = 1$ the test is reminiscent of a likelihood ratio test where instead of P_1, we have the perturbed training type $\hat{T}'_{\boldsymbol{z}}(a)$. Intuitively, as long as we have enough training samples, the training type $\hat{T}'_{\boldsymbol{z}}(a)$ will be close to P_1 and we will attain the optimal error exponent tradeoff.

Next, we find a refined expression for the type-I error probability and show that the error probability prefactor is of order $O(\frac{1}{\sqrt{n}})$, i.e., of the same order of the prefactor achieved by the likelihood ratio test.

Theorem 1: For $P_0, P_1, 0 < \beta \le 1$ and fixed E_0, the classifier ϕ_β defined in (16) attains a type-I error probability such that

$$\epsilon_0(\phi_\beta) = \frac{1}{\sqrt{n}}e^{-nE_0}(c + o(1)), \tag{21}$$

In addition, for every $P_0, P_1, E_0, \beta \in (0, 1]$, there exists a finite training to sample size ratio α^*_β such that for any $\alpha > \alpha^*_\beta$

$$\epsilon_1(\phi_\beta) = \frac{1}{\sqrt{n}}e^{-nE_1^*(E_0)}(c' + o(1)), \tag{22}$$

where c, c' are positive constants that only depend on the data distributions and E_0.

Theorem 1 shows that the classifier proposed in (16) not only achieves the optimal error exponent tradeoff for $\alpha > \alpha^*_\beta$ but also achieves the same prefactor of the type-I error probability of the likelihood ratio test. This is a significant improvement with respect to the Hoeffding's universal test for observation alphabets $|\mathcal{X}| > 2$, cf. (15). The result also shows that the proposed classifier achieves the same type-II error probability prefactor as the likelihood ratio test, establishing the optimality of the proposed classifier up to a constant. The proof of the result, as well as upper and lower bounds to α^*_β and an extension to the sequential case can be found in [9].

REFERENCES

[1] J. Neyman and E. S. Pearson, “On the problem of the most efficient tests of statistical hypotheses,” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933. [Online]. Available: http://www.jstor.org/stable/91247

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[3] I. Csiszár and P. Shields, “Information theory and statistics: A tutorial,” *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004. [Online]. Available: http://dx.doi.org/10.1561/0100000004

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 01 2010, vol. 95.

[5] W. Hoeffding, *Asymptotically Optimal Tests for Multinomial Distributions*. New York, NY: Springer New York, 1994, pp. 431–471.

[6] M. Iltis, “Sharp asymptotics of large deviations,” *Journal of Theoretical Probability*, vol. 8, no. 3, pp. 501–522, Jul 1995. [Online]. Available: https://doi.org/10.1007/BF02218041

[7] G. Vazquez-Vilar, A. Guillén i Fàbregas, T. Koch, and A. Lancho, “Saddlepoint approximation of the error probability of binary hypothesis testing,” *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 2306–2310, 2018.

[8] Y. Li, S. Nitinawarat, and V. V. Veeravalli, “Universal outlier hypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, July 2014.

[9] P. Boroumand and A. Guillén i Fàbregas, “Universal Neyman–Pearson classification with a partially known hypothesis,” *submitted to Information and Inference, a journal of the IMA, https://arxiv.org/abs/2206.11700*, 2022.

Second-Order Asymptotics of Divergence Tests

K. V. Harsha,* Jithin Ravi,† and Tobias Koch‡

*Gandhi Institute of Technology and Management, Hyderabad, India
†Indian Institute of Technology Kharagpur, India
‡Universidad Carlos III de Madrid, Leganés, Spain and Gregorio Marañón Health Research Institute, Madrid, Spain.
Emails: hkalluma@gitam.edu, jithin@ece.iitkgp.ac.in, koch@tsc.uc3m.es

***Abstract*—Consider a binary statistical hypothesis testing problem, where n independent and identically distributed random variables Z^n are either distributed according to the null hypothesis P or the alternate hypothesis Q, and only P is known. A well-known test that is suitable for this case is the so-called Hoeffding test, which accepts P if the Kullback-Leibler (KL) divergence between the empirical distribution of Z^n and P is below some threshold. In this work, we characterize the first and second-order terms of the type-II error probability for a fixed type-I error probability for the Hoeffding test as well as for divergence tests, where the KL divergence is replaced by a general divergence. We demonstrate that, irrespective of the divergence, divergence tests achieve the first-order term of the Neyman-Pearson test, which is the optimal test when both P and Q are known. In contrast, the second-order term of divergence tests is strictly worse than that of the Neyman-Pearson test. We further demonstrate that divergence tests with an invariant divergence achieve the same second-order term as the Hoeffding test, but divergence tests with a non-invariant divergence may outperform the Hoeffding test for some alternate hypotheses Q.**

I. Introduction

Consider a binary hypothesis testing problem that decides whether a sequence of independent and identically distributed (i.i.d.) random variables Z^n is either generated from distribution P or from distribution Q. Assume that both distributions are discrete and the hypothesis test has access to P but not to Q. A suitable test for this case is the well-known *Hoeffding test* [1], which accepts P if $D_{\mathrm{KL}}(T_{Z^n}\|P) < c$, for some $c > 0$, and otherwise accepts Q. Here, T_{Z^n} is the type (the empirical distribution) of Z^n and $D_{\mathrm{KL}}(P\|Q)$ is the Kullback-Leibler (KL) divergence between P and Q [2]. In this paper, we analyze the second-order performance of the Hoeffding test as well as of Hoeffding-like tests, referred to as *divergence tests*, where the KL divergence is replaced by other divergences (see Section II for a rigorous definition).

We focus on the asymptotic behaviour of the type-II error β_n (the probability of declaring hypothesis P under hypothesis Q) for a fixed type-I error α_n (the probability of declaring hypothesis Q under hypothesis P). When both P and Q are known, the optimal test is the likelihood ratio test, also known as the Neyman-Pearson test. For this test, the smallest type-II error β_n for which $\alpha_n \leq \epsilon$ satisfies [3, Prop. 2.3]

$$-\ln\beta_n = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV(P\|Q)}\,\mathsf{Q}^{-1}(\epsilon) + o(\sqrt{n}) \quad (1)$$

as $n \to \infty$, where

$$V(P\|Q) \triangleq \sum_{i=1}^{k} P_i \left[\left(\ln\frac{P_i}{Q_i} - D_{\mathrm{KL}}(P\|Q)\right)^2\right] \quad (2)$$

denotes the divergence variance; $\mathsf{Q}^{-1}(\cdot)$ denotes the inverse of the tail probability of the standard Normal distribution; P_i and Q_i denote the i-th components of P and Q; and k denotes their dimension. Here and throughout this paper, we write $a_n = o(b_n)$ for two sequences $\{a_n\}$ and $\{b_n\}$ of real numbers if $\lim_{n\to\infty}\frac{a_n}{b_n} = 0$. We write $a_n = O(b_n)$ if $\overline{\lim}_{n\to\infty}|\frac{a_n}{b_n}| < \infty$. By inspecting the expansion of $-\ln\beta_n$ in (1), one can define the first-order term β' and the second-order term β'' of any hypothesis test $\mathbb{T}$ as

$$\beta' \triangleq \lim_{n\to\infty}\frac{-\ln\beta_n(\mathbb{T})}{n} \quad (3)$$

and

$$\beta'' \triangleq \lim_{n\to\infty}\frac{-\ln\beta_n(\mathbb{T}) - n\beta'}{\sqrt{n}} \quad (4)$$

if the limits exist. The first-order term β' is sometimes referred to as the *error exponent*. For the Neyman-Pearson test, we have $\beta' = D_{\mathrm{KL}}(P\|Q)$ and $\beta'' = -\sqrt{V(P\|Q)}\mathsf{Q}^{-1}(\epsilon)$.

It was shown in [1] that the first-order term β' of the Hoeffding test is also $D_{\mathrm{KL}}(P\|Q)$. In other words, the Hoeffding test is first-order optimal. Recently, we have demonstrated [4] that the second-order term of the Hoeffding test is $\beta'' = -\sqrt{V(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$, where $\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\cdot)$ denotes the inverse of the tail probability of the chi-square distribution with $k-1$ degrees of freedom. Since $\sqrt{\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} > \mathsf{Q}^{-1}(\epsilon)$, it follows that the second-order performance of the Hoeffding test is worse than that of the Neyman-Pearson test.

In this paper, we analyze the second-order performance of the divergence test $\mathbb{T}^D$, which accepts P if $D(T_{Z^n}\|P) < c$,

Part of this work was done while K. V. Harsha and J. Ravi were with the Signal Theory and Communications Department, Universidad Carlos III de Madrid, Leganés, Spain. K. V. Harsha, J. Ravi, and T. Koch have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 714161). J. Ravi has further received funding from the Science and Engineering Research Board (SERB) under Start-up Research Grant (SRG) and also from IIT Kharagpur under Faculty Start-up Research Grant (FSRG). T. Koch has further received funding from the Spanish Ministerio de Ciencia e Innovación under Grant PID2020-116683GB-C21 / AEI / 10.13039/501100011033.

for some $c > 0$, and otherwise accepts Q. The divergence D of the divergence test $\mathbb{T}^D$ is arbitrary, so $\mathbb{T}^D$ includes the Hoeffding test as a special case when $D = D_{\text{KL}}$. We demonstrate that the divergence test $\mathbb{T}^D$ achieves the same first-order term β' as the Neyman-Pearson test, irrespective of the divergence D. Hence, $\mathbb{T}^D$ is first-order optimal for every divergence D. We further demonstrate that, for the class of *invariant divergences* [5], which includes the Rényi divergence and the f-divergence (and, hence, also the KL divergence), the divergence test $\mathbb{T}^D$ achieves the same second-order term β'' as the Hoeffding test. In contrast, we show that a divergence test $\mathbb{T}^D$ with a non-invariant divergence may achieve a second-order term β'' that is strictly better than that of the Hoeffding test for some Q and ϵ.

A. Related Work

The considered hypothesis testing problem falls under the category of *composite hypothesis testing* [6]. Indeed, in composite hypothesis testing, the test has no access to the distribution P of the null hypothesis and the distribution Q of the alternate hypothesis, but it has the knowledge that P and Q belong to the sets of distributions $\mathcal{P}$ and $\mathcal{Q}$, respectively. Our setting corresponds to the case where $\mathcal{P} = \{P\}$ and $\mathcal{Q} = \mathcal{P}^c$ (where we use the notation $\mathcal{A}^c$ to denote the complement of a set $\mathcal{A}$).

The Hoeffding test is a particular instance of the *generalized likelihood-ratio test (GLRT)* [7], which is arguably the most common test used in composite hypothesis testing. A useful benchmark for the Hoeffding test is the Neyman-Pearson test, which is the optimal test when both P and Q are known. As mentioned before, the Hoeffding test achieves the same first-order term β' as the Neyman-Pearson test, both in *Stein's regime*, where the type-I error satisfies $\alpha_n \leq \epsilon$, as well as in the *doubly-exponential regime*, where $\alpha_n \leq e^{-n\gamma}$, $\gamma > 0$; see, e.g., [1], [8]–[11]. Thus, the first-order term of the Neyman-Pearson test can be achieved without having access to the distribution Q of the alternate hypothesis. However, not having access to Q negatively affects higher-order terms. For example, for a given threshold γ, the type-I error of the Hoeffding test satisfies [11, Eq. (10)]

$$\alpha_n = n^{\frac{k-3}{2}} e^{-n\gamma}(c' + o(1)) \tag{5}$$

whereas for the corresponding Neyman-Pearson test [11, Eq. (9)]

$$\alpha_n = n^{-\frac{1}{2}} e^{-n\gamma}(c + o(1)). \tag{6}$$

Here, c and c' are constants that only depend on P, Q, and γ. Moreover, it was demonstrated in [9] that the variance of the normalized Hoeffding test statistic $nD_{\text{KL}}(T_{Z^n}\|P)$ converges to $\frac{1}{2}(k-1)$ as $n \to \infty$. Both results suggest that, for moderate n, the Hoeffding test scales unfavorably with the cardinality of P and Q, which motivated the authors of [9] to propose their *test via mismatched divergence*. The same observation can be made for Stein's regime. Indeed, as mentioned before, the second-order term of the Hoeffding test is [4]

$$\beta'' = -\sqrt{V(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} \tag{7}$$

whereas the second-order term of the Neyman-Pearson test is [3, Prop. 2.3]

$$\beta'' = -\sqrt{V(P\|Q)}\mathsf{Q}^{-1}(\epsilon). \tag{8}$$

Since $\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)$ is monotonically increasing in k, this again suggests an unfavorable scaling with the cardinality of P and Q.

Our setting where $\mathcal{P} = \{P\}$ and $\mathcal{Q} = \mathcal{P}^c$ was also studied by Watanabe [12], who proposed a test that is second-order optimal in some sense. The related case where only training sequences are available for both P and Q was considered in [13]. The test proposed in [13] was later shown to be second-order optimal [14].

II. Divergence and Divergence Test

A. Divergence

Let us consider a random variable Z that takes value in a discrete set $\mathcal{Z} = \{a_1, \cdots, a_k\}$ with cardinality $|\mathcal{Z}| = k \geq 2$. Let $\overline{\mathcal{P}}(\mathcal{Z})$ denote the set of probability distributions on $\mathcal{Z}$, and let $\mathcal{P}(\mathcal{Z})$ denote the set of probability distributions with strictly positive probabilities. Any probability distribution $R \in \mathcal{P}(\mathcal{Z})$ can be written as a length-k vector $R = (R_1, \cdots, R_k)^\mathsf{T}$, where $R_i \triangleq \Pr\{Z = a_i\}, i = 1, \cdots, k$. Note that this R can also be represented by its first $(k-1)$ components, denoted by the vector $\mathbf{R} = (R_1, \cdots, R_{k-1})^\mathsf{T}$, which takes value in the coordinate space

$$\Xi \triangleq \left\{(R_1, \cdots, R_{k-1})^\mathsf{T} \colon R_i > 0, \sum_{i=1}^{k-1} R_i < 1\right\}. \tag{9}$$

Given any two probability distributions $S, R \in \mathcal{P}(\mathcal{Z})$, one can define a non-negative function $D(S\|R)$, called a *divergence*, which represents a measure of discrepancy between them. A divergence is not necessarily symmetric in its arguments and also need not satisfy the triangle inequality; see [15], [16] for more details. More precisely, a divergence is defined as follows [15]:

Definition 1: Consider two distributions S and R in $\mathcal{P}(\mathcal{Z})$. A *divergence* $D \colon \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty)$ between S and R, denoted by $D(S\|R)$, is a smooth function[1] of $\mathbf{S} \in \Xi$ and $\mathbf{R} \in \Xi$ (we may write $D(S\|R) = D(\mathbf{S}\|\mathbf{R})$) satisfying the following conditions:

1) $D(S\|R) \geq 0$ for every $S, R \in \mathcal{P}(\mathcal{Z})$.
2) $D(S\|R) = 0$ if, and only if, $S = R$.
3) When $\mathbf{S} = \mathbf{R} + \boldsymbol{\varepsilon}$ for some $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_{k-1})^\mathsf{T}$, the Taylor expansion of D satisfies

$$D(\mathbf{R} + \boldsymbol{\varepsilon}\|\mathbf{R}) = \frac{1}{2}\sum_{i,j=1}^{k-1} g_{ij}(\mathbf{R})\varepsilon_i\varepsilon_j + O(\|\boldsymbol{\varepsilon}\|_2^3) \tag{10}$$

as $\|\boldsymbol{\varepsilon}\|_2 \to 0$ for some $(k-1) \times (k-1)$-dimensional positive-definite matrix $G(\mathbf{R}) = [g_{ij}(\mathbf{R})]$ that depends on $\mathbf{R}$. In (10), $\|\boldsymbol{\varepsilon}\|_2$ is the Euclidean norm of $\boldsymbol{\varepsilon}$.

[1] We shall say that a function is *smooth* if it has partial derivatives of all orders.

4) Let $R \in \mathcal{P}(\mathcal{Z})$, and let $\{S_n\}$ be a sequence of distributions in $\mathcal{P}(\mathcal{Z})$ that converges to a distribution S on the boundary of $\mathcal{P}(\mathcal{Z})$. Then,

$$\varliminf_{n\to\infty} D(S_n \| R) > 0. \tag{11}$$

Remark 1: We follow the definition of divergence from the information geometry literature. In particular, according to [15, Def. 1.1], a divergence must satisfy the first three conditions in Definition 1. Often, the behavior of divergence on the boundary of $\mathcal{P}(\mathcal{Z})$ is not specified. In Definition 1, we add the fourth condition to treat the case of sequences of distributions $\{S_n\}$ that lie in $\mathcal{P}(\mathcal{Z})$ but converge to a distribution on the boundary of $\mathcal{P}(\mathcal{Z})$. Note that condition 4) is consistent with conditions 1) and 2).

Given a divergence D and $R \in \mathcal{P}(\mathcal{Z})$, consider the function $D(\cdot\|R)\colon \mathbb{R}^{k-1} \to \mathbb{R}$. By computing the partial derivatives of $D(S\|R)$ with respect to the first variable $\mathbf{S} = (S_1, \cdots, S_{k-1})^{\mathsf{T}}$, it follows from the third condition in Definition 1 that

$$D(S\|R) = (\mathbf{S}-\mathbf{R})^T \boldsymbol{A}_{D,\mathbf{R}}(\mathbf{S}-\mathbf{R}) + O(\|\mathbf{S}-\mathbf{R}\|_2^3) \tag{12}$$

as $\|\mathbf{S}-\mathbf{R}\|_2 \to 0$, where $\boldsymbol{A}_{D,\mathbf{R}}$ is the matrix associated with the divergence D at $\mathbf{R}$, which has components

$$a_{ij}(\mathbf{R}) \triangleq \frac{1}{2}\, \frac{\partial^2}{\partial S_i \partial S_j} D(S\|R)\bigg|_{S=R}, \quad i,j = 1,\cdots,k-1. \tag{13}$$

Based on $\boldsymbol{A}_{D,\mathbf{R}}$, we can introduce the notion of an *invariant divergence*.

Definition 2: Let D be a divergence, and let $R \in \mathcal{P}(\mathcal{Z})$. Then, D is said to be an *invariant divergence* on $\mathcal{P}(\mathcal{Z})$ if the matrix associated with the divergence D at $\mathbf{R}$ is of the form $\boldsymbol{A}_{D,\mathbf{R}} = \eta \boldsymbol{\Sigma}_{\mathbf{R}}$ for a constant $\eta > 0$ (possibly depending on $\mathbf{R}$) and a matrix $\boldsymbol{\Sigma}_{\mathbf{R}}$ with components

$$\boldsymbol{\Sigma}_{ij}(\mathbf{R}) = \begin{cases} \frac{1}{R_i} + \frac{1}{R_k}, & i = j \\ \frac{1}{R_k}, & i \neq j. \end{cases} \tag{14}$$

The notion of an invariant divergence is adapted from the notion of invariance of geometric structures in information geometry; see [15], [17] for more details. The matrix $\boldsymbol{\Sigma}_{\mathbf{R}}$ represents the unique invariant Riemannian metric in $\mathcal{P}(\mathcal{Z})$ with respect to the coordinate system Ξ; see [18, Eq. (47)], [5] for more details. However, in the information geometry literature, the constant η is often required to be independent of $\mathbf{R}$. Well-known divergences, such as the KL divergence, the f-divergence, and the Rényi divergence, are invariant [19]. For an invariant divergence, (12) becomes

$$D(S\|R) = \eta(\mathbf{S}-\mathbf{R})^T \boldsymbol{\Sigma}_{\mathbf{R}}(\mathbf{S}-\mathbf{R}) + O(\|\mathbf{S}-\mathbf{R}\|_2^3) \tag{15}$$

as $\|\mathbf{S}-\mathbf{R}\|_2 \to 0$, where η is a positive constant.

There are many divergences that do not satisfy (15). An example is the *squared Mahalanobis distance*, which is of the form

$$D_{\text{SM}}(S\|R) = (\mathbf{S}-\mathbf{R})^{\mathsf{T}} \boldsymbol{W}_{\mathbf{R}}(\mathbf{S}-\mathbf{R}) \tag{16}$$

for some positive-definite matrix $\boldsymbol{W}_{\mathbf{R}}$. This divergence is non-invariant if $\boldsymbol{W}_{\mathbf{R}}$ is not a constant multiple of $\boldsymbol{\Sigma}_{\mathbf{R}}$.

For a detailed list of divergences and their properties, we refer to [19, Ch. 2].

B. General Setting and Divergence Test

We consider a binary hypothesis testing problem with null hypothesis H_0 and alternate hypothesis H_1. We assume that, under hypothesis H_0, the length-n sequence Z^n of observations is i.i.d. according to $P \in \mathcal{P}(\mathcal{Z})$; under hypothesis H_1, the sequence of observations Z^n is i.i.d. according to Q, where $Q \in \mathcal{P}(\mathcal{Z}) \setminus \{P\}$.

We next define the divergence test. To this end, we first introduce the *type distribution*, which for every sequence z^n is defined as

$$T_{z^n}(a_i) \triangleq \frac{1}{n}\sum_{\ell=1}^{n} \mathbf{1}\{z_\ell = a_i\}, \quad i = 1,\ldots,k \tag{17}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

For a divergence D and a threshold $r > 0$, a *divergence test* $\mathbb{T}_n^D(r)$ for testing H_0 against the alternative H_1 is defined as follows:

Observe Z^n: if $D(T_{Z^n}\|P) < r$, then H_0 is accepted;
else H_1 is accepted.

When the divergence D is the Kullback-Leibler divergence D_{KL}, the divergence test becomes the Hoeffding test, proposed by Hoeffding in [1].

For $r > 0$, define the acceptance region for H_0 as

$$\mathcal{A}_n^D(r) \triangleq \{z^n \colon D(T_{z^n}\|P) < r\}. \tag{18}$$

Then, the type-I and the type-II errors are given by

$$\alpha_n\big(\mathbb{T}_n^D(r)\big) \triangleq P^n\big(\mathcal{A}_n^D(r)^c\big) \tag{19}$$

$$\beta_n\big(\mathbb{T}_n^D(r)\big) \triangleq Q^n\big(\mathcal{A}_n^D(r)\big). \tag{20}$$

Our goal is to analyze the asymptotic behavior of the type-II error β_n when the type-I error satisfies $\alpha_n \leq \epsilon$, $0 < \epsilon < 1$.

III. Main Results

The asymptotic behavior of the divergence test depends on the asymptotic behavior of the random variable $nD(T_{Z^n}\|P)$ in the limit as $n \to \infty$. For certain divergences, the limiting distribution of $nD(T_{Z^n}\|P)$ has been analyzed in the literature. For example, when D is the KL divergence, a well-known result by Wilks [20] states that $2nD_{\text{KL}}(T_{Z^n}\|P)$ converges in distribution to a chi-square random variable with $k-1$ degrees of freedom. This result generalizes to the α-divergence [21, Th. 3.1], [22, Th. 3]. In Lemma 1, we show that, for a general divergence D, $nD(T_{Z^n}\|P)$ converges in distribution to a *generalized chi-square random variable*, defined as follows:

Definition 3: The *generalized chi-square distribution* is the distribution of the random variable

$$\xi = \sum_{i=1}^{m} w_i \Upsilon_i \tag{21}$$

where w_i, $i = 1, \cdots, m$ are deterministic weight parameters and $\Upsilon_i, i = 1, \cdots, m$ are independent chi-square random variables with degree of freedom 1. We shall denote the generalized chi-square distribution with weight vector $\mathbf{w} = (w_1, \cdots, w_m)^\mathsf{T}$ and degrees of freedom m by $\chi^2_{\boldsymbol{w},m}$. If $w_i = 1$ for all i, then the generalized chi-square distribution becomes the chi-square distribution χ^2_m with degrees of freedom m.

Lemma 1: Let Z^n be a sequence of i.i.d. random variables distributed according to the distribution P of the null hypothesis, and let D be a divergence. Further let $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_{k-1})^\mathsf{T}$ be a vector that contains the eigenvalues of the matrix $\boldsymbol{\Sigma}_\mathbf{P}^{-1/2}\boldsymbol{A}_{D,\mathbf{P}}\boldsymbol{\Sigma}_\mathbf{P}^{-1/2}$, where $\boldsymbol{A}_{D,\mathbf{P}}$ is the matrix associated with the divergence D at $\mathbf{P}$ and the matrix $\boldsymbol{\Sigma}_\mathbf{P}$ is defined in (14). Then, the tail probability of the random variable $nD(T_{Z^n}\|P)$ satisfies

$$P^n(nD(T_{Z^n}\|P) \geq c) = \mathsf{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}(c) + O(\delta_n), \quad c \geq 0 \quad (22)$$

for some positive sequence $\{\delta_n\}$ that is independent of c and satisfies $\lim_{n\to\infty}\delta_n = 0$. Here, $\mathsf{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}(c) \triangleq \Pr(\xi \geq c)$ is the tail probability of the generalized chi-square random variable ξ with weight vector $\boldsymbol{\lambda}$ and degrees of freedom $k-1$.

Proof: Omitted due to space limitations. ■

We are now ready to present the main result of this paper:

Theorem 1: Let D be a divergence as defined in Definition 1, and let $0 < \epsilon < 1$. Further let $P, Q \in \mathcal{P}(\mathcal{Z})$ and $P \neq Q$. Recall that the cardinality of $\mathcal{Z}$ is $k \geq 2$. Then, for all sequences of thresholds $\{r_n\}$ satisfying

$$\alpha_n(\mathbb{T}_n^D(r_n)) \leq \epsilon \quad (23)$$

the divergence test $\mathbb{T}_n^D$ introduced in Section II-B satisfies

$$\begin{aligned}&\sup_{r_n\,:\,\alpha_n(\mathbb{T}_n^D(r_n))\leq\epsilon} -\ln\beta_n\big(\mathbb{T}_n^D(r_n)\big)\\ &\quad= nD_{\mathrm{KL}}(P\|Q) - \sqrt{n}\sqrt{\mathbf{c}^\mathsf{T}\boldsymbol{A}_{D,\mathbf{P}}^{-1}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)}\\ &\qquad+ O(\max\{\delta_n\sqrt{n}, \ln n\}).\end{aligned} \quad (24)$$

Here, $\boldsymbol{A}_{D,\mathbf{P}}$ is the matrix associated with the divergence D at $\mathbf{P}$; the sequence $\{\delta_n\}$ was defined in (22); $\mathbf{c} = (c_1, \cdots, c_{k-1})^\mathsf{T}$ is a vector with components

$$c_i \triangleq \ln\left(\frac{P_i}{Q_i}\right) - \ln\left(\frac{P_k}{Q_k}\right), \quad i = 1, \cdots, k-1; \quad (25)$$

and $\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}$ is the inverse of the tail probability $\mathsf{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}$ introduced in Lemma 1.

Proof: Omitted due to space limitations. ■

Remark 2: Since the sequence $\{\delta_n\}$ tends to zero as $n \to \infty$, we have that $O(\max\{\delta_n\sqrt{n}, \ln n\}) = o(\sqrt{n})$.

Corollary 1: For the class of invariant divergences, (24) in Theorem 1 becomes

$$\begin{aligned}&\sup_{r_n\,:\,\alpha_n(\mathbb{T}_n^D(r_n))\leq\epsilon} -\ln\beta_n\big(\mathbb{T}_n^D(r_n)\big)\\ &\quad= nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o(\sqrt{n}).\end{aligned} \quad (26)$$

Since the KL divergence belongs to the class of invariant divergences, it follows that (26) also characterizes the second-order performance of the Hoeffding test.

We observe from Theorem 1 that the divergence test $\mathbb{T}_n^D$ achieves the same first-order term β' as the Neyman-Pearson test, irrespective of D. In contrast, it can be shown that

$$-\sqrt{\mathbf{c}^\mathsf{T}\boldsymbol{A}_{D,\mathbf{P}}^{-1}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} < -\sqrt{V(P\|Q)}\mathsf{Q}_{\mathcal{N}}^{-1}(\epsilon). \quad (27)$$

Thus, the second-order term β'' of the divergence test $\mathbb{T}^D$ is strictly smaller than the second-order term of the Neyman-Pearson test.

In the next section, we show that there are divergences for which the divergence test outperforms the Hoeffding test for certain distributions Q of the alternate hypothesis.

IV. Second-Order Performance Comparison

In order to contrast the performances of different divergence tests, we numerically evaluate the second-order performances of $\mathbb{T}^{D_{\mathrm{KL}}}$ and $\mathbb{T}^{D_{\mathrm{SM}}}$, where D_{KL} is the KL divergence and D_{SM} is the squared Mahalanobis distance. Recall that the KL divergence is an invariant divergence. For the squared Mahalanobis distance, we shall consider (16) with $\boldsymbol{W}_\mathbf{P}$ having components

$$\boldsymbol{W}_{ij}(\mathbf{P}) = \begin{cases}\frac{1}{2P_i^2} + \frac{1}{2P_k^2}, & i = j\\ \frac{1}{2P_k^2}, & i \neq j\end{cases} \quad (28)$$

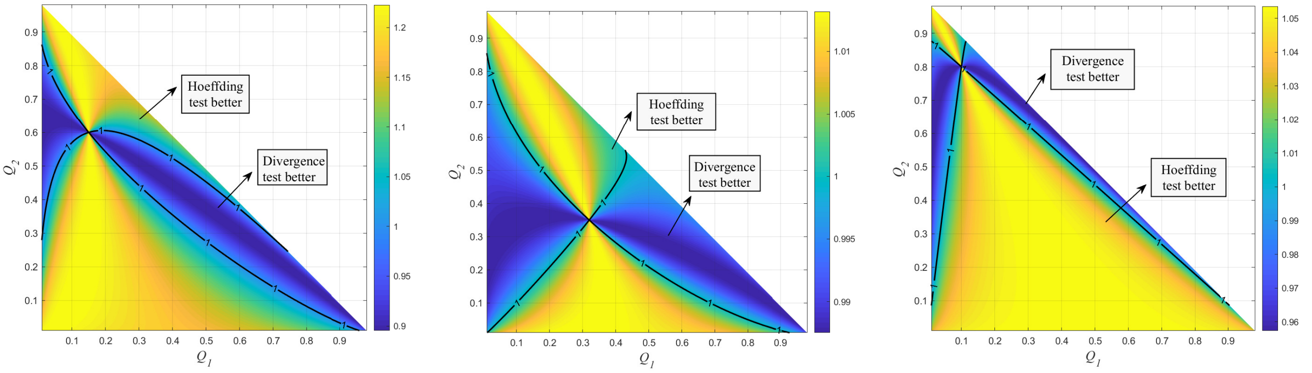
which is a non-invariant divergence. To better visualize the second-order performances, we focus on distributions with dimension $k = 3$ and represent them by the two-dimensional vectors $\mathbf{P} = (P_1, P_2)^\mathsf{T}$ and $\mathbf{Q} = (Q_1, Q_2)^\mathsf{T}$ in the coordinate space Ξ.

Since the first-order term β' of the divergence test $\mathbb{T}^D$ is not affected by the choice of D, we shall compare the second-order performances of $\mathbb{T}^{D_{\mathrm{KL}}}$ and $\mathbb{T}^{D_{\mathrm{SM}}}$ by considering the ratio of the second-order terms β'' as a function of P, Q, and ϵ:

$$\rho(P, Q, \epsilon) \triangleq \frac{\sqrt{\mathbf{c}^\mathsf{T}(\boldsymbol{W}_\mathbf{P})^{-1}\mathbf{c}}\sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},2}}(\epsilon)}}{\sqrt{V(P\|Q)}\sqrt{\mathsf{Q}^{-1}_{\chi^2_2}(\epsilon)}}. \quad (29)$$

If $\rho(P, Q, \epsilon) > 1$, then the second-order term of the divergence test is strictly smaller than the second-order term of the Hoeffding test, hence the Hoeffding test has a better second-order performance. In contrast, if $\rho(P, Q, \epsilon) < 1$, then the divergence test has a better second-order performance.

In Fig. 1, we plot the contour lines of the ratio $\rho(P, Q, \epsilon)$ as a function of $\mathbf{Q} \in \Xi$ for $\epsilon = 0.02$ and the three different null hypotheses $\mathbf{P} = (0.15, 0.6)$, $\mathbf{P} = (0.32, 0.35)$, and $\mathbf{P} = (0.1, 0.8)$. In the figure, the coordinate space Ξ is divided into two regions: one region is labeled as "Hoeffding test better" and includes the points $\mathbf{Q} \in \Xi$ for which $\rho(P, Q, \epsilon) > 1$; the other region is labeled as "Divergence test better" and includes the points $\mathbf{Q} \in \Xi$ for which $\rho(P, Q, \epsilon) < 1$. The solid contour line drawn in all three sub-figures shows all the points $\mathbf{Q} \in \Xi$ for which the Hoeffding test and the divergence



(a) $P = (0.15, 0.6, 0.25)$

(b) $P = (0.32, 0.35, 0.33)$

(c) $P = (0.1, 0.8, 0.1)$

Fig. 1: Second-order performance comparison between the Hoeffding test $\mathbb{T}^{D_{\mathrm{KL}}}$ and the divergence test $\mathbb{T}^{D_{\mathrm{SM}}}$ for the three different null hypotheses $P = (0.15, 0.6, 0.25)$, $P = (0.32, 0.35, 0.33)$, and $P = (0.1, 0.8, 0.1)$ and $\epsilon = 0.02$.

test have the same second-order performance. For each subfigure, the color bar on the right indicates the values of the ratio $\rho(P, Q, \epsilon)$.

Observe that there are distributions Q of the alternate hypothesis for which the Hoeffding test has a better second-order performance than the divergence test, and there are distributions Q for which the opposite is true. The set of distributions Q for which one test outperforms the other typically depends on the distribution P of the null hypothesis and on ϵ. Potentially, this behavior could be exploited in a composite hypothesis testing problem by tailoring the divergence D of the divergence test $\mathbb{T}^D$ to the set $\mathcal{Q}$ of possible alternate distributions.

Acknowledgment

The authors would like to thank Shun Watanabe for his helpful comments during the initial phase of this work and for pointing out relevant literature.

References

[1] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” *The Annals of Mathematical Statistics*, pp. 369–401, Apr. 1965.

[2] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2006.

[3] V. Y. Tan, “Asymptotic estimates in information theory with non-vanishing error probabilities,” in *Foundations and Trends® in Communications and Information Theory*, vol. 11, no. 1-2, pp. 1–184, 2014.

[4] K. V. Harsha, J. Ravi, and T. Koch, “Second-order asymptotics of Hoeffding-like hypothesis tests,” in *Proc. 2022 IEEE Information Theory Workshop (ITW)*, Mumbai, India, Nov. 2022, pp. 654–659.

[5] N. N. Cencov, “Statistical decision rules and optimal inference," in *Translations of Mathematical Monographs*, Vol. 53, 1982.

[6] M. Feder and N. Merhav, “Universal composite hypothesis testing: A competitive minimax approach,” *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1504–1517, Jun. 2002.

[7] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* John Wiley & Sons, 1968.

[8] O. Zeitouni, J. Ziv, and N. Merhav, “When is the generalized likelihood ratio test optimal?” *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1597–1602, Sep. 1992.

[9] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli, “Universal and composite hypothesis testing via mismatched divergence,” *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1587–1603, Mar. 2011.

[10] P. Boroumand and A. Guillén i Fàbregas, “Mismatched binary hypothesis testing: Error exponent sensitivity," *IEEE Transactions on Information Theory*, vol. 68, no. 10, pp. 6738–6761, Oct. 2022.

[11] ——, “Composite Neyman-Pearson hypothesis testing with a known hypothesis," in *Proc. 2022 IEEE Information Theory Workshop (ITW)*, Mumbai, India, Nov. 2022, pp. 131–136.

[12] S. Watanabe, “Second-order optimal test in composite hypothesis testing,” in *Proc. 2018 International Symposium on Information Theory and Its Applications (ISITA)*, Singapore, Oct. 2018, pp. 722–726.

[13] M. Gutman, “Asymptotically optimal classification for multiple tests with empirically observed statistics,” *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[14] Y. Li and V. Y. F. Tan, “Second-order asymptotics of sequential hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 7222–7230, Nov. 2020.

[15] S.-I. Amari, *Information geometry and its applications*. Springer, 2016, vol. 194.

[16] S. Eguchi, “A differential geometric approach to statistical inference on the basis of contrast functionals,” *Hiroshima Mathematical Journal*, vol. 15, no. 2, pp. 341–391, 1985.

[17] L. L. Campbell, “An extended Cencov characterization of the information metric,” *Proceedings of the American Mathematical Society*, vol. 98, no. 1, pp. 135–141, Sep. 1986.

[18] S.-I. Amari and A. Cichocki, “Information geometry of divergence functions,” *Bulletin of the Polish Academy of Sciences. Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.

[19] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[20] S. S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses,” *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, Mar. 1938.

[21] T. R. Read, “Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics,” *Annals of the Institute of Statistical Mathematics*, vol. 36, pp. 59–69, Dec. 1984.

[22] J. K. Yarnold, “Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set,” *The Annals of Mathematical Statistics*, pp. 1566–1580, Oct. 1972.

Error Probability Trade-off in Quantum Hypothesis Testing via the Nussbaum-Szkoła Mapping

Gonzalo Vazquez-Vilar[†] and Jorge Lizarribar-Carrillo
Universidad Carlos III de Madrid, Spain
Email: gvazquez@ieee.org

Abstract—**The error probability trade-off of quantum hypothesis testing is related to that of a certain surrogate classical hypothesis test via the Nussbaum-Szkoła mapping. This connection was used in the information-theoretic literature to establish the asymptotic error exponent of Bayesian quantum hypothesis testing and asymmetric quantum hypothesis testing (Hoeffding bound). In this work, we analyze the non-asymptotic gap between the error probability of a quantum test and the corresponding classical test via the Nussbaum-Szkoła mapping.**

I. Introduction

We study the problem of discriminating between two quantum states. Specifically, let us consider the density operators[1] ρ and σ, acting on some finite dimensional complex Hilbert space $\mathcal{H}$ with dimension d, and define the hypotheses

$$\mathrm{H}_0:\ \rho,\qquad \mathrm{H}_1:\ \sigma. \tag{1}$$

In this binary setting we distinguish between two error types:

- The type-I error occurs when accepting H_1 when the true state is the null hypothesis $\mathrm{H}_0\colon \rho$.
- The type-II error is the error of accepting H_0 when the true system state is $\mathrm{H}_1\colon \sigma$.

A binary test is defined by a positive self-adjoint operator Π acting on $\mathcal{H}$ such that $0 \preceq \Pi \preceq \mathbb{1}$, where $\mathbb{1}$ denotes the identity matrix and the notation $A \preceq B$ means that $B - A$ is positive semidefinite. For a test Π associated to H_1, let $\bar{\Pi} \triangleq \mathbb{1} - \Pi$. The type-I and type-II error probabilities are, respectively,

$$\alpha(\Pi) = \mathrm{Tr}[\Pi\rho], \tag{2}$$
$$\beta(\Pi) = \mathrm{Tr}[\bar{\Pi}\sigma] = 1 - \mathrm{Tr}[\Pi\sigma]. \tag{3}$$

The two error probabilities cannot be made arbitrarily small at the same time. The best achievable trade-off between these probabilities is given by the Pareto optimal boundary

$$\alpha_\beta^\star(\rho,\sigma) = \inf_{\Pi:\beta(\Pi)\le\beta} \alpha(\Pi). \tag{4}$$

When the alternatives are n-fold tensor products, i.e., $\rho \equiv \rho^{\otimes n}$ and $\sigma \equiv \sigma^{\otimes n}$, previous results established the asymptotic exponential behavior of the type-I and type-II error probabilities as $n \to \infty$. Several of these asymptotic results were obtained using a mapping, first proposed by Nussbaum and Szkoła in [1], that relates the quantum testing problem to a classical one with the same asymptotic exponential behavior.

In this work, we study the Nussbaum-Szkoła mapping in the non-asymptotic setting of fixed n. We analyze its properties and highlight the distinctions between quantum and classical testing problems through specific examples.

The organization of the remainder of the article is as follows. In Sec. II, we summarize some relevant asymptotic results and introduce the Nussbaum-Szkoła mapping. In Sec. III we present bounds on the error probability trade-off and show their tightness under certain conditions. Finally, Sec. IV closes this work with several numerical examples and some final remarks.

II. Preliminaries

For a test between the alternatives $\mathrm{H}_0\colon \rho^{\otimes n}$ and $\mathrm{H}_1 : \sigma^{\otimes n}$, we consider three significant asymptotic regimes as $n \to \infty$:

1) In a Bayesian setting with prior probabilities $\Pr[\mathrm{H}_0] = \eta$ and $\Pr[\mathrm{H}_1] = 1 - \eta$, the optimal average error probability is:

$$\epsilon_\eta^\star\big(\rho^{\otimes n}, \sigma^{\otimes n}\big) = \inf_{0\preceq\Pi\preceq\mathbb{1}} \big\{\eta\alpha(\Pi) + (1-\eta)\beta(\Pi)\big\}. \tag{5}$$

The asymptotic exponential analysis of this probability leads to the quantum Chernoff bound [1], [2] (see also [3, Sec. 3]):

$$\limsup_{n\to\infty} -\frac{1}{n}\log \epsilon_\eta^\star\big(\rho^{\otimes n}, \sigma^{\otimes n}\big) = \sup_{0\le s\le 1}\Big\{-\log \mathrm{Tr}\big[\rho^{1-s}\sigma^s\big]\Big\}. \tag{6}$$

2) In a non-Bayesian setting with a fixed type-II error β, the optimal type-I error is given by $\alpha_\beta^\star(\rho^{\otimes n}, \sigma^{\otimes n})$. Its exponential behavior corresponds to the quantum Stein's Lemma [4], [5]:

$$\limsup_{n\to\infty} -\frac{1}{n}\log \alpha_\beta^\star(\rho^{\otimes n}, \sigma^{\otimes n}) = \mathrm{Tr}\big[\sigma\big(\log\sigma - \log\rho\big)\big]. \tag{7}$$

3) Enforcing an exponential decrease in the type-II error as $\beta_n = e^{-nr}$, the Hoeffding bound asserts that [6], [7]:

$$\limsup_{n\to\infty} -\frac{1}{n}\log \alpha_{\beta_n}^\star(\rho^{\otimes n}, \sigma^{\otimes n}) = \sup_{0\le s\le 1}\left\{\frac{1}{s-1}\log\mathrm{Tr}\big[\rho^{1-s}\sigma^s\big] + \frac{s}{s-1}r\right\}. \tag{8}$$

Two important information metrics appear in these results: the quantum extension of the Renyi and the Kullback-Leibler divergences between density operators σ and ρ are defined as

$$D_s(\sigma\|\rho) \triangleq \frac{1}{s-1}\log\mathrm{Tr}\big[\rho^{1-s}\sigma^s\big], \tag{9}$$
$$D_{\mathrm{KL}}(\sigma\|\rho) \triangleq \mathrm{Tr}\big[\sigma\big(\log\sigma - \log\rho\big)\big] = \lim_{s\to1} D_s(\sigma\|\rho). \tag{10}$$

[†] G. Vazquez-Vilar is also with the Gregorio Marañón Health Research Institute, Madrid, Spain. This work has been funded in part by the Spanish Ministry for Science and Innovation under Grant PID2020116683GB-C21 / AEI / 10.13039/501100011033.

[1] Density operators are self-adjoint, positive semidefinite and have unit trace.

A. The Nussbaum-Szkoła Mapping

We consider the eigen-decomposition of the quantum states:

$$\rho = \sum_{i=1}^{d} \lambda_i \left|x_i\right\rangle \left\langle x_i\right| , \qquad \sigma = \sum_{j=1}^{d} \mu_j \left|y_j\right\rangle \left\langle y_j\right| . \tag{11}$$

The Nussbaum-Szkoła mapping transforms the states ρ and σ in two classical distributions P and Q which are defined as

$$p_{i,j} = \lambda_i \left|\langle x_i \,|\, y_j\rangle\right|^2 , \qquad q_{i,j} = \mu_j \left|\langle x_i \,|\, y_j\rangle\right|^2 , \tag{12}$$

for $i, j = 1, \ldots, d$. For this mapping, it follows that [3, Prop. 1]

$$\operatorname{Tr}\left[\rho^{1-s}\sigma^{s}\right] = \sum_{i,j} p_{i,j}^{1-s} q_{i,j}^{s}. \tag{13}$$

Then, the quantum Renyi and Kullback-Leibler divergences (9)-(10) coincide with their classical counterparts:

$$D_s(\sigma\|\rho) = D_s(Q\|P) \triangleq \frac{1}{s-1}\log\sum_{i,j} p_{i,j}^{1-s} q_{i,j}^{s}, \tag{14}$$

$$D_{\mathrm{KL}}(\sigma\|\rho) = D_{\mathrm{KL}}(Q\|P) \triangleq \sum_{i,j} q_{i,j}\left(\log q_{i,j} - \log p_{i,j}\right). \tag{15}$$

It follows that the exponential behavior of the quantum test $\rho^{\otimes n}$ v. $\sigma^{\otimes n}$ and that of the classical test $P^{\otimes n}$ v. $Q^{\otimes n}$ coincide in the three asymptotic regimes considered above. Given this (maybe) surprising property, one may wonder about how these tests compare in the non-asymptotic setting of fixed n.

III. Non-asymptotic Analysis

Let $\alpha^{\star}_{\beta}(P, Q)$ denote the error probability trade-off of a classical hypothesis test between the distributions P and Q.[2]

Theorem 1: For a binary quantum hypothesis test between states ρ and σ, and for the classical distributions P and Q defined via the Nussbaum-Szkoła mapping (12), it follows that

$$\alpha^{\star}_{\beta}(\rho, \sigma) \geq \frac{1}{2}\alpha^{\star}_{2\beta}(P, Q), \tag{16}$$

for any $\beta \in \left[0, \frac{1}{2}\right]$, and trivially $\alpha^{\star}_{\beta}(\rho, \sigma) \geq 0$ for $\beta \in \left(\frac{1}{2}, 1\right]$.

Proof: This result corresponds to [3, Prop. 2], which is stated for the average error probability in a Bayesian setting. Using the same technique, in Sec. III-A we give a direct proof for the bound on the error probability trade-off $\alpha^{\star}_{\beta}(\cdot)$. ■

The inequality (16) implies that the optimal error probability trade-off of the quantum test ρ v. σ is lower bounded by that of the classical test when both the type-I and type-II error probabilities α and β are multiplied by 1/2. Obviously, this lower bound also applies to curve of the classical test P v. Q.

Analogously, applying a change of variable $\alpha' \leftrightarrow 2\alpha$, $\beta' \leftrightarrow 2\beta$ in (16), we conclude that the optimal error probability trade-off of both the quantum test and that of the classical test is upper bounded by the quantum curve when both the type-I and type-II error probabilities are multiplied by 2.

In Sec. IV, we illustrate the accuracy of these bounds through numerical experiments. Prior to that, we prove the main result, and we show that this non-asymptotic bound is indeed tight for specific symmetric discrimination problems.

A. Proof of Theorem 1

The proof of Theorem 1 is based on the following variational formulation of the optimal trade-off $\alpha^{\star}_{\beta}(\cdot)$. For fixed $t \geq 0$, let $\Pi_t \triangleq \left\{t\sigma - \rho \geq 0\right\}$ be the projector onto the non-negative eigenspace of $t\sigma - \rho$, and $\bar{\Pi}_t \triangleq \mathbb{1} - \Pi_t$. Then, [8, Lemma 2]

$$\alpha^{\star}_{\beta}(\rho, \sigma) = \sup_{t\geq 0}\Big\{\operatorname{Tr}\big(\rho\Pi_t\big) + t\big(\operatorname{Tr}\big(\sigma\bar{\Pi}_t\big) - \beta\big)\Big\}. \tag{17}$$

Using the eigendecompositions of ρ and σ from (11), together with the cyclic property of the trace, then (17) yields

$$\alpha^{\star}_{\beta}(\rho, \sigma) = \sup_{t\geq 0}\Big\{\sum\nolimits_{i} \lambda_i \left\langle x_i\right| \Pi_t \left|x_i\right\rangle + t\sum\nolimits_{j} \mu_j \left\langle y_j\right| \bar{\Pi}_t \left|y_j\right\rangle - t\beta\Big\}. \tag{18}$$

For the projectors Π_t and $\bar{\Pi}_t$, it holds that $\Pi_t = \Pi_t\mathbb{1}\Pi_t$ and $\bar{\Pi}_t = \bar{\Pi}_t\mathbb{1}\bar{\Pi}_t$. Moreover, the identity operator can be decomposed as $\mathbb{1} = \sum_i \left|x_i\right\rangle\left\langle x_i\right| = \sum_j \left|y_j\right\rangle\left\langle y_j\right|$. Therefore, after some algebra, we shall rewrite (18) as:

$$\alpha^{\star}_{\beta}(\rho, \sigma) = \sup_{t\geq 0}\Big\{\sum\nolimits_{i,j} \lambda_i \big|\left\langle x_i\right| \Pi_t \left|y_j\right\rangle\big|^2 + t\sum\nolimits_{i,j} \mu_j \big|\left\langle x_i\right| \bar{\Pi}_t \left|y_j\right\rangle\big|^2 - t\beta\Big\}. \tag{19}$$

We group the two sums and we focus on the (i, j)-th addend

$$\begin{aligned}
&\lambda_i\big|\left\langle x_i\right| \Pi_t \left|y_j\right\rangle\big|^2 + t\mu_j\big|\left\langle x_i\right| \bar{\Pi}_t \left|y_j\right\rangle\big|^2 \\
&\quad\geq \min(\lambda_i, t\mu_j)\Big(\big|\left\langle x_i\right| \Pi_t \left|y_j\right\rangle\big|^2 + \big|\left\langle x_i\right| \bar{\Pi}_t \left|y_j\right\rangle\big|^2\Big) && (20)\\
&\quad\geq \frac{1}{2}\min(\lambda_i, t\mu_j)\left(\big|\left\langle x_i\right| \Pi_t \left|y_j\right\rangle\big| + \big|\left\langle x_i\right| \bar{\Pi}_t \left|y_j\right\rangle\big|\right)^2 && (21)\\
&\quad\geq \frac{1}{2}\min(\lambda_i, t\mu_j)\left|\left\langle x_i\right| \Pi_t \left|y_j\right\rangle + \left\langle x_i\right| \bar{\Pi}_t \left|y_j\right\rangle\right|^2 , && (22)
\end{aligned}$$

where in (20) we used that both λ_i and $t\mu_j$ are lower bounded by $\min(\lambda_i, t\mu_j)$; in (21) we defined the vector $u = [\left\langle x_i\right| \Pi_t \left|y_j\right\rangle \ \left\langle x_i\right| \bar{\Pi}_t \left|y_j\right\rangle]^T$ featuring $k = 2$ dimensions, and applied the norm inequality $\|u\|_2 \geq \frac{1}{\sqrt{k}}\|u\|_1$, $u \in \mathbb{C}^k$; and in the last step (22) we used that $|u_1| + |u_2| \geq |u_1 + u_2|$.

Applying the inequality chain (20)-(22) to the addends in (19) for each (i, j), and recalling that $\Pi_t + \bar{\Pi}_t = \mathbb{1}$, hence $\left\langle x_i\right| \Pi_t \left|y_j\right\rangle + \left\langle x_i\right| \bar{\Pi}_t \left|y_j\right\rangle = \langle x_i|y_j\rangle$, we obtain

$$\alpha^{\star}_{\beta}(\rho, \sigma) \geq \sup_{t\geq 0}\left\{\frac{1}{2}\sum\nolimits_{i,j}\min(\lambda_i, t\mu_j)\big|\langle x_i|y_j\rangle\big|^2 - t\beta\right\}. \tag{23}$$

Using the definitions of P and Q in (12), we note that

$$\min(\lambda_i, t\mu_j)\big|\langle x_i|y_j\rangle\big|^2 = p_{i,j}1_{[\lambda_i\leq t\mu_j]} + tq_{i,j}1_{[\lambda_i>t\mu_j]}, \tag{24}$$

where $1_{\mathcal{E}}$ denotes the indicator function for the event $\mathcal{E}$.

Particularizing the variational formulation (17) for ρ, σ being diagonal operators with P, Q in their diagonal, it yields:

$$\alpha^{\star}_{\beta}\left(P, Q\right) = \sup_{t\geq 0}\Big\{\sum\nolimits_{i,j} p_{i,j}1_{[p_{i,j}\leq tq_{i,j}]} + t\Big(\sum\nolimits_{i,j} q_{i,j}1_{[p_{i,j}>tq_{i,j}]} - \beta\Big)\Big\}, \tag{25}$$

Therefore, noting that for the distributions P and Q in (12), $[p_{i,j} \leq tq_{i,j}] \Leftrightarrow [\lambda_i \leq t\mu_j]$, moving the factor $\frac{1}{2}$ out of the maximization in (23) (using that $\beta = \frac{1}{2}2\beta$), we obtain the desired lower bound (16) from (23)-(24) using (25).

[2] The function $\alpha^{\star}_{\beta}(P, Q)$ coincides with (4) when ρ and σ are diagonal operators with the distributions P and Q in their respective diagonals.

B. Pure-state discrimination and symmetric error probability

We now consider a testing problem between two pure states,

$$\mathrm{H}_0: \ \rho = |x_1\rangle\langle x_1|, \tag{26}$$

$$\mathrm{H}_1: \ \sigma = |y_1\rangle\langle y_1|, \tag{27}$$

where $|x_1\rangle$ and $|y_1\rangle$ are assumed to satisfy $0 < |\langle x_1|y_1\rangle|^2 < 1$. We apply one step of the Gram-Schmidt process and define:

$$|x_2\rangle = \frac{|y_1\rangle - |x_1\rangle\langle x_1\,|\,y_1\rangle}{\||y_1\rangle - |x_1\rangle\langle x_1\,|\,y_1\rangle\|}, \tag{28}$$

$$|y_2\rangle = \frac{|x_1\rangle - |y_1\rangle\langle y_1\,|\,x_1\rangle}{\||x_1\rangle - |y_1\rangle\langle y_1\,|\,x_1\rangle\|}. \tag{29}$$

Both the orthonormal basis $\{|x_1\rangle, |x_2\rangle\}$ and $\{|y_1\rangle, |y_2\rangle\}$ span the same 2-dimensional subspace encompassing $|x_1\rangle$ and $|y_1\rangle$. If the dimension of the underlying Hilbert space is $d > 2$, the remaining eigenvectors $|x_3\rangle, \ldots, |x_d\rangle$ and $|y_3\rangle, \ldots, |y_d\rangle$ are orthogonal to both $|x_1\rangle$ and $|y_1\rangle$, and they become irrelevant in the sequel. In Fig. 1(a), we illustrate a 2-dimensional example of these bases for certain $\rho = |x_1\rangle\langle x_1|$ and $\sigma = |y_1\rangle\langle y_1|$.

1) Classical test: For the eigendecompositions of ρ and σ defined above, the Nussbaum-Szkoła mapping from (12) yields

$$p_{i,j} = \begin{cases} |\langle x_1|y_j\rangle|^2, & i=1,\ j=1,2,\\ 0, & \text{otherwise,}\end{cases} \tag{30}$$

$$q_{i,j} = \begin{cases} |\langle x_i|y_1\rangle|^2, & i=1,2,\ j=1,\\ 0, & \text{otherwise.}\end{cases} \tag{31}$$

The distributions P and Q exhibit non-overlapping supports, except in the singular case $(i,j) = (1,1)$, under which

$$p_{1,1} = q_{1,1} = |\langle x_1|y_1\rangle|^2 = \mathrm{Tr}[\rho\sigma] \triangleq a. \tag{32}$$

Here we defined $a = |\langle x_1|y_1\rangle|^2$ for future convenience.

The optimal classical test for this problem decides the correct hypothesis with no error, except when $(i,j) = (1,1)$. For this observation, in the symmetric setting, the optimal test may select between H_0 and H_1 at random with equal probability, hence incurring an error with probabilities

$$\alpha^{\mathrm{c}} = \tfrac{1}{2}p_{1,1} = \tfrac{1}{2}a, \qquad \beta^{\mathrm{c}} = \tfrac{1}{2}q_{1,1} = \tfrac{1}{2}a. \tag{33}$$

2) Quantum test: A binary test $\Pi = |x_2\rangle\langle x_2|$ does not yield a symmetric error probability in the measurement process. Neither it does the test $\Pi = |y_1\rangle\langle y_1|$. Instead, we construct a symmetric measurement $\Pi = |v_y\rangle\langle v_y|$, $\bar{\Pi} \triangleq \mathbb{1} - \Pi$, with

$$|v_x\rangle \triangleq \frac{|x_1\rangle + |y_2\rangle}{\||x_1\rangle + |y_2\rangle\|}, \tag{34}$$

$$|v_y\rangle \triangleq \frac{|y_1\rangle + |x_2\rangle}{\||y_1\rangle + |x_2\rangle\|}. \tag{35}$$

The vector $|v_x\rangle$ (resp. $|v_y\rangle$) corresponds to the normalized vector which is exactly at the midpoint between $|x_1\rangle$ and $|y_2\rangle$ (resp. between $|y_1\rangle$ and $|x_2\rangle$). It can be verified that these vectors are orthogonal, $\langle v_x|v_y\rangle = 0$, and that they define an orthonormal basis of the subspace spanned by $\{|x_1\rangle, |y_1\rangle\}$. This basis is depicted in Fig. 1(b) for illustration purposes.

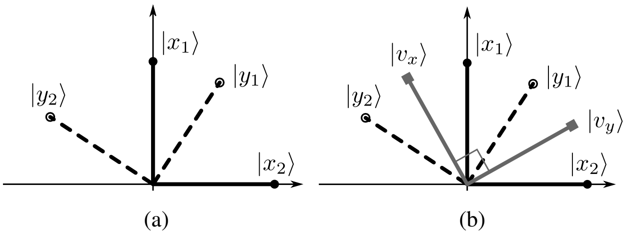


Fig. 1: Hypothesis test between pure states $|x_1\rangle$ and $|y_1\rangle$, with $0 < |\langle x_1|y_1\rangle|^2 \leq \frac{1}{2}$. (a) Basis $\{|x_1\rangle, |x_2\rangle\}$ (solid) and $\{|y_1\rangle, |y_2\rangle\}$ (dashed). (b) Orthogonal symmetric measurement $\{|v_x\rangle, |v_y\rangle\}$ (solid gray) for testing between $|x_1\rangle$ and $|y_1\rangle$.

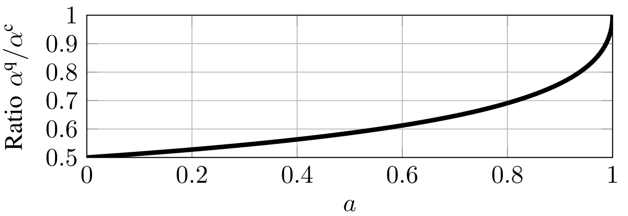


Fig. 2: Ratio between $\alpha^{\mathrm{q}} = \beta^{\mathrm{q}}$ and $\alpha^{\mathrm{c}} = \beta^{\mathrm{c}}$ for a hypothesis test between pure states $|x_1\rangle$ and $|y_1\rangle$, versus $a = |\langle x_1|y_1\rangle|^2$.

We now derive the error probabilities for this quantum test:

$$\alpha^{\mathrm{q}} = \mathrm{Tr}\big[\rho\,|v_y\rangle\langle v_y|\big], \qquad \beta^{\mathrm{q}} = \mathrm{Tr}\big[\sigma\,|v_x\rangle\langle v_x|\big]. \tag{36}$$

We first note that

$$\mathrm{Tr}\big[\rho\,|v_y\rangle\langle v_y|\big] = \mathrm{Tr}\big[\,|x_1\rangle\langle x_1| \cdot |v_y\rangle\langle v_y|\big] = |\langle x_1|v_y\rangle|^2, \tag{37}$$

and, using (35), we write

$$|\langle x_1|v_y\rangle|^2 = \frac{\big|\langle x_1|y_1\rangle + \langle x_1|x_2\rangle\big|^2}{\||y_1\rangle + |x_2\rangle\|^2} = \frac{a}{2\big(1+\sqrt{1-a}\big)}. \tag{38}$$

In the last step we used that $\langle x_1|x_2\rangle = 0$ and $\big|\langle x_1|y_1\rangle\big|^2 = a$; and then we used (28) to obtain, after some straightforward algebra, that $\||y_1\rangle + |x_2\rangle\|^2 = 2\big(1+\sqrt{1-a}\big)$.

According to (36)-(38) and given the symmetry of the problem, the type-I and type-II error probabilities are thus

$$\alpha^{\mathrm{q}} = \beta^{\mathrm{q}} = \frac{a}{2\big(1+\sqrt{1-a}\big)}, \tag{39}$$

which depend only on $a = |\langle x_1|y_1\rangle|^2$.

Figure 2 shows the ratio between $\alpha^{\mathrm{q}} = \beta^{\mathrm{q}}$ and the classical error probability $\alpha^{\mathrm{c}} = \beta^{\mathrm{c}} = \frac{1}{2}a$ as a function of a. We observe that, as a tends to 0 (i.e., states $|x_1\rangle$ and $|y_1\rangle$ approaching orthogonality), this ratio tends to $\frac{1}{2}$. Indeed, using the Taylor expansion of $f(a) \triangleq \frac{a}{2(1+\sqrt{1-a})}$ around $a = 0$, it yields

$$\alpha^{\mathrm{q}} = \beta^{\mathrm{q}} = \tfrac{1}{4}a + o(a), \tag{40}$$

where $o(a)$ satisfies $\lim_{a\to 0}\frac{o(a)}{a} = 0$ (little-o notation).

Therefore, up to a vanishing term $o(a)$, the quantum error probabilities $\alpha^{\mathrm{q}} = \beta^{\mathrm{q}}$ coincide with the lower bound from Theorem 1, given by $\frac{1}{2}\alpha^{\mathrm{c}} = \frac{1}{2}\beta^{\mathrm{c}}$. We conclude that the bound in Theorem 1 is tight in certain scenarios, even when $n = 1$.

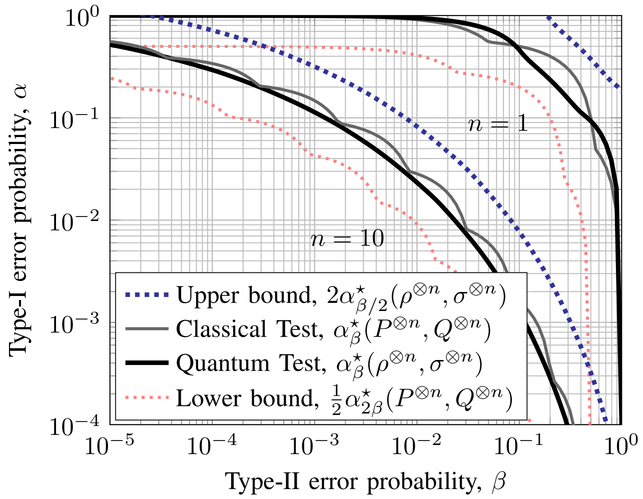


Fig. 3: Error probability trade-off of a hypothesis test between mixed states $\rho^{\otimes n}$ and $\sigma^{\otimes n}$, with ρ and σ given in (41).

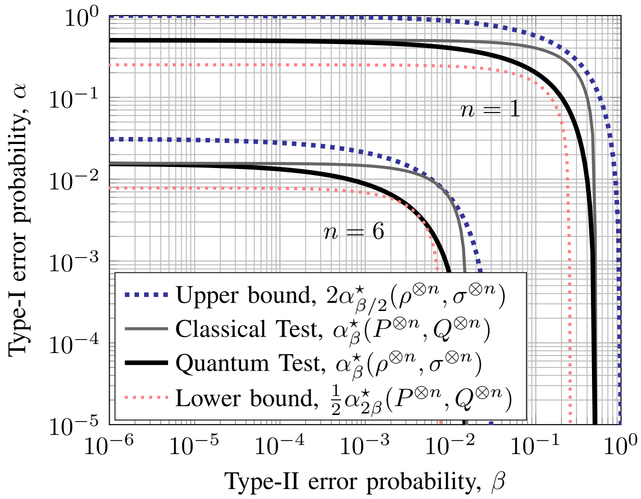


Fig. 4: Error probability trade-off of a hypothesis test between pure states $\rho^{\otimes n}$ and $\sigma^{\otimes n}$, with ρ and σ from (42) when $\phi = \frac{\pi}{4}$.

IV. Numerical Results and Conclusions

We now compare the type-I and type-II error probability trade-off of a quantum hypothesis test with that of the classical hypothesis test resulting from the Nussbaum-Szkoła mapping.

A. Mixed-state discrimination

Consider the quantum states defined by the density operators

$$\rho = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.1 \end{bmatrix}, \qquad \sigma = \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}. \tag{41}$$

Both ρ and σ are mixed states with overlapping supports.

Figure 3 compares the error probability trade-off of the quantum hypothesis test $\rho^{\otimes n}$ v. $\sigma^{\otimes n}$ with that of the surrogate classical test $P^{\otimes n}$ v. $Q^{\otimes n}$ defined via the Nussbaum-Szkoła mapping (11)-(12). Even when $n = 1$ both curves exhibit a similar behavior. Note that $P^{\otimes n}$ and $Q^{\otimes n}$ correspond to discrete distributions defined over d^{2n} points, hence their staggered shape when depicted in logarithmic scale (due to the corresponding affine segments in linear scale).

For comparison, we also depict the upper and lower bounds that follow from Theorem 1. Note that in general these bounds are not tight. Moreover, the gap between the upper and lower bound (when plotted in logarithmic scale) is approximately constant with n, as it could be expected due to the multiplicative nature of the bounds that follow from Theorem 1.

B. Pure-state discrimination

We now consider two pure states defined by

$$\rho = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \sigma = \begin{bmatrix} \cos(\phi)^2 & \cos(\phi)\sin(\phi) \\ \sin(\phi)\cos(\phi) & \sin(\phi)^2 \end{bmatrix}. \tag{42}$$

Figure 4 shows the error probability trade-off for the test $\rho^{\otimes n}$ v. $\sigma^{\otimes n}$ with $n = 1$ and $n = 6$, for the states in (42) with $\phi = \frac{\pi}{4}$. For $n = 1$ the inner product between the two states is $a = \mathrm{Tr}[\rho\sigma] = \frac{1}{2}$, and the gap from the error trade-off to the upper and lower bounds is still significant. As the value of n increases, the pure states $\rho^{\otimes n}$ and $\sigma^{\otimes n}$ exhibit a growing degree of orthogonality. In this regime, as shown in Sec. III-B, the lower and upper bound become increasingly tight for the symmetric error probability. This is apparent from Fig. 4, since for $n = 6$ (i.e., $a \approx 0.0156$) the gap between the error curves and the bounds becomes negligible in the region where $\alpha \approx \beta$.

The Nussbaum-Szkoła mapping transforms a hypothesis test between two quantum states into a test between two classical probability distributions. While this mapping was primarily used to study the asymptotics of quantum hypothesis testing as $n \to \infty$, it also approximates its non-asymptotic performance for fixed n. In this work we examine and illustrate the gap between the error probability trade-off of the quantum and classical hypothesis tests in certain settings of interest, laying the groundwork for potential future research in this direction.

References

[1] M. Nussbaum and A. Szkoła, "A lower bound of Chernoff type in quantum hypothesis testing," *arXiv preprint quant-ph/0607216*, 2006.

[2] K. M. R. Audenaert, J. Calsamiglia, R. Munoz-Tapia, E. Bagan, L. Masanes, A. Acin, and F. Verstraete, "Discriminating states: The quantum Chernoff bound," *Physical Review Letters*, vol. 98, no. 16, p. 160501, 2007.

[3] K. M. R. Audenaert, M. Nussbaum, A. Szkoła, and F. Verstraete, "Asymptotic error rates in quantum hypothesis testing," *Communications in Mathematical Physics*, vol. 279, no. 1, pp. 251–283, 04 2008. [Online]. Available: https://doi.org/10.1007/s00220-008-0417-5

[4] F. Hiai and D. Petz, "The proper formula for relative entropy and its asymptotics in quantum probability," *Communications in Mathematical Physics*, vol. 143, no. 1, pp. 99–114, 1991.

[5] T. Ogawa and M. Hayashi, "On error exponents in quantum hypothesis testing," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1368–1372, 2004.

[6] H. Nagaoka, "The converse part of the theorem for quantum Hoeffding bound," *arXiv preprint quant-ph/0611289*, 2006.

[7] M. Hayashi, "Error exponent in asymmetric quantum hypothesis testing and its application to classical-quantum channel coding," *arXiv preprint quant-ph/0611013*, 2006.

[8] A. B. Coll, G. Vazquez-Vilar, and J. R. Fonollosa, "Generalized perfect codes for symmetric classical-quantum channels," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5923–5936, Sep. 2022.

Graph-based Algorithms for Linear Computation Coding

Hans Rosenberger*, Ali Bereyhi†, Ralf R. Müller*
*Institute for Digital Communications, Friedrich-Alexander-Universität (FAU), Erlangen, Germany
{hans.rosenberger, ralf.r.mueller}@fau.de
†Wireless Computing Lab, University of Toronto, Canada
ali.bereyhi@utoronto.ca

***Abstract*—We revisit existing linear computation coding (LCC) algorithms, and introduce a new framework that measures the computational cost of computing multidimensional linear functions, not only in terms of the number of additions, but also with respect to their suitability for parallel processing. Utilizing directed acyclic graphs, which correspond to signal flow graphs in hardware, we propose a novel LCC algorithm that controls the trade-off between the total number of operations and their parallel executability. Numerical evaluations show that the proposed algorithm, constrained to a fully parallel structure, outperforms existing schemes.**

I. Introduction

Over-parameterized neural networks (NNs) have achieved many of the recent advancements in improving inference accuracy. Many real-world applications of these very large NNs require both real-time inference and operate in a resource constrained environment. It is therefore of great importance to implement them with minimal computational complexity. Various research efforts have been directed towards improving NN efficiency, including pruning, knowledge distillation, quantization and NN-hardware co-design [1], [2].

Linear computation coding (LCC) introduces an analytical framework that invokes the idea of sparse matrix decomposition to reduce the computational cost of computing matrix-vector products, i.e. the lossy compression of a multidimensional linear function with constant coefficients. Earlier studies on LCC mainly focus on the number of additions as the metric of computational complexity [3]–[6]. Though important, this metric is not the only concern in many applications.

In this paper, we revisit the earlier LCC studies from a new perspective on computational complexity, in which not only the number of operations, but also their order matters. Our interest follows from a simple fact: optimizing the order in which the operations are carried out enables us to fully exploit the potential of *parallel processing*. We use the notion of a directed acyclic graph (DAG), closely corresponding to the signal flow graph of a hardware implementation, to develop a new LCC algorithm. The proposed scheme explicitly tunes the structure of the DAG and outperforms existing algorithms on parallel processing units.

A. Notation

Vectors and matrices are denoted by lower- and upper-case boldface letters, e.g. $\boldsymbol{x}$ and $\boldsymbol{X}$, respectively. The Euclidean and Frobenius norms are shown by $\|\cdot\|_2$ and $\|\cdot\|_\mathrm{F}$, respectively. The matrix transpose is denoted by $(\cdot)^\mathrm{T}$. The augmented identity matrix with dimension $N \times K$ is denoted by $\boldsymbol{I}_{N\times K}$, and the j-th row unit vector in K dimensions by $\mathbf{1}_{j,K}$. The function $\mathrm{supp}(\boldsymbol{x})$ returns the indices in the support of $\boldsymbol{x}$, i.e. the set of all indices i where $x_i \neq 0$.

Sets are specified by upper case caligraphic letters, e.g. $\mathcal{A}$. We use the notation $|\mathcal{A}|$ to represent the cardinality of $\mathcal{A}$. A DAG is denoted by $D = (\mathcal{C}, \mathcal{A})$, where $\mathcal{C} \subset \mathbb{R}^{1\times K}$ is the ordered set of all vertices and $\mathcal{A}$ the set of arcs (directed edges). The indegree and outdegree of a vertex $\boldsymbol{c} \in \mathcal{C}$ are denoted by $\mathrm{d}^-_\mathrm{D}(\boldsymbol{c})$ and $\mathrm{d}^+_\mathrm{D}(\boldsymbol{c})$, respectively. Given a DAG $D = (\mathcal{C}, \mathcal{A})$ and a vertex $\boldsymbol{c} \in \mathcal{C}$, $\mu_\mathrm{D}(\boldsymbol{c})$ denotes the depth of $\boldsymbol{c}$, i.e. the longest path from any node $\boldsymbol{c}' \in \mathcal{C}$ to node $\boldsymbol{c}$. The operator $\mathrm{mat}(\cdot)$ converts a vertex set $\mathcal{C} = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_L\} \subset \mathbb{R}^{1\times K}$ with $|\mathcal{C}| = L$ to its corresponding matrix, i.e. $\boldsymbol{C} = \mathrm{mat}(\mathcal{C}) = [\boldsymbol{c}_1, \dots, \boldsymbol{c}_L] \in \mathbb{R}^{L\times K}$. Unless otherwise specified, $\boldsymbol{c}_i$ denotes the i-th element in the set $\mathcal{C}$ or the i-th row vector of the corresponding matrix $\boldsymbol{C} = \mathrm{mat}(\mathcal{C})$. The notation $[N]$ is an abbreviation for the set $\{1, \dots, N\}$.

II. Preliminaries

Consider the matrix vector product

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{x} \tag{1}$$

with the arbitrary, but constant, matrix $\boldsymbol{T} \in \mathbb{R}^{N\times K}$ and the arbitrary input vector $\boldsymbol{x} \in \mathbb{R}^{K\times 1}$. Our goal is to approximately compute $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ with minimum effort. Calculating the matrix-vector product straightforwardly requires NK multiplications and $N(K-1)$ additions. Using a finite-precision representation of $\boldsymbol{T}$, a multiplication can be reduced to additions and bitshifts. Quantizing the matrix entries independently, it is well known that each additional bit on average improves the signal to quantization noise ratio (SQNR) by $6\,\mathrm{dB}$ while requiring half an extra addition. Using the canonically signed digit (CSD) representation [7], i.e. allowing for subtractions as well, the SQNR even improves by $14.5\,\mathrm{dB}$ per digit. However, by quantizing the operations of a matrix-vector product jointly, far larger gains are possible [3], [8].

A. Addition as a Fundamental Operation

Definition 1 (Fundamental Operation): Let $\mathcal{C} \subset \mathbb{R}^{1\times K}$ denote a set of L vectors and be called a codebook. We define the fundamental operation as the linear combination of at most S vectors contained in $\mathcal{C}$, or, more formally:

$$\mathrm{add}_S(\boldsymbol{\omega}_S, \mathcal{C}) = \boldsymbol{\omega}_S\, \mathrm{mat}(\mathcal{C}) \tag{2}$$

This work was supported by Deutsche Forschungsgemeinschaft (DFG) under the project Computation Coding (MU-3735-/8-1).

with $\boldsymbol{\omega}_S \in \mathcal{W}_S$, where

$$\mathcal{W}_S = \left\{ \boldsymbol{\omega} = \sum_{s=1}^{S} i_s \mathbf{1}_{j_s,L} : i_s \in \mathcal{M} \subseteq \{0, \pm 2^{\mathbb{Z}}\}, j_s \in [L] \;\; \forall s \right\}. \tag{3}$$

The nonzero coefficients of $\boldsymbol{\omega}_S \in \mathcal{W}_S$ are restricted to the set of (sums of) signed powers of two, corresponding only to bitshifts in hardware, which can be considered computationally cheap.[1] The computational cost of a fundamental operation is governed by the at most $S-1$ additions needed to form the linear combination.

Given a codebook $\mathcal{C}$ and using the notion of the fundamental operation, our aim is now to approximate a target vector $\boldsymbol{t}$ by a single fundamental operation. We call this objective wiring. Mathematically we aim to solve the following least squares (LS) problem:

$$w(\boldsymbol{t}, \mathcal{C}, S) = \underset{\boldsymbol{\omega}_S \in \mathcal{W}_S}{\operatorname{argmin}} \|\boldsymbol{t} - \boldsymbol{\omega}_S \operatorname{mat}(\mathcal{C})\|_2, \tag{4}$$

which can be equivalently seen as a sparse recovery problem [9] due to the restricted support of $\boldsymbol{\omega}_S$.

The minimization over the set of discrete vectors $\mathcal{W}_S$ in (4) is an NP-hard problem. Hence, an optimal solution is generally computationally intractable. Therefore, we resort to the following two suboptimal approaches:

- *Discrete matching pursuit (DMP)* [3]: Start with $\boldsymbol{\omega} \leftarrow \mathbf{0}$. Find the vector in $\boldsymbol{c}_i \in \mathcal{C}$ scaled by a signed power of two that reduces the error to $\boldsymbol{t}$ maximally and update $\boldsymbol{\omega}$ in the i-th component. Repeat S times.
- *Reduced state (RS) approach* [5]: Procedure similar to DMP. However, instead of choosing in each iteration the best vector minimizing the error, we retain a list of the Q best linear combinations in each iteration and choose the combination with minimum error at termination. This procedure enables a performance close to full search at a reasonable time complexity [5].

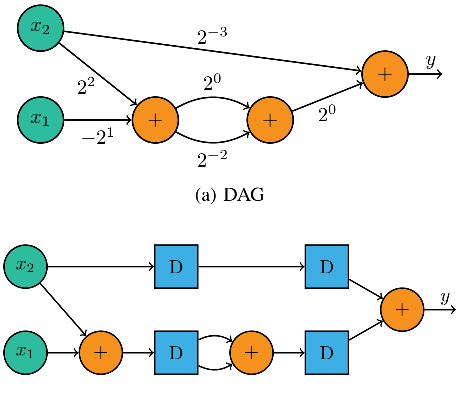
To quantify the ability of a codebook $\mathcal{C}$ to approximate the matrix $\boldsymbol{T}$ with row vectors $\boldsymbol{t}_n$, we use the SQNR defined as

$$\operatorname{SQNR}(\boldsymbol{T}, \mathcal{C}) = \frac{\|\boldsymbol{T}\|_{\mathrm{F}}^2}{\sum_{n=1}^{N} \|\boldsymbol{t}_n - w(\boldsymbol{t}_n, \mathcal{C}, 1) \operatorname{mat}(\mathcal{C})\|_2^2}. \tag{5}$$

Note that $w(\boldsymbol{t}_n, \mathcal{C}, 1) \operatorname{mat}(\mathcal{C})$ finds the vector in $\mathcal{C}$ scaled by a signed power of two, that approximates $\boldsymbol{t}_n$ best. As $S=1$, this is only a selection and potentially a bitshift, no additions are required.

B. Constant matrix vector multiplication (CMVM)

Using the notion of a fundamental operation, any matrix-vector product with finite precision can now be expressed as a DAG with K input and N output vertices. Input vertices are all vertices with no preceding fundamental operations, i.e. $\{\boldsymbol{c} \in \mathcal{C} | \, \mathrm{d}_{\mathrm{D}}^{-}(\boldsymbol{c}) = 0\}$. Likewise, output vertices have no arcs directed to subsequent vertices ($\{\boldsymbol{c} \in \mathcal{C} | \, \mathrm{d}_{\mathrm{D}}^{+}(\boldsymbol{c}) = 0\}$). In such a graph, each vertex, except the input vertices, corresponds to one fundamental operation, and each directed arc is labeled with a signed power of two. An example of such a DAG is depicted in Fig. 1a. It is our goal, given some target matrix $\boldsymbol{T}$, to find a DAG requiring a minimum of computations given some fidelity constraint. We can therefore define now a CMVM problem.



(a) DAG

(b) Pipelined DAG

Fig. 1: A DAG realizing the function $y(x_1, x_2) = (21/8)x_2 - (5/4)x_1$ is depicted in (a). The same DAG is extended in (b) with delay elements to allow for pipelining.

Definition 2 (CMVM Problem): For all fundamental operations assume without loss of generality $S=2$. Given a target matrix $\boldsymbol{T}$ and some positive parameter ϵ, find a DAG $D = (\mathcal{C}, \mathcal{A})$ with vertex set $\mathcal{C} \subset \mathbb{R}^{1 \times K}$, that solves

$$\begin{aligned} \min \quad & |\mathcal{C}| && \text{(6a)} \\ \text{s.t.} \quad & \operatorname{SQNR}(\boldsymbol{T}, \mathcal{C}) > \epsilon && \text{(6b)} \\ & \boldsymbol{c}_m = \mathbf{1}_{m,K} \qquad \forall m \in [K] && \text{(6c)} \\ & \boldsymbol{c}_l = \operatorname{add}_2(\boldsymbol{\omega}_2, \{\boldsymbol{c}_i \in \mathcal{C} : i \in [l-1]\}) \quad \forall l > K && \text{(6d)} \end{aligned}$$

The CMVM problem is at least NP-complete. Similar to multiple constant multiplication (MCM) [10], it is an even broader generalization of the single constant multiplication (SCM) problem,[2] which is known to be NP-complete [11], [12]. Hence, by polynomial reduction the CMVM problem has to be at least as difficult. As the optimal solution is generally computationally intractable, we focus for the remainder of this paper on the development of efficient heuristics for obtaining decomposition DAGs.

Remark 1: Throughout the paper we do not specify the set of arcs $\mathcal{A}$ of a DAG explicitly for reasons of brevity. As new vertices are created from an initial codebook, i.e. the set of unit vectors, by means of fundamental operations, implicitly $\mathcal{A}$ is defined uniquely[3] by $\mathcal{C}$ for any decomposition DAG $D = (\mathcal{C}, \mathcal{A})$ as well.

C. Computational Cost

Three terms contribute to the overall computational cost

$$C_{\text{total}} = C_{\text{add}} N_{\text{add}} + C_{\text{delay}} N_{\text{delay}} + C_{\text{inv}} N_{\text{inv}}. \tag{7}$$

The number of additions N_{add}, the number of delay elements (latches) N_{delay} and the number of sign inverters N_{inv} required. Further, C_{add}, C_{delay} and C_{inv} are the effective cost for an

[1] In this paper we consider the set of wiring coefficients to be unrestricted, i.e. $\mathcal{M} = \{0, \pm 2^{\mathbb{Z}}\}$. For some applications, it is beneficial to restrict the coefficients to a subset. Efficient strategies for such cases are investigated in [6].

[2] The optimization of the multiplication of a constant scalar to a scalar variable.

[3] Uniqueness only refers in that context to the start and endpoint of individual arcs, not their labeling. For example two different fundamental operations, differing in their labeling/bitshift, might produce the same result, i.e. $\boldsymbol{c}_2 = \boldsymbol{c}_1 - 1/4\boldsymbol{c}_1 = 1/2\boldsymbol{c}_1 + 1/4\boldsymbol{c}_1$.

addition, a delay element and an inverter, respectively. Inspired by the CMOS implementation of these basic functions, we assume for simplicity that the cost for an adder and a delay element are approximately equal and set to[4] $C_{\mathrm{add}} = C_{\mathrm{delay}} = 20$. For an inverter we assume a cost of $C_{\mathrm{inv}} = 2$, since these can be easily implemented by two transistors [13].

The number of additions in computing a DAG is upper bounded, as zeros are allowed for coefficients as well, by

$$N_{\mathrm{add}} = \sum_{i=K+1}^{|\mathcal{C}|} \left(\mathrm{d}_{\mathrm{D}}^{-}(\boldsymbol{c}_i) - 1\right) \tag{8a}$$

$$\overset{\text{(a)}}{=} (|\mathcal{C}| - K)(S-1) \tag{8b}$$

where (a) follows from the fact that the number of additions $S-1$ for all vertices is constant.

For medium to large matrices it may not be desirable to straightforwardly implement the DAG in hardware, apply a realisation of $\boldsymbol{x}$, and wait for the output $\boldsymbol{y}$ to be computed. Particularly, for a DAG with many logical operations in sequence, this may take some time and is not an optimal use of resources. Instead, a pipelined approach is desirable,[5] each adder is followed by a latch or delay element that is able to store the intermediate result produced by that adder. For example, after an addition is completed, and the result is stored, the following input realization can already be forwarded to the adder. The stored result is then forwarded to the subsequent adder. The schematic of a pipelined design is depicted in Fig. 1b. There, a pipelined signal flow graph/DAG with two inputs x_1 and x_2 computes a single output y. The second input is required for the final addition computing the output. Thus, two additional delay elements are required in the upper branch to delay the input accordingly, adding to the overall hardware cost.

Pipelining largely improves overall throughput, keeping each adder busy and reducing idle times of resources. However, to enable that, idle paths require additional delay elements that contribute to the overall hardware cost. Hence, for a practical algorithm it is desirable to not only minimize the number of adders but to find a DAG structure that limits the number of delay elements. The overall number of delay elements required for a pipelined implementation of a decomposition DAG can be computed by

$$N_{\mathrm{delay}} = N_{\mathrm{add}} + \sum_{\forall \tilde{\boldsymbol{c}} \in \tilde{\mathcal{C}}} \left(\max_{\boldsymbol{c} \in \mathcal{D}(\tilde{\boldsymbol{c}})} \mu_{\mathrm{D}}(\boldsymbol{c}) - \mu_{\mathrm{D}}(\tilde{\boldsymbol{c}}) - 1 \right) \tag{9a}$$

with

$$\tilde{\mathcal{C}} = \left\{ \boldsymbol{c} \in \mathcal{C} \,|\, \mathrm{d}_{\mathrm{D}}^{+}(\boldsymbol{c}) > 0 \right\} \tag{9b}$$

$$\mathcal{D}(\tilde{\boldsymbol{c}}) = \left\{ \boldsymbol{c} \in \mathcal{C} \,|\, (\tilde{\boldsymbol{c}}, \boldsymbol{c}) \in \mathcal{A} \right\} \tag{9c}$$

The set $\mathcal{D}(\tilde{\boldsymbol{c}})$ contains all vertices $\boldsymbol{c}$ that are connected by a directed arc in $\mathcal{A}$ from $\tilde{\boldsymbol{c}}$ to $\boldsymbol{c}$. The total number of delay elements is the sum of the number of adders, as each adder needs a buffer at the output, and for each node with outgoing arcs the longest path difference minus one that needs to be equalized.

The number of inverters depends on the specific algorithm used. For brevity, we will not discuss inverters in detail. A reduction algorithm for the number of inverters in parallel LCC algorithms is discussed in [15].

III. Algorithmic Approaches

We now discuss two existing algorithmic approaches for LCC, namely a fully sequential and fully parallel algorithm. Utilizing the best of both worlds, we introduce a new mixed algorithm (MA) that enables us to tune the DAG structure for further analysis.

A. Fully sequential (FS) Algorithm

Given the set of all unit vectors in K dimensions as our initial codebook set $\mathcal{C} = \{\mathbf{1}_{1,K}, \ldots, \mathbf{1}_{K,K}\}$, we recursively add vertices to the DAG using the following update rule [4]:

$$\mathcal{C} \leftarrow \mathcal{C} \cup \{ w(\boldsymbol{t}_{\tilde{n}}, \mathcal{C}, S) \operatorname{mat}(\mathcal{C}) \}. \tag{10}$$

This means that we find the best linear combination of vectors in $\mathcal{C}$ that approximates $\boldsymbol{t}_{\tilde{n}}$ well and requires $S-1$ additions. We choose the row vector with index $\tilde{n}$ from $\boldsymbol{T}$ that provides us with the largest reduction of the squared error for the update:

$$\tilde{n} = \operatorname*{argmin}_{n \in [N]} \Bigg(\| \boldsymbol{t}_n - w(\boldsymbol{t}_n, \mathcal{C}, S) \operatorname{mat}(\mathcal{C}) \|_2^2 + \sum_{k \neq n} \| \boldsymbol{t}_k - w(\boldsymbol{t}_k, \mathcal{C}, 1) \operatorname{mat}(\mathcal{C}) \|_2^2 \Bigg) \tag{11}$$

Although this approach shows excellent performance when looking at the tradeoff between distortion and the number of additions required, it is in many cases not suited for pipelining. This follows from the fact that any S vertices in a given codebook can be combined in each iteration, the obtained graph has an arbitrary structure (c.f. Fig. 2a). Assuming for simplicity that each fundamental operation takes time t_{f} to compute,[6] it is concluded that the delay at any node $\boldsymbol{c}$ is $\mu_{\mathrm{D}}(\boldsymbol{c}) t_{\mathrm{f}}$. Thus, if the depth $\mu_{\mathrm{D}}(\boldsymbol{c})$ varies in $\boldsymbol{c}$, delays are introduced that need to be compensated for. The additional hardware resources and overhead required by the FS algorithm are typically not acceptable, especially for large matrices. Therefore, algorithms that take these hardware constraints into account are desirable.

B. Fully parallel (FP) algorithm

Instead of performing updates sequentially, we now successively refine the codebook for all vectors of the target matrix in parallel and then forget the old codebook. Such a fully parallel algorithm can be written as a product of matrices [3]:

$$\boldsymbol{T} \approx \boldsymbol{W}_L \boldsymbol{W}_{L-1} \cdots \boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{C}_0. \tag{12}$$

The n-th row of the l-th matrix factor $\boldsymbol{W}_l$ is recursively obtained by

$$\boldsymbol{w}_{l,n} = w(\boldsymbol{t}_n, \boldsymbol{C}_{l-1}, S) \quad \forall n \in [N] \tag{13}$$

with

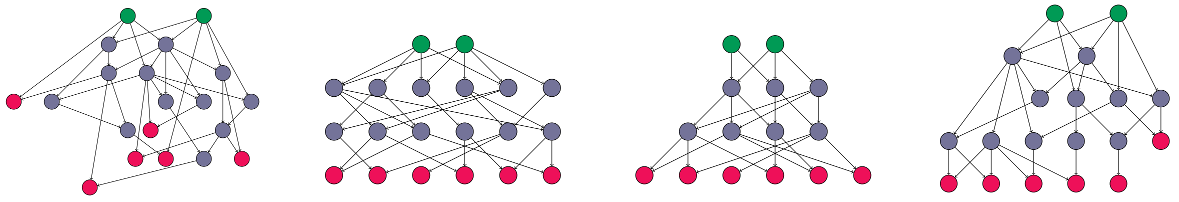
$$\boldsymbol{C}_{l-1} = \boldsymbol{W}_{l-1} \boldsymbol{W}_{l-2} \cdots \boldsymbol{W}_2 \boldsymbol{W}_1 \boldsymbol{C}_0. \tag{14}$$

Each layer l refines the approximation for each $\boldsymbol{t}_n$ using the codebook obtained in the previous iteration $l-1$. Using our DAG based interpretation, this is the same as effectively

[4] The cost of a full adder ranges around 20 transistors and can vary depending on the specific implementation used, clock speed, etc. This cost only considers a full adder for the addition of two inputs of a single bit. For larger bitwidths the cost scales accordingly and simplifications in the implementation are possible. For simplicity we only consider the cost per bit.

[5] For a detailed discussion of pipelining, refer to [14].

[6] This assumption is valid as long as we use the same type of adder throughout a DAG, i.e. S is fixed.



(a) Fully Sequential Algorithm (b) Fully Parallel Algorithm (c) Mixed Alg. ($\Delta\mu_{\max} = 0$) (d) Mixed Alg. ($\Delta\mu_{\max} = 1$)

Fig. 2: Resulting graph topologies of different algorithmic approaches for decomposing a target matrix $\boldsymbol{T}$ of dimension 6 × 2. Green nodes represent input vertices corresponding to elements of the input vector $\boldsymbol{x}$, red nodes represent output vertices of the resulting matrix-vector product $\boldsymbol{y}$ and blue nodes are intermediary vertices of the decomposition graph.

restricting the codebook used in iteration l to the subset of vectors in $\mathcal{C}$ at depth $l-1$, i.e. the matrix $\boldsymbol{C}_{l-1}$ contains all row vectors that are also included in the set $\{\boldsymbol{c} \in \mathcal{C} | \mu_{\mathrm{D}}(\boldsymbol{c}) = l-1\}$. As for the FS algorithm, we use the set of all unit vectors as the initial codebook $\boldsymbol{C}_0 = \boldsymbol{I}_{N\times K}$. The structure of the DAG generated by this algorithm is depicted in Fig. 2b. Compared to the FS algorithm, a fully parallel implementation in hardware can be achieved, no delays by differing path lengths are introduced. However, this algorithm is not without drawbacks. First, previous work [5] showed that refinement of the initial codebook during the first few iteration comes with a drop in performance. Second, this algorithm does not scale to arbitrarily small matrices. As the effective codebook scales with the target matrix size this can lead to convergence issues when decomposing smaller matrices.

C. Mixed algorithm (MA)

Using the ideas of the FS and FP algorithms we introduce a new MA enabling us tune the structure of the computation DAG. We reuse the sequential update rule from the FS algorithm in (10) to update $\mathcal{C}$.

$$\tilde{n} = \underset{n\in[N]}{\operatorname{argmin}}\, \lambda_n \Bigg(\| \boldsymbol{t}_n - w(\boldsymbol{t}_n, \mathcal{C}, S) \operatorname{mat}(\mathcal{C}) \|_2^2 + \sum_{k\neq n} \| \boldsymbol{t}_k - w(\boldsymbol{t}_k, \mathcal{C}, 1) \operatorname{mat}(\mathcal{C}) \|_2^2 \Bigg) \tag{15a}$$

with

$$\lambda_n = \max_{j\in\mathcal{S}} \mu_{\mathrm{D}}(\boldsymbol{c}_j) \quad \text{and} \quad \mathcal{S} = \operatorname{supp}\left(w(\boldsymbol{t}_n, \mathcal{C}, S)\right).$$

To obtain the index $\tilde{n}$ for the target vector to be approximated, we amend the objective to update the approximation with the largest drop in error in (11) by a multiplicative penalty factor λ_n. This factor penalizes the absolute depth of approximations for different target vectors, i.e. updating a codeword at a higher depth leads to a larger penalty. Moreover, to be able to limit the number and depth of idle paths in the DAG, we introduce a side constraint limiting the difference in depth for any linear combination of codewords, which is

$$\max_{j\in\mathcal{S}} \left(\mu_{\mathrm{D}}(\boldsymbol{c}_j)\right) - \min_{j\in\mathcal{S}} \left(\mu_{\mathrm{D}}(\boldsymbol{c}_j)\right) \leq \Delta\mu_{\max}. \tag{15b}$$

The parameter $\Delta\mu_{\max}$ controls the maximum difference in depth for the codewords used in each update. For $\Delta\mu_{\max} \to \infty$ and $\lambda_n = 1$ the algorithm is equal to the FS algorithm. Constraining $\Delta\mu_{\max} = 0$ we obtain a parallel structure of the decomposition DAG, similar to the FP algorithm; however, codewords are added sequentially with a constraint on a parallel structure. In general, the constraint on depth lets us tune the structure of the graph with respect to parallelism. In Fig. 2c and 2d, the resulting graph structures for a graph constraint to a fully parallel structure and a depth difference of $\Delta\mu_{\max} = 1$ are depicted, respectively.

D. Related Algorithms

Most competing algorithms for CMVM have a decent time complexity for small matrices. However as they solve complex underlying problems, such as 0-1 integer linear programming [8], they do not scale well with growing matrix size and/or precision. They are hence often intractable. Instead, we use as a benchmark the best-performing MCM algorithm known, presented in [10], that has reasonable polynomial time complexity and is thus tractable for larger matrices as well. Note that MCM, the multiplication of a variable scalar to a arbitrary constant vector, is a special case of CMVM. Any CMVM problem can therefore be rewritten as a sum of K MCM problems, i.e.

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{x} = \sum_{k=1}^{K} \boldsymbol{t}_k x_k, \tag{16}$$

that are solved independently. Here, $\boldsymbol{t}_k$ and x_k are the k-th column vector in $\boldsymbol{T}$ and k-th element in $\boldsymbol{x}$, respectively. Due to the reduced search space the benchmark MCM algorithm has excellent performance. However, the adder tree required for the summation of the K partial results, as well as a DAG structure, similar to the FS algorithm, limit the performance when pipelined.

IV. Numerical Experiments

The entries of all target matrices in the subsequent evaluations are drawn from an i.i.d. Gaussian distribution with zero mean and unit variance. We expect that for practical matrices, e.g. weight matrices of NNs, similar performance is observed for LCC algorithms [17]. A Python implementation of all algorithms discussed in this paper is available in our *github* repository: https://github.com/hansrosenberger/computationcoding.

As the first experiment, we compare the different algorithms for target matrices of dimension 64 × 4 in Fig. 3. The figure shows, the FS algorithm achieves the highest SQNR, considering only the cost of additions (dashed lines). However, when considering the total hardware cost, the FS performance massively deteriorates, leaving this algorithm impractical for a pipelined implementation. The overall hardware cost in this case is dominated by delay elements required to equalize path differences within the DAG. The MA constrained to a FP structure shows the best overall performance, when considering

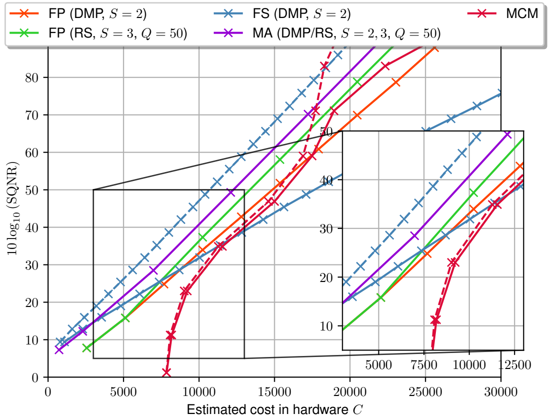


Fig. 3: Comparison of different algorithmic approaches for decomposing a 64×4 target matrix $\boldsymbol{T}$. Solid lines indicate results considering the total cost C_{total}. Dashed lines only consider the cost of adders $C_{\text{add}} N_{\text{add}}$. MCM refers to the algorithm presented in [10] (using the C++ implementation available on [16] and extended by our hardware model). The results for each algorithm are averaged over 10^5 matrix entries.

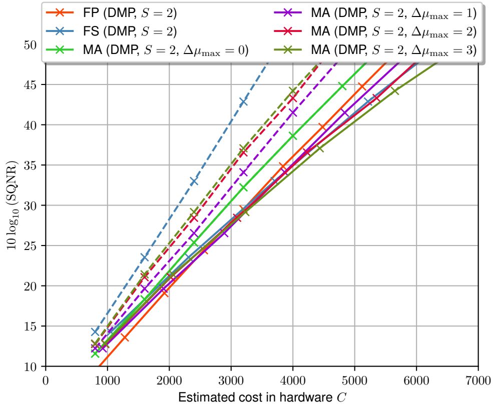


Fig. 4: Comparison of different depth parameters $\Delta\mu_{\max}$ of the MA given a 16×4 target matrix $\boldsymbol{T}$. Solid lines indicate results considering the total cost C_{total}. Dashed lines only consider the cost of adders $C_{\text{add}} N_{\text{add}}$. The results for each algorithm are averaged over 10^5 matrix entries.

the total hardware cost. It outperforms the FP algorithm, both the DMP and RS versions. Relative gains are particulary large for the low SQNR regime. This is achieved by first setting $S = 2$ and utilizing the DMP to build up a coarse codebook from the initial codebook, and then dynamically switching to $S = 3$ via the RS approach. The savings of MA to the FS result from an improved structure of the DAG for the first few layers. The FP algorithm is forced to find an approximation for each target vector separately. This creates codewords that are correlated and unnecessary for the computation. The MA eliminates this redundancy (cf. Figs. 2b and 2c).

As the second experiment we compare the performance of the MA using different depth parameters $\Delta\mu_{\max}$ for target matrices of dimension 16×4 in Fig. 4. Considering only the cost of the adders (dashed lines), we can clearly observe a tradeoff between parallelism and performance, i.e. decreasing $\Delta\mu_{\max}$ leads to a performance degradation. However, when considering the total hardware cost (solid lines) the MA performs best when constrained to a FP structure ($\Delta\mu_{\max} = 0$). For $\Delta\mu_{\max} > 0$ the MA performs worse than its FP counterpart and for some instances even worse than the FS algorithm. This result seems somewhat intuitive: Elements that incur a hardware cost that is not vanishingly small should also improve the SQNR. Hence, a fully parallel structure seems to be the best option.

Remark 2: LCC works best for matrices with an exponential aspect ratio, i.e. $K \approx \log N$. Therefore, we only consider in the evaluation matrices with that property. For approximately square matrices it is beneficial to cut these into rectangular matrices with more extreme aspect ratios and apply an LCC algorithm to each slice individually [18]. For example, to decompose a 64×64 matrix with a target SQNR of $47\,\text{dB}$, a slicing into submatrices of size 64×4 is a good choice.

V. Conclusion

By interpreting the decomposition of a matrix as a DAG, we proposed a new MA for LCC. The proposed algorithm is able to significantly outperform existing schemes. Using a realistic hardware model for pipelining, we show that in almost all cases it is best to decompose a target matrix constraining the resulting DAG to a parallel structure.

References

[1] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," 2021. arXiv:2103.13630.

[2] J. O. Neill, "An overview of neural network compression," 2020. arXiv:2006.03669.

[3] R. R. Müller, B. M. W. Gäde, and A. Bereyhi, "Linear computation coding: A framework for joint quantization and computing," *Algorithms*, vol. 15, no. 7, p. 253, 2022.

[4] R. R. Müller, "Linear computation coding inspired by the Lempel-Ziv algorithm," in *2022 IEEE Information Theory Workshop (ITW)*, IEEE, 2022.

[5] H. Rosenberger, J. S. Fröhlich, A. Bereyhi, and R. R. Müller, "Linear computation coding: Exponential search and reduced-state algorithms," in *2023 Data Compression Conference (DCC)*, IEEE, 2023.

[6] A. Karataev, H. Rosenberger, A. Bereyhi, and R. R. Müller, "Storage constrained linear computation coding," in *2023 Data Compression Conference (DCC)*, IEEE, 2023.

[7] A. D. Booth, "A signed binary multiplication technique," *The Quarterly Journal of Mechanics and Applied Mathematics*, vol. 4, no. 2, pp. 236–240, 1951.

[8] L. Aksoy, P. Flores, and J. Monteiro, "A novel method for the approximation of multiplierless constant matrix vector multiplication," in *2015 IEEE 13th International Conference on Embedded and Ubiquitous Computing*, IEEE, 2015.

[9] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.

[10] Y. Voronenko and M. Püschel, "Multiplierless multiple constant multiplication," *ACM Transactions on Algorithms*, vol. 3, no. 2, p. 11, 2007.

[11] P. Cappello and K. Steiglitz, "Some complexity issues in digital signal processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 5, pp. 1037–1041, 1984.

[12] M. R. Garey and D. S. Johnson, *Computers and Intractability*. W. H. Freeman and Company, 1979.

[13] U. Tietze, C. Schenk, and E. Gamm, *Halbleiter-Schaltungstechnik*. Springer Vieweg Berlin, Heidelberg, 16 ed., 2019.

[14] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*. Elsevier, Morgan Kaufmann, 5 ed., 2012.

[15] A. Lehnert, H. Rosenberger, R. Müller, and M. Reichenbach, "More efficient CMMs on FPGAs: Instantiated ternary adders for computation coding," in *Applied Reconfigurable Computing. Architectures, Tools, and Applications*, pp. 275–289, Springer Nature Switzerland, 2023.

[16] "Spiral: Software/hardware generation for performance: Multiplier block generator." https://spiral.ece.cmu.edu/mcm/gen.html. Accessed: 15.04.2023.

[17] R. R. Müller, H. Rosenberger, and M. Reichenbach, "Linear computation coding for convolutional neural networks," in *Statistical Signal Processing (SSP) Workshop*, (Hanoi, Vietnam), 2023.

[18] A. Lehnert, P. Holzinger, S. Pfenning, R. Müller, and M. Reichenbach, "Most resource efficient matrix vector multiplication on FPGAs," *IEEE Access*, vol. 11, pp. 3881–3898, 2023.

Optimizing IRS-Assisted SIMO/MISO Channels: An Analytical Approach

Milad Dabiri, Sergey Loyka

***Abstract*—Intelligent reflective surfaces (IRS) have recently emerged as a tool to improve the energy and spectral efficiencies of wireless systems and networks at reasonable cost. The underlying IRS optimization problems are difficult due to their non-convex nature, complicated analytical structure and the lack of appropriate analytical tools. While a number of algorithms were proposed to obtain locally-optimal or approximate solutions, globally-optimal ones are out of reach at the moment. This paper considers single-input multiple-output (SIMO) or multiple-input single-output (MISO) IRS-assisted channels and develops an analytical approach for their global optimization exploiting some special features of the problem. A number of closed-form solutions for globally-optimal IRS phase shifts are obtained in some special but practically-important cases, which show that the globally-optimal IRS gain scales either linearly or quadratically with the number of its elements. For a distributed multi-IRS channel, it scales linearly with the number of IRSs. From this, a minimum number of elements can be determined for IRS to have a significant impact. Upper bounds to the globally-optimal IRS gain are established in the general case, which are tight for some channels thus establishing globally-optimal phase shifts in those cases.**

I. Introduction

Intelligent reflecting surfaces (IRS, also known as reconfigurable intelligent surfaces) have recently emerged as a low complexity/cost tool to improve spectral and energy efficiencies of modern wireless networks at reasonable complexity/cost [1]-[4]. Experimental studies and prototypes demonstrate the feasibility of this approach and report notable gains [4][5][9], at least in certain scenarios. A number of studies of IRS-assisted MIMO wireless systems/channels have been reported as well - we refer the reader to the recent surveys [1]-[3] for more details and review some of these studies below.

Since analytical solutions are not feasible in most cases due to the complexity of underlying optimization problems and other factors, a number of numerical algorithms for IRS optimization in various configurations and using various criteria have been proposed [3][10]-[13]. While these algorithms do optimize IRS phase shifts in various ways and are valuable from a practical perspective, they suffer from the same fundamental weakness: their convergence point is locally-optimal at best and can be far away from a globally-optimal one; it also depends on an algorithm's initial point and the gap to a globally-optimal solution is not known. It should be emphasized that these weaknesses are not due to some deficiencies of the above algorithms but rather due to non-convexity of the underlying optimization problems and their complicated analytical structure, which is typical for non-convex problems in general [16][17]. While algorithm for global non-convex optimization do exist, their complexity is exponential in the number of variables and constraints so that only small-size problems can be solved in reasonable time; non-convex problems are known to be hard in general [17]. Since IRS offers significant advantages when the number of elements is large (100s or 1000s) [9][10], using generic algorithms for global optimization is ruled out.

From analytical perspective, no closed-form solutions to the considered optimization problems are known either (due to their difficult analytical structure and the lack of appropriate tools). This makes it difficult to evaluate fundamental performance limits from communication/information-theoretic perspective.

Thus, the major challenge for the IRS optimization problems is to determine reflector phase shifts that are *globally* rather than locally optimal, preferably analytically and in a manageable closed-form (amendable to further communication/information-theoretic analysis), or with algorithms of reasonable (polynomial) complexity. The key difficulty is due to the non-convexity of the related optimization problems so that any standard approach (e.g. using KKT conditions, gradient-decent or similar algorithms) will result in locally-optimal solutions at best [16][17]. New tools are needed to overcome this difficulty by exploiting the problem's structure.

In this paper, we consider SIMO/MISO channels assisted by single or multiple IRS(s) (made of passive reflectors with adjustable phase shifts). The considered channel model is general enough to include line-of-sight (LOS) and non-LOS scenarios, multipath and fading, propagation and reflection losses, in near or far-field scenarios. To address the above-mentioned issues, we adopt an analytical approach (reminiscent of the approaches used in information theory, which are based on upper bounds and achievability schemes) and revisit the IRS optimization problem in terms of its *global* (rather than local) optima in two ways:

(i) we present compact closed-form upper and lower bounds to the *globally*-optimal IRS gain in the general case, which are tight in some special cases, thus providing a *globally*-optimal solution in those cases; the upper bounds can also be used as a benchmark to evaluate the performance of known algorithms in terms of their *global* optimality gap (which remains unknown at this time),

(ii) we obtain closed-form *globally*-optimal solutions for some special cases typical for massive MIMO and distributed multi-IRS channels; these solutions indicate that the IRS-assisted globally-optimal gain scales either linearly or quadratically with the number of reflectors and linearly with the number of IRSs, which is consistent with measurements in [5][7]; since the presented solutions are in compact closed-forms, they can be used for online optimization of large IRSs.

Notations: bold capitals and bold lower-case letters denote matrices and vectors, respectively, while regular lower case

M. Dabiri and S. Loyka are with the School of Electrical Engineering and Computer Science, University of Ottawa, Canada, e-mail: sergey.loyka@uottawa.ca

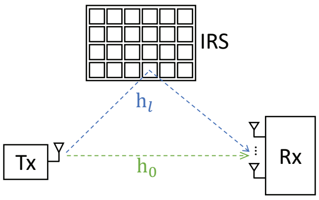


Fig. 1. An illustration of IRS-assisted SIMO channel; $\mathbf{h}_0$ represents the direct Tx-Rx link while $\mathbf{h}_l$ - the l-th reflected link.

letters denote scalars; $|\mathbf{h}|$, $\mathbf{h}^T$ and $\mathbf{h}^+$ denote Euclidean norm (length), transposition and Hermitian conjugation of column vector $\mathbf{h}$ (for scalar h, h^+ denotes complex conjugation), h_i is i-th entry of vector $\mathbf{h}$; $\arg\{z\}$ is the argument (phase) of a complex number z.

II. Channel Model and Problem Formulation

Let us consider an IRS-assisted SIMO (uplink) channel as shown in Fig. 1, consisting of a single-antenna transmitter (Tx, e.g. user equipment), an IRS equipped with L passive reflectors, and a receiver (Rx) equipped with N antennas (e.g. a base station). Following the standard discrete-time baseband model [3][10]-[13], the received signal is

$$\mathbf{y} = \big(\mathbf{h}_0 + \sum\nolimits_{l=1}^{L} e^{j\phi_l}\mathbf{h}_l\big)x + \mathbf{z} \tag{1}$$

where $\mathbf{y} = [y_1, ..., y_N]^T \in \mathbb{C}^{N\times 1}$ is the Rx signal vector and its n-th entry y_n is the Rx signal of n-th antenna, $\mathbf{z} \in \mathbb{C}^{N\times 1}$ is the circularly-symmetric, complex Gaussian noise vector, i.i.d. across antennas, of zero mean and variance σ_0^2 per Rx antenna; x is the scalar transmitted (Tx) signal satisfying the Tx power constraint $\mathbb{E}\{|x|^2\} = P$; $\mathbf{h}_0, \mathbf{h}_l \in \mathbb{C}^{N\times 1}$ are the channel vectors representing the direct Tx-Rx and the reflected Tx-IRS-Rx links (via l-th reflector), n-th entry of $\mathbf{h}_l$ is $h_{l,n} = h_l^{Tx} h_{l,n}^{Rx}$, where h_l^{Tx} and $h_{l,n}^{Rx}$ represent the Tx-IRS and IRS-Rx links via l-th reflector (all including the average propagation loss and the reflection loss), and ϕ_l is the reflector-induced phase shift (to be optimized later on).

The channel is further assumed to be static or quasi-static (stays fixed for a sufficiently long time), frequency-flat, with full channel information (CSI) available to the Rx and IRS controller[1]. Note that this model accommodates single-IRS as well as multi-IRS settings, where L is the total number of reflectors and $\mathbf{h}_l$, $l = 1..L$, represent the respective IRS-assisted links, and it applies to LOS, non-LOS (or partially-blocked LOS) and multi-path scenarios (since we do not make any specific assumptions on $\mathbf{h}_0$, $\mathbf{h}_l$ at this point).

For the no-IRS case, the Rx SNR or power is maximized via matched filtering (MF)/beamforming (also known as maximal-ratio combining, MRC) and can be expressed as [15]

$$\gamma_{no-IRS} = |\mathbf{h}_0|^2\gamma_0, \ \gamma_0 = P/\sigma_0^2 \tag{2}$$

Absorbing the average propagation path loss into P, γ_0 becomes the Rx SNR in the unit-gain channel. Likewise, for the IRS-assisted channel in (1), $\mathbf{h}_{eq} \triangleq \mathbf{h}_0 + \sum_{l=1}^{L} e^{j\phi_l}\mathbf{h}_l$ is the equivalent channel vector and the MF beamforming does maximize its Rx SNR/power for given $\boldsymbol{\phi}$,

$$\gamma(\boldsymbol{\phi}) = g(\boldsymbol{\phi})\gamma_0, \ g(\boldsymbol{\phi}) = |\mathbf{h}_{eq}|^2 = \Big|\mathbf{h}_0 + \sum_{l=1}^{L} e^{j\phi_l}\mathbf{h}_l\Big|^2 \tag{3}$$

where $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_L]^T$ is the vector of IRS phase shifts, and we emphasize that the SNR $\gamma(\boldsymbol{\phi})$ depends on $\boldsymbol{\phi}$; $g(\boldsymbol{\phi})$ is the IRS-assisted gain for given $\boldsymbol{\phi}$; with some abuse of terminology, we call it simply "IRS gain" in the rest of this paper.

For a given $\boldsymbol{\phi}$, the maximum achievable rate per unit bandwidth (spectral efficiency) supported by this IRS-assisted SIMO channel is $R(\boldsymbol{\phi}) = \log(1 + g(\boldsymbol{\phi})\gamma_0)$ and the IRS-assisted channel capacity is

$$C_{IRS} = \max_{\boldsymbol{\phi}} R(\boldsymbol{\phi}) = \log(1 + \gamma_0 g^*), \tag{4}$$

$$g^* = \max_{\boldsymbol{\phi}} \big|\mathbf{h}_0 + \sum\nolimits_{l=1}^{L} e^{j\phi_l}\mathbf{h}_l\big|^2 \tag{5}$$

where $g^* = g(\boldsymbol{\phi}^*) = \max_{\boldsymbol{\phi}} g(\boldsymbol{\phi})$ is the globally-optimal IRS gain and $\boldsymbol{\phi}^*$ are the respective globally-optimal phase shifts. This SIMO channel setting can also be extended to a MISO channel via the uplink-downlink duality and channel reciprocity [15].

III. Prior Results

Despite its apparent simplicity, no closed-form solution of the problem in (5) is known to date in the general SIMO case, i.e. for arbitrary $\mathbf{h}_0$, $\mathbf{h}_l$, L and N. The key difficulties are non-convexity of the problem and the lack of appropriate analytical tools. While a number of numerical algorithms have been proposed [3][10]-[12], they lack insights, exhibit local convergence at best (since the problem is not convex) and the gap to global optima is not known. The relevant analytical results for global optima are rare. We briefly review them below.

The first special case for which the globally-optimal solution is known (and easy to establish) is that of the SISO channel, i.e. $N = 1$ so that h_0, h_l are scalars. The globally-optimal phase shifts and IRS gain are [1][3]

$$\phi_l^* = \arg\{h_0\} - \arg\{h_l\}, \ g^* = \big(|h_0| + \sum\nolimits_{l=1}^{L} |h_l|\big)^2 \tag{6}$$

so that the signals of the direct and all reflected links add up constructively. Unfortunately, this result does not extend to a general SIMO channel (when $\mathbf{h}_0$, $\mathbf{h}_l$ are arbitrary vectors). However, closed-form solutions can be obtained in some special cases.

Specifically, for SIMO channels typical for millimeter waves (mmWave) or THz propagation conditions, where the dominant propagation mode is via LOS, the globally-optimal closed-form solutions can be obtained as follows [14].

Proposition 1. *Let the IRS-assisted channel satisfy $\mathbf{h}_l = e^{j\alpha_l}\mathbf{h}_1$, $l = 1 \cdots L$, for some real α_l and $\alpha_1 = 0$, and let $\mathbf{h}_0$ be arbitrary. Then, the globally-optimal IRS gain and the respective phase shifts are as follows:*

$$g^* = |\mathbf{h}_0|^2 + 2L|\mathbf{h}_0^+\mathbf{h}_1| + L^2|\mathbf{h}_1|^2 \tag{7}$$

$$\phi_l^* = \arg\{\mathbf{h}_1^+\mathbf{h}_0\} - \alpha_l = \arg\{\mathbf{h}_l^+\mathbf{h}_0\} \tag{8}$$

[1] the case of partial CSI can be handled via the compound channel approach, similarly to the standard MIMO channel [20]. This, however, is beyond the scope of the present paper.

where 2nd equality in (8) *holds if* $\mathbf{h}_1^+\mathbf{h}_0 \neq 0$. *If* $\mathbf{h}_1^+\mathbf{h}_0 = 0$, *then* $\phi_l^* = \alpha_0 - \alpha_l$, *where* α_0 *is arbitrary, i.e. the optimal phase shifts are not unique.*

It follows from (7) that the globally-optimal IRS gain g^* scales either linearly or quadratically with the number L of reflectors, depending on how strong the direct link $\mathbf{h}_0$ is, e.g. $g^* = L^2|\mathbf{h}_1|^2$ if $\mathbf{h}_0 = 0$ (blocked direct link). Note that the conditions of this Proposition are always satisfied if $L = 1$, so it gives the globally-optimal solution in this case.

Proposition 1 can be further extended to a slightly more general case of alighted reflector links with $\mathbf{h}_l = a_l\mathbf{h}_1$ for some complex a_l but, unfortunately, not beyond that. The general case of arbitrary $\mathbf{h}_0, \mathbf{h}_l$, N and L can be addressed via lower and upper bounds. To this end, let

$$\mathbf{H} = [\mathbf{h}_0, \cdots, \mathbf{h}_L],\ \mathbf{w} = [1, e^{j\phi_1}, \cdots, e^{j\phi_L}]^T \tag{9}$$

and $\mathbf{H} = \mathbf{U\Sigma V}^+$ be the singular value decomposition (SVD), where $\mathbf{U}, \mathbf{V}$ are the unitary matrices of left and right singular vectors of $\mathbf{H}$, respectively, and $\mathbf{\Sigma} = diag\{\sigma_l(\mathbf{H})\}$ is the diagonal matrix of its singular values $\sigma_l(\mathbf{H})$ sorted in decreasing order, i.e. $\sigma_1(\mathbf{H})$ is the largest one; $\mathbf{v}_l$ is the l-th column of $\mathbf{V}$ (i.e. the right singular vector corresponding to the l-th largest singular value $\sigma_l(\mathbf{H})$). The following Proposition presents the desired lower and upper bounds for the general SIMO (and, by duality, MISO) case using the SVD of $\mathbf{H}$ [14].

Proposition 2. *The globally-optimal IRS gain* $g^* = g(\boldsymbol{\phi}^*)$ *for the channel in* (1) *is bounded as follows:*

$$g_{lb} \leq g^* \leq g_{ub} = \sigma_1^2(\mathbf{H})(L+1) \tag{10}$$

$$g_{lb} = \sigma_1^2(\mathbf{H})|\mathbf{v}_1|_1^2 + \sum_{l=2}^{r(\mathbf{H})} \sigma_l^2(\mathbf{H})|\mathbf{w}_1^+\mathbf{v}_l|^2 \tag{11}$$

$$w_{1l} = \exp\{j\arg(v_{1l}) - j\arg(v_{11})\},\ l = 1...L+1 \tag{12}$$

where $r(\mathbf{H})$ *is the rank of* $\mathbf{H}$, $|\mathbf{v}_1|_1 = \sum_{l=1}^{L+1}|v_{1l}|$ *is* l_1 *norm;* v_{1l} *and* w_{1l} *are* l*-th entry of* $\mathbf{v}_1$ *and* $\mathbf{w}_1$*, respectively.*

The lower bound is tight, i.e. $g_{lb} = g^*$, *if* $\mathbf{H}$ *is rank-one,* $r(\mathbf{H}) = 1$,

$$g^* = \sigma_1^2(\mathbf{H})|\mathbf{v}_1|_1^2,\ \phi_l^* = \arg\{v_{1(l+1)}\} - \arg\{v_{11}\} \tag{13}$$

where ϕ_l^* *are globally-optimal phase shifts for this case.*

If $\mathbf{v}_1$ *has equal-magnitude entries, i.e.* $|v_{1l}| = |v_{11}|$ *for* $l = 1..L+1$, *then the upper and lower bounds coincide and are therefore tight,*

$$g^* = g_{lb} = g_{ub} = \sigma_1^2(\mathbf{H})(L+1) \tag{14}$$

and the globally-optimal phase shifts are as in (13).

The lower bound is close to the globally-optimal IRS gain, $g^* \approx g_{lb}$, if $\sigma_1(\mathbf{H}) \gg \sigma_2(\mathbf{H})$ and this becomes exact equality if $r(\mathbf{H}) = 1$. Note also from (10) and (11) that the globally-optimal IRS gain g^* (or Rx SNR/power) scales at least as $\sigma_1^2(\mathbf{H})$ in the general case,

$$\sigma_1^2(\mathbf{H}) \leq \sigma_1^2(\mathbf{H})|\mathbf{v}_1|_1^2 \leq g^* \leq \sigma_1^2(\mathbf{H})(L+1) \tag{15}$$

where 1st inequality is due to $|\mathbf{v}_1|_1 \geq 1$. This is somewhat similar to the regular MIMO channel with channel matrix $\mathbf{H}$, where the maximum SNR gain achievable with Tx/Rx beamforming is $\sigma_1^2(\mathbf{H})$ so that the globally-optimal IRS gain is at least as large (assuming both channels have the same channel matrix $\mathbf{H}$).

To the best of our knowledge, no other analytical results for the problem in (5) are known. The next Section presents new analytical results for this problem, for which the proofs are outlined in the Appendix.

IV. New bounds and closed-form solutions

The next Proposition presents novel upper bounds to g^* with explicit dependence on channel vectors $\mathbf{h}_l$ (since the SVD, while being a useful tool from computational and information-theoretic perspectives, essentially ”hides” such dependence).

Proposition 3. *The globally-optimal IRS gain* g^* *is upper bounded in the general case as follows:*

$$g^* \leq g_{UB} \triangleq |\mathbf{h}_0|^2 + 2\sum_{l=1}^{L}|\mathbf{h}_0^+\mathbf{h}_l| + \sum_{l,k=1}^{L}|\mathbf{h}_l^+\mathbf{h}_k| \tag{16}$$

$$\leq \left(|\mathbf{h}_0| + \sum\nolimits_{l=1}^{L}|\mathbf{h}_l|\right)^2 \tag{17}$$

where both inequalities hold with equality (i.e. the upper bounds are attained) if $\mathbf{h}_l = a_l\mathbf{h}_0$ *for all* l *and some complex* a_l*, and the respective globally-optimal phase shifts are* $\phi_l^* = -\arg\{a_l\}$.

Comparing the upper bounds in (16) and (10), it can be shown (by examples) that neither is tighter in general, so that they are complementary to each other.

Next, we present closed-form solutions for the problem in (5) and identify additional cases where the upper bound in (16) is tight, i.e. where $g^* = g_{UB}$.

A. Massive MIMO

When the number N of Rx antennas is large, as in massive MIMO, and the condition known as ”favorable propagation” holds, individual propagation paths become resolvable and the respective channel vectors become orthogonal to each other, $\mathbf{h}_l^+\mathbf{h}_k = 0,\ l \neq k$ [19]. The next Proposition presents a closed-form globally-optimal solution of (5) in this case.

Proposition 4. *Let the channel vectors of reflected paths be mutually-orthogonal,* $\mathbf{h}_l^+\mathbf{h}_k = 0,\ l \neq k$, *where* $l, k = 1...L$ *(no such assumption is made for the LOS path). Then, the globally optimal phase shifts are* $\phi_l^* = \arg\{\mathbf{h}_l^+\mathbf{h}_0\}$ *and the respective IRS gain is*

$$g^* = |\mathbf{h}_0|^2 + 2\sum\nolimits_{l=1}^{L}|\mathbf{h}_l^+\mathbf{h}_0| + \sum\nolimits_{l=1}^{L}|\mathbf{h}_l|^2 \tag{18}$$

If $\mathbf{h}_l^+\mathbf{h}_0 = 0$, *then* ϕ_l^* *is arbitrary.*

Note that, in this case, the upper bound in (16) is also tight, $g^* = g_{UB}$. Comparing the above result to (8), we note that $\phi_l^* = \arg\{\mathbf{h}_l^+\mathbf{h}_0\}$ is globally-optimal in the cases where $\mathbf{h}_l$ are either orthogonal or parallel to each other. This feature can be used to obtain a globally-optimal solution for a distributed multi-IRS channel as follows.

B. Multi-IRS channel

A multi-IRS setup is often considered in the literature as an inexpensive way to enhance system performance. In this section, we consider a scenario where several IRSs are

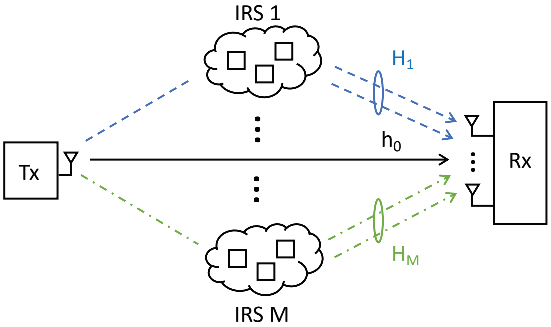


Fig. 2. An illustration of multi-IRS SIMO channel with M IRSs.

spatially distributed, as in Fig. 2. When the number N of Rx antennas is large and, therefore, the Rx antenna array angular resolution is high [18], reflectors from different IRSs can be resolved at the Rx (since their spacing is large) so that their channel vectors are orthogonal to each other. On the other hand, reflectors from the same IRS are not resolvable since their spacing is not large enough so that their Rx angles of arrival are almost the same.

Based on this consideration, the overall multi-IRS channel matrix $\mathbf{H}$ can be block-partitioned as $\mathbf{H} = [\mathbf{h}_0\ \mathbf{H}_1\ ...\ \mathbf{H}_M]$, where each block represents the respective IRS (collects channel vectors from that IRS):

$$\mathbf{H}_1 = [\mathbf{h}_1,\ ...,\ \mathbf{h}_{L_1}] \tag{19}$$

$$\mathbf{H}_m = [\mathbf{h}_{L_1+...+L_{m-1}+1},\ ...,\ \mathbf{h}_{L_1+...+L_m}] \tag{20}$$

where $\mathbf{H}_m$ represents m-th IRS, L_m is its number of reflectors, $m = 1...M$, and M is the number of IRSs. The total number of reflectors (in all IRSs) is $L = \sum_{m=1}^{M} L_m$. For further use, let us define the index set $\mathcal{I}_m$ of columns in $\mathbf{H}_m$:

$$\mathcal{I}_m = \{l : L_1 + ... + L_{m-1} + 1 \leq l \leq L_1 + ... + L_m\} \tag{21}$$

Since reflectors from different (distant) IRSs are resolvable, their channel vectors are orthogonal to each other, so that $\mathbf{h}_l^+\mathbf{h}_k = 0$ for $l \in \mathcal{I}_{m_1}$, $k \in \mathcal{I}_{m_2}$, $m_1 \neq m_2$, and therefore $\mathbf{H}_{m_1}^+\mathbf{H}_{m_2} = \mathbf{0}$. On the other hand, since reflectors of the same IRS are not resolvable due to their proximity to each other, $r(\mathbf{H}_m) = 1$ and therefore

$$\mathbf{h}_l = a_l\mathbf{u}_m \text{ for some } a_l\ \forall l \in \mathcal{I}_m,\ m = 1...M \tag{22}$$

where $\mathbf{u}_m$ is the unit basis vector of span$\{\mathbf{H}_m\}$; without loss of generality, we further assume that $a_l \neq 0$ for all l. Since different blocks are orthogonal to each other, it follows that $\{\mathbf{u}_1...\mathbf{u}_M\}$ is an orthonormal set.

For this multi-IRS channel, the following Proposition provides an explicit globally-optimal solution of (5).

Proposition 5. *The globally-optimal phase shifts and the respective IRS gain of the multi-IRS channel described above are as follows:*

$$\phi_l^* = \arg\{\mathbf{u}_m^+\mathbf{h}_0\} - \arg\{a_l\} = \arg\{\mathbf{h}_l^+\mathbf{h}_0\}\ \forall l \in \mathcal{I}_m \tag{23}$$

$$g^* = g_{UB} = |\mathbf{h}_0|^2 + 2\sum_{m=1}^{M} A_m|\mathbf{h}_0^+\mathbf{u}_m| + \sum_{m=1}^{M} A_m^2 \tag{24}$$

where 2nd equality in (23) *holds if* $\mathbf{u}_m^+\mathbf{h}_0 \neq 0$*;* A_m *is the combined amplitude gain of* m*-th IRS,*

$$A_m = \sum_{l\in\mathcal{I}_m} |a_l| = \sum_{l\in\mathcal{I}_m} |\mathbf{h}_l| \tag{25}$$

If $\mathbf{u}_m^+\mathbf{h}_0 = 0$*, then* $\phi_l^* = \psi_m - \arg\{a_l\}$ *for all* $l \in \mathcal{I}_m$ *and arbitrary (real)* ψ_m*.*

It should be noted that the globally-optimal phase shifts in (23) are of the same form as in (8) and (18), and, for this multi-IRS channel, the upper bound in (16) is also tight, $g^* = g_{UB}$. (25) represents an equal-gain combiner (EGC) and A_m is its amplitude gain.

To get some insight, let us consider the case of absent (blocked) direct link, $\mathbf{h}_0 = 0$, for which (24) reduces to

$$g^* = \sum_{m=1}^{M} A_m^2 = \sum_{m=1}^{M} \Big(\sum_{l\in\mathcal{I}_m} |\mathbf{h}_l|\Big)^2 \tag{26}$$

Note that the internal summation represents amplitude-wise combining (or EGC) for each IRS across its reflectors while the external sum is a power-wise combining across different IRSs. This difference is due to the fact that the reflectors within the same IRS are not resolvable (so that their channel vectors are parallel to each other) while the reflectors of different IRSs are resolvable (so that their channel vectors are orthogonal to each other).

To reveal the scaling of the globally-optimal IRS gain g^* with the number M of IRSs and the numbers L_m of reflectors in each IRSs, let us consider the case when all IRSs are identical and have similar channels to their reflectors so that $L_m = L_1$, $|\mathbf{u}_m^+\mathbf{h}_0| = |\mathbf{u}_1^+\mathbf{h}_0|$ for all m, and $|\mathbf{h}_l| = |\mathbf{h}_1|$ for all l. In this case, (24) reduces to

$$g^* = |\mathbf{h}_0|^2 + 2|\mathbf{h}_1^+\mathbf{h}_0|ML_1 + |\mathbf{h}_1|^2ML_1^2 \tag{27}$$

In the case of weak or absent direct link, the last term dominates so that

$$g^* \approx |\mathbf{h}_1|^2ML_1^2 \ \text{ if } \ \frac{|\mathbf{h}_0|^2}{ML_1} + 2|\mathbf{h}_1^+\mathbf{h}_0| \ll |\mathbf{h}_1|^2L_1 \tag{28}$$

i.e. g^* scales quadratically with L_1 but only linearly with M and this scaling holds provided the number L_1 of reflectors per IRS is large enough. If the opposite is true,

$$g^* \approx |\mathbf{h}_0|^2 + 2|\mathbf{h}_1^+\mathbf{h}_0|ML_1 \ \text{ if } \ \frac{|\mathbf{h}_0|^2}{ML_1} + 2|\mathbf{h}_1^+\mathbf{h}_0| \gg |\mathbf{h}_1|^2L_1$$

i.e. the scaling with M, L_1 is linear at best, and this holds provided the direct link is not too strong,

$$g^* \approx 2|\mathbf{h}_1^+\mathbf{h}_0|ML_1 \ \text{ if } \ |\mathbf{h}_0|^2 \ll 2|\mathbf{h}_1^+\mathbf{h}_0|ML_1 \tag{29}$$

If the opposite is true, then $g^* \approx |\mathbf{h}_0|^2$, i.e. IRSs are useless. Thus, of all three cases considered, the first one is most favorable in terms of IRSs impact, i.e., the stronger the direct link, the smaller the impact of the IRSs. For IRSs to be effective, either (28) or (29) has to hold, which can be used as design guidelines as to (i) how many reflectors per IRS or (ii) how many IRSs are needed to make a significant impact.

Finally, one can consider the case of resolvable direct and reflected paths so that their channel vectors are orthogonal to each other, $\mathbf{h}_1^+\mathbf{h}_0 = 0$. In this case, (27) reduces to

$$g^* = |\mathbf{h}_0|^2 + |\mathbf{h}_1|^2ML_1^2 \tag{30}$$

and IRSs have significant impact if 2nd term is dominant, i.e. $g^* \approx |\mathbf{h}_1|^2 ML_1^2$ if $ML_1^2 \gg |\mathbf{h}_0|^2 |\mathbf{h}_1|^{-2}$. The latter condition can be combined with physically-based models in [8]-[10] to evaluate $\mathbf{h}_0, \mathbf{h}_1$ and thus to determine the required M, L_1 for IRSs to have substantial impact.

We further note that the above scalings in (27)-(30) are consistent with the measurements in [5, Fig. 12(b)][7, Fig. 18], where quadratic scaling of the IRS gain with the number of reflectors was experimentally verified in a certain environment for large (single) IRS.

C. Variable-gain reflectors

One can further consider a more general setting where IRS reflectors, while being passive, have variable gains β_l, where $\beta_l \leq 1$ reflects their passive nature. In this setting, the IRS gain is

$$g(\boldsymbol{\phi}, \boldsymbol{\beta}) = |\mathbf{h}_0 + \sum\nolimits_l \beta_l e^{j\phi_l} \mathbf{h}_l|^2 \tag{31}$$

and it can be jointly optimized over $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$:

$$g^* = \max_{\boldsymbol{\phi}, \boldsymbol{\beta}} \left|\mathbf{h}_0 + \sum\nolimits_{l=1}^{L} \beta_l e^{j\phi_l} \mathbf{h}_l\right|^2 \text{ s.t. } \beta_l \leq 1,\ l = 1...L$$

It follows that using the largest possible gains $\beta_l = 1$ is optimal in the cases of Propositions 1, 3-5 provided phase shifts are also optimized and, therefore, there is no loss of optimality in assuming $\beta_l = 1$. However, it can be shown (by examples) that $\beta_l = 1$ are not necessarily optimal if phase shifts are not optimized as well, i.e. the largest possible gains are optimal for the joint optimization only.

V. Appendix: Outlines of Proofs

Proof of Proposition 3: Using (3), one obtains, after some manipulations:

$$\begin{aligned} g(\boldsymbol{\phi}) &= |\mathbf{h}_0|^2 + 2\sum\nolimits_{l=1}^{L} |\mathbf{h}_0^+ \mathbf{h}_l| \cos(\phi_l + \varphi_{0l}) \qquad (32)\\ &\quad + \sum\nolimits_{l,k=1}^{L} |\mathbf{h}_l^+ \mathbf{h}_k| \cos(\phi_k - \phi_l + \varphi_{lk})\\ &\leq |\mathbf{h}_0|^2 + 2\sum\nolimits_l |\mathbf{h}_0^+ \mathbf{h}_l| + \sum\nolimits_{k,l} |\mathbf{h}_l^+ \mathbf{h}_k| \qquad (33)\\ &\leq |\mathbf{h}_0|^2 + 2\sum\nolimits_l |\mathbf{h}_0||\mathbf{h}_l| + \sum\nolimits_{k,l} |\mathbf{h}_l||\mathbf{h}_k| \qquad (34)\\ &= \left(|\mathbf{h}_0| + \sum\nolimits_l |\mathbf{h}_l|\right)^2 \qquad (35) \end{aligned}$$

where $\varphi_{lk} = \arg\{\mathbf{h}_l^+ \mathbf{h}_k\}$; (33) is due to $\cos(x) \leq 1$ and (34) is due to Cauchy-Schwarz inequality $|\mathbf{h}_l^+ \mathbf{h}_k| \leq |\mathbf{h}_l||\mathbf{h}_k|$, which holds with equality if $\mathbf{h}_l = a_l \mathbf{h}_0$ for all l. Since the upper bounds in (33), (34) are independent of $\boldsymbol{\phi}$, taking $\max_{\boldsymbol{\phi}}$ on all sides, (16) and (17) follow. If $\mathbf{h}_l = a_l \mathbf{h}_0$ for all l, (34) holds with equality and

$$\varphi_{lk} = \arg\{\mathbf{h}_l^+ \mathbf{h}_k\} = \arg\{a_k\} - \arg\{a_l\} \tag{36}$$

so that setting $\phi_l = -\arg\{a_l\} = -\varphi_{0l}$,

$$\phi_l + \varphi_{0l} = 0,\ \phi_k - \phi_l + \varphi_{lk} = 0 \tag{37}$$

and therefore (i) (33) holds with equality as well, and (ii) $\phi_l = -\arg\{a_l\}$ are the globally-optimal phase shifts in this case, since they attain both upper bounds.

Proof of Proposition 4: In this case, using (32),

$$\begin{aligned} g(\boldsymbol{\phi}) &= |\mathbf{h}_0|^2 + 2\sum_l |\mathbf{h}_0^+ \mathbf{h}_l| \cos(\phi_l + \varphi_{0l}) + \sum_l |\mathbf{h}_l|^2 \qquad (38)\\ &\leq |\mathbf{h}_0|^2 + 2\sum\nolimits_l |\mathbf{h}_0^+ \mathbf{h}_l| + \sum\nolimits_l |\mathbf{h}_l|^2 \qquad (39) \end{aligned}$$

where (38) is due to $\mathbf{h}_l^+ \mathbf{h}_k = 0$ for $l \neq k$. Note that the upper bound holds for any $\boldsymbol{\phi}$, including the optimal one, is independent of $\boldsymbol{\phi}$ and is attained by $\phi_l = -\varphi_{0l} = \arg\{\mathbf{h}_l^+ \mathbf{h}_0\}$, which are therefore globally-optimal. If $\mathbf{h}_l^+ \mathbf{h}_0 = 0$, then ϕ_l^* is arbitrary since $\arg\{0\}$ is arbitrary and the upper bound in (39) is achieved for any ϕ_l in this case.

Proof of Proposition 5: Observe that the upper bound g_{UB} in (16) also applies to the multi-IRS case here, where L is the total number of reflectors. It can be verified, after some lengthy but otherwise straightforward manipulations, that the upper bound is attained by the phase shifts in (23) under the conditions of this Proposition, which are therefore globally-optimal. The same manipulations verify (24) and (25) as well.

References

[1] E. Basar et al, Wireless Communications Through Reconfigurable Intelligent Surfaces, IEEE Access, v. 7, pp. 116753–116773, Aug. 2019.

[2] M. Di Renzo et al, Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How it Works, State of Research, and The Road Ahead, IEEE JSAC, v. 38, no. 11, pp. 2450–2525, Nov. 2020.

[3] Q. Wu et al, Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial, IEEE Trans. Comm., vol. 69, no. 5, pp. 3313–3351, May 2021.

[4] M. Di Renzo, A.I. Aravanis, Catching the 6G Wave by Using Metamaterials: A Reconfigurable Intelligent Surface Paradigm, in ”Shaping Future 6G Networks: Needs, Impacts, and Technologies”, IEEE Press and Wiley, 2022, pp.69–87.

[5] V. Arun, H. Balakrishnan, RFocus: Beamforming Using Thousands of Passive Antennas, 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI’20), Feb. 25-27, 2020, Santa Clara, CA, pp. 1047–1061.

[6] W. Tang et al., MIMO Transmission Through Reconfigurable Intelligent Surface: System Design, Analysis, and Implementation, IEEE JSAC, vol. 38, no. 11, Nov. 2020, pp. 2683–2699.

[7] M. Rossanese et al, Designing, Building, and Characterizing RF Switch-Based Reconfigurable Intelligent Surfaces, 16th ACM Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization, Sydney, Australia, Oct. 17, 2022, pp. 69–76.

[8] E. Bjornson, L. Sanguinetti, Power Scaling Laws and Near-Field Behaviors of Massive MIMO and Intelligent Reflecting Surfaces, IEEE Open Journal of the Communications Society, vol. 1, pp. 1306–1324, Sep. 2020.

[9] W. Tang et al., Wireless Communications With Reconfigurable Intelligent Surface: Path Loss Modeling and Experimental Measurement, IEEE Trans. on Wireless Comm., vol. 20, no. 1, pp. 421-439, Jan. 2021.

[10] M. Najafi et al, Physics-Based Modeling and Scalable Optimization of Large Intelligent Reflecting Surfaces, IEEE Trans. Comm., vol. 69, no. 4, pp. 2673–2691, Apr. 2021.

[11] Q. Wu, R. Zhang, Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming, IEEE Trans. on Wireless Comm., vol. 18, no. 11, pp. 5394-5409, Nov. 2019.

[12] X. Yu, R. Schober, MISO Wireless Communication Systems via Intelligent Reflecting Surfaces (Invited Paper), IEEE Int. Conf. on Comm. in China, pp. 735-740, Aug. 2019.

[13] T. Jiang, W. Yu, Interference Nulling Using Reconfigurable Intelligent Surface, IEEE JSAC, vol. 40, no. 5, pp. 1392–1406, May 2022.

[14] M. Dabiri, S. Loyka, On Globally-Optimal IRS Design for SIMO/MISO Channels, IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC-23), Toronto, Canada, Sep. 2023.

[15] D. Tse and P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.

[16] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[17] S.A. Vavasis, Complexity Issues in Global Optimization: A Survey. In R. Horst, P.M. Pardalos (Eds.), “Handbook of Global Optimization”, Springer, Boston, MA, 1995 (pp. 27-41).

[18] H. L. Van Trees, Optimum Array Processing, Wiley, 2002.

[19] T. L. Marzetta et al., Fundamentals of Massive MIMO, Cambridge University Press, 2016.

[20] S. Loyka, C.D. Charalambous, Novel Matrix Singular Value Inequalities and Their Applications to Uncertain MIMO Channels, IEEE Trans. Info. Theory, v. 61, N. 12, pp. 6623–6634, Dec. 2015.

Active Eavesdropper Mitigation via Orthogonal Channel Estimation

Gian Marti and Christoph Studer
Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland
email: gimarti@ethz.ch and studer@ethz.ch

***Abstract*—Beamforming is a powerful tool for physical layer security, as it can be used for steering signals towards legitimate receivers and away from eavesdroppers. An active eavesdropper, however, can interfere with the pilot phase that the transmitter needs to acquire the channel knowledge necessary for beamforming. By doing so, the eavesdropper can make the transmitter form beams towards the eavesdropper rather than towards the legitimate receiver. To mitigate active eavesdroppers, we propose VILLAIN, a novel channel estimator that uses secret pilots. When an eavesdropper interferes with the pilot phase, VILLAIN produces a channel estimate that is orthogonal to the eavesdropper's channel (in the noiseless case). We prove that beamforming based on this channel estimate delivers the highest possible signal power to the legitimate receiver without delivering any signal power to the eavesdropper. Simulations show that VILLAIN mitigates active eavesdroppers also in the noisy case.**

I. Introduction

Security is a concern of paramount importance in modern communication systems [1]. Physical layer security (PLS) is emerging as a powerful alternative to classical cryptography [2], [3]. While classical cryptography is based on the assumption that certain computational problems are hard, PLS builds on the characteristics of the channel itself and offers information-theoretic security. In multi-antenna transmission systems, PLS can leverage beamforming to steer communication signals towards the intended recipient and away from an eavesdropper [4].

Beamforming requires the transmitter to know the channels to the receivers—to the legitimate receiver for steering signals towards it and to the eavesdropper for steering signals away from it. Most research on eavesdropper mitigation simply assumes perfect knowledge of all channels [4]–[7]. However, it is unclear how the transmitter could obtain channel knowledge of an eavesdropper, which—if not declared otherwise—is usually understood to be a *passive* eavesdropper, i.e., one which emits no signals. Thankfully, transmitters with a large number of antennas (as are used, e.g., in the massive multiple-input multiple-output (MIMO) downlink) are intrinsically resistant to passive eavesdropping due to the narrow beams that such transmitters form towards the receiver. Such narrow beams entail significantly higher signal strength at the legitimate receiver than at the eavesdropper [8]–[10].

In contrast to passive eavesdroppers, *active* eavesdroppers try to influence communication in their favor by emitting signals themselves. For instance, an active eavesdropper may contaminate the channel estimation (or *pilot*) phase such that the transmitter forms beams towards the eavesdropper instead of the receiver [11]–[13]. This pilot contamination renders active eavesdroppers effective also against transmitters with many antennas. On the flip side, the signals that an active eavesdropper emits give the transmitter an opportunity to *somehow* estimate the eavesdropper's channel and then using beamforming to steer signals away. Much research on active eavesdropper mitigation therefore simply assumes that the channel of an active eavesdropper is perfectly [14]–[18] or imperfectly [19], [20] known at the transmitter, or that at least its second-order statistics are known [12], [13]. *How* this channel knowledge should be obtained is usually not discussed, however. In particular, a sophisticated active eavesdropper might only transmit during the pilot phase and thus never provide the transmitter with snapshots of its channel that are uncontaminated by the pilot signals of the legitimate receiver.

A. Contributions

We propose VILLAIN (short for eaVesdropper resILient channeL estimatIoN) for active eavesdropper mitigation in the single-user MIMO downlink. When an active eavesdropper contaminates the pilot phase, VILLAIN produces an estimate of the legitimate receiver's channel that is orthogonal to the eavesdropper's channel (if the noise at the basestation (BS) is negligible). We prove that using this channel estimate for maximum-ratio transmission (MRT) results in a beamformer that is optimal in the sense that it delivers the highest possible signal power to the legitimate receiver while simultaneously ensuring that the received signal power at the eavesdropper is zero. VILLAIN therefore guarantees perfect secrecy against active eavesdroppers,[1] and it does so without requiring wiretap coding. Using numerical simulations, we show that VILLAIN succeeds in mitigating active eavesdroppers also when the noise at the BS is not negligible.

B. Notation

Column vectors and matrices are denoted by lowercase boldface (e.g., $\mathbf{a}$) and uppercase boldface (e.g. $\mathbf{A}$) letters, respectively. The transpose is denoted $(\cdot)^{\text{T}}$, the complex conjugate $(\cdot)^*$, the conjugate transpose $(\cdot)^{\text{H}}$, and the Moore-Penrose pseudoinverse $(\cdot)^\dagger$. The Frobenius-norm is $\|\cdot\|_F$, the 2-norm $\|\cdot\|_2$, and the absolute value $|\cdot|$. The subspace spanned by $\mathbf{a}$ is $\text{span}(\mathbf{a})$ and its orthogonal complement is $\text{span}(\mathbf{a})^\perp$. The circularly-symmetric complex Gaussian distribution with variance Q is $\mathcal{CN}(0,Q)$. The expectation operator is $\mathbb{E}[\cdot]$.

[1] By perfect secrecy, we mean that $H(s|y_{\text{ed}}) = H(s)$, where s and y_{ed} are defined in Sec. II, and where $H(\cdot)$ and $H(\cdot|\cdot)$ denote (conditional) entropy.

II. System Model

We consider the case where a B-antenna BS wants to transmit data (from a constellation $\mathcal{S}$ with unit average symbol energy) to a single-antenna user equipment (UE) in the presence of a single-antenna eavesdropper. The receive signals at the UE and the eavesdropper can be written as

$$y_{\text{ue}} = \mathbf{h}^{\text{T}}\mathbf{x} + n_{\text{ue}}, \tag{1}$$

$$y_{\text{ed}} = \mathbf{j}^{\text{T}}\mathbf{x} + n_{\text{ed}}, \tag{2}$$

respectively. Here, $\mathbf{x} \in \mathbb{C}^B$ is the BS transmit vector that must satisfy a power constraint $\mathbb{E}\big[\|\mathbf{x}\|_2^2\big] \leq P$, $\mathbf{h}^{\text{T}}, \mathbf{j}^{\text{T}} \in \mathbb{C}^B$ are the downlink channel vectors (which include the effects of large-scale as well as of small-scale fading) between the BS and the UE and the eavesdropper, respectively, and $n_{\text{ue}} \sim \mathcal{CN}(0, \mathsf{N}_{\text{ue}})$ and $n_{\text{ed}} \sim \mathcal{CN}(0, \mathsf{N}_{\text{ed}})$ model the noise at the UE and the eavesdropper, respectively. The transmit vector $\mathbf{x}$ is a linear function of the data symbol $s \in \mathcal{S}$ to be sent to the UE, i.e.,

$$\mathbf{x} = \mathbf{w}s, \tag{3}$$

where $\mathbf{w}$ is the BS's precoding vector.[2] This precoding vector has two objectives: First, it should ensure that the UE can easily recover s based on y_{ue}. Second, it should ensure that the eavesdropper *cannot* recover s based on y_{ed} (not even if the eavesdropper knows both $\mathbf{j}$ and $\mathbf{w}$).

The BS determines its precoding vector $\mathbf{w}$ based on a pilot phase in which the UE transmits a length-T pilot sequence $\mathbf{s}_T \in \mathbb{C}^T$ that is known to the BS. In other words, $\mathbf{w} = f(\mathbf{Y}_T)$ for some function $f: \mathbb{C}^{B\times T} \to \mathbb{C}^B$, where $\mathbf{Y}_T \in \mathbb{C}^{B\times T}$ is the BS's pilot receive signal. We assume that $\mathbf{s}_T$ is *secret* (i.e., unknown to the eavesdropper) and potentially *random*.[3]

In a no-eavesdropper or passive eavesdropper scenario, the receive matrix $\mathbf{Y}_T$ from the pilot phase can be written as

$$\mathbf{Y}_T = \mathbf{h}\mathbf{s}_T^{\text{T}} + \mathbf{N}_T, \tag{4}$$

where the UE uplink channel vector $\mathbf{h}$ is the transpose of the UE downlink channel vector due to channel reciprocity, and where $\mathbf{N}_T \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, \mathsf{N}_{\text{bs}})$ models the receive noise at the BS.

The precoding vector $\mathbf{w}$ is often determined by first forming an estimate $\hat{\mathbf{h}} = g(\mathbf{Y}_T)$ of the UE channel vector $\mathbf{h}$, and then setting $\mathbf{w} = h(\hat{\mathbf{h}})$ for some function h (i.e., $f = h \circ g$). A classic example is least squares (LS) channel estimation

$$\hat{\mathbf{h}} = g(\mathbf{Y}_T) = \mathbf{Y}_T(\mathbf{s}_T^{\text{T}})^{\dagger} \tag{5}$$

followed by maximum ratio transmission (MRT) precoding

$$\mathbf{w} = h(\hat{\mathbf{h}}) = \sqrt{P}\,\hat{\mathbf{h}}^{*}/\|\hat{\mathbf{h}}\|_2. \tag{6}$$

The UE can then simply estimate s by rescaling y_{ue} as

$$\hat{s} = \beta y_{\text{ue}}, \tag{7}$$

where $\beta = 1/|\mathbf{h}^{\text{T}}\mathbf{w}|$ recovers the scale of the transmit signal.[4] With these choices of $\mathbf{w}$ and β, we have $\hat{s} \to s$ as $\|\mathbf{N}_T\|_F \to 0$.

In this paper, we consider an *active* eavesdropper that transmits a signal $\mathbf{z} \in \mathbb{C}^T$ during the pilot phase to make the BS use a precoding vector $\mathbf{w}$ that makes it easy for the eavesdropper to detect s based on y_{ed}. Thus, the pilot receive signal does not have the form of (4), but instead can be written as

$$\mathbf{Y}_T = \mathbf{h}\mathbf{s}_T^{\text{T}} + \mathbf{j}\mathbf{z}^{\text{T}} + \mathbf{N}_T. \tag{8}$$

In the active eavesdropper literature, it is typically assumed that the eavesdropper knows the pilot sequence $\mathbf{s}_T$. In that case, it is natural for the eavesdropper to also transmit the pilot sequence (potentially at higher power), i.e., $\mathbf{z}^{\text{T}} = \alpha\mathbf{s}_T^{\text{T}}$ for some $\alpha \geq 1$. A BS that uses an LS channel estimator as in (5) with MRT beamforming as in (6) will then effectively form a least-square estimate not of $\mathbf{h}$, but of $\mathbf{h} + \alpha\mathbf{j}$. As $\alpha \to \infty$, this becomes a (scaled) estimate of $\mathbf{j}$, so that the BS optimizes its precoding vector for the eavesdropper instead of the UE and, consequently, forms its beam towards the eavesdropper rather than towards the UE.

In contrast, we assume the eavesdropper does not know $\mathbf{s}_T$, so that $\mathbf{z}$ cannot depend on $\mathbf{s}_T$. However, the eavesdropper can still influence the BS to its advantage by sending $\mathbf{z} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, Q)$ for some $Q > 0$. In that case, the LS channel estimator in (5) gives $\mathbf{h}\mathbf{s}_T^{\text{T}}(\mathbf{s}_T^{\text{T}})^{\dagger} = \mathbf{h}$ in the UE term, but $\mathbf{j}\mathbf{z}^{\text{T}}(\mathbf{s}_T^{\text{T}})^{\dagger} = \omega\mathbf{j}$ in the eavesdropper term, where $\omega \sim \mathcal{CN}(0, Q/\|\mathbf{s}_T\|_2^2)$. So, even if the eavesdropper transmits Gaussian noise, a BS with a LS channel estimator effectively forms a least-square estimate of $\mathbf{h} + \omega\mathbf{j}$, where ω is complex Gaussian with variance $Q/\|\mathbf{s}_T\|_2^2$. If $Q \gg \|\mathbf{s}_T\|_2^2$, then with high probability $|\omega| > 1$, so that the BS chooses $\mathbf{w}$ mainly as a function of the eavesdropper channel vector $\mathbf{j}$ and not the UE channel vector $\mathbf{h}$.

III. VILLAIN: A Channel Estimator for Active Eavesdropper Mitigation

Before we present VILLAIN, a remark is in order. VILLAIN is a channel estimator—how can a channel estimator mitigate an eavesdropper? The answer is that VILLAIN does not produce an unbiased estimate of $\mathbf{h}$, but one that is projected onto (an estimate of) the orthogonal complement span$(\mathbf{j})^{\perp}$ of the eavesdropper subspace span$(\mathbf{j})$. The following result shows that if such a channel estimate is combined with MRT precoding as in (6), then the eavesdropper receives no signal.

Proposition 1. *If the BS obtains a channel estimate $\hat{\mathbf{h}} \neq \mathbf{0}$ from the image of the projection $\mathbf{P} = \mathbf{I} - \mathbf{j}\mathbf{j}^{\dagger}$ onto span$(\mathbf{j})^{\perp}$, (i.e., $\hat{\mathbf{h}} = \mathbf{P}\tilde{\mathbf{h}}$ for some $\tilde{\mathbf{h}} \in \mathbb{C}^B$) and uses MRT precoding as in* (6)*, then the eavesdropper receives no signal,* $\mathbf{j}^{\text{T}}\mathbf{w} = 0$.

[2] Since we assume that the constellation $\mathcal{S}$ has unit average symbol energy, the power constraint $\mathbb{E}\big[\|\mathbf{x}\|_2^2\big] \leq P$ is equivalent to $\|\mathbf{w}\|_2^2 \leq P$.

[3] To prevent the eavesdropper from learning the pilot sequence, $\mathbf{s}_T$ should be changed every coherence time. For information-theoretic security, this would require reading from a one-time pad (OTP). However, this OTP could be much shorter than a message-encrypting OTP, since pilots are transmitted infrequently. Moreover, even if one abandons information-theoretic security and uses a cryptographic random number generator (CRNG) to update $\mathbf{s}_T$, the resulting security is much better than in classical cryptography, since the eavesdropper has to break the CRNG *in real time* to create an attack opening (while in classical cryptography, the received message can be stored and cracked offline).

[4] This way of expressing the scaling factor β assumes—unrealistically—that the UE knows both $\mathbf{h}$ and $\mathbf{w}$. However, β depends primarily on the large-scale fading between the UE and the BS, which changes slowly in time. We therefore assume that the UE can estimate β from the receive signals [21].

All proofs are in the Appendix. Motivated by this result, we now present VILLAIN. In VILLAIN, the UE sends a *redundant* pilot sequence, i.e., a pilot sequence of length $T > 1$. As we will see, this redundancy allows the BS to estimate two things: the eavesdropper subspace and the projection of the UE's channel onto the orthogonal complement of the eavesdropper subspace.

Let the pilot phase be given as in (8). Then VILLAIN solves

$$\min_{\substack{\tilde{\mathbf{h}}\in\mathbb{C}^B,\\ \tilde{\mathbf{P}}\in\mathscr{G}_{B-1}(\mathbb{C}^B)}} \|\tilde{\mathbf{P}}\mathbf{Y}_T - \tilde{\mathbf{h}}\mathbf{s}_T^{\mathrm{T}}\|_F^2. \tag{9}$$

Here, $\mathscr{G}_{B-1}(\mathbb{C}^B) = \{\mathbf{I}_B - \mathbf{a}\mathbf{a}^\dagger : \mathbf{a}\in\mathbb{C}^B\}$ is the Grassmannian manifold, i.e., the set of orthogonal projections onto $(B-1)$-dimensional subspaces of $\mathbb{C}^B$. The channel estimate $\hat{\mathbf{h}}$ that is obtained from solving (9) can then be used for MRT precoding as in (6). Even though the problem in (9) is non-convex, it has a closed-form solution:

Proposition 2. *The problem in* (9) *is solved by*

$$\hat{\mathbf{P}} = \mathbf{I}_B - \mathbf{u}\mathbf{u}^{\mathrm{H}} \quad \textit{and} \quad \hat{\mathbf{h}} = \hat{\mathbf{P}}\mathbf{Y}_T(\mathbf{s}_T^{\mathrm{T}})^\dagger, \tag{10}$$

where $\mathbf{u}\in\mathbb{C}^B$ *is the left singular vector which corresponds to the largest singular value of* $\mathbf{Y}_T(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger\mathbf{s}_T^{\mathrm{T}})$.

In (10), $\mathbf{u}$ should be understood as an estimate of the eavesdropper subspace (i.e., $\mathbf{u}\approx\alpha\mathbf{j}$ for some $\alpha\in\mathbb{C}$), $\hat{\mathbf{P}}$ is the orthogonal projection onto the orthogonal complement of that subspace (i.e., $\hat{\mathbf{P}}\approx\mathbf{I}_B-\mathbf{j}\mathbf{j}^\dagger$), and $\hat{\mathbf{h}}$ is an estimate of the projection of the UE's channel vector onto the orthogonal complement of the eavesdropper subspace (i.e., $\hat{\mathbf{h}}\approx(\mathbf{I}_B-\mathbf{j}\mathbf{j}^\dagger)\mathbf{h}$). If the pilot sequence $\mathbf{s}_T$ is chosen at random and unknown to the eavesdropper, and if there is no noise at the BS, then we have the following guarantee:

Theorem 1. *Assume that* $T>1$, *that* $\mathbf{s}_T\overset{\text{i.i.d.}}{\sim}\mathcal{CN}(0,1)$ *and* $\mathbf{z}\neq\mathbf{0}$ *are independent of each other, and that* $\mathsf{N}_{\text{bs}}=0$. *Then, almost surely, the precoding vector* $\mathbf{w}$ *that results from the VILLAIN channel estimator in conjunction with MRT precoding solves*

$$\max_{\tilde{\mathbf{w}}\in\mathbb{C}^B:\|\tilde{\mathbf{w}}\|_2^2\le P} |\mathbf{h}^{\mathrm{T}}\tilde{\mathbf{w}}|^2 \quad \textit{such that} \quad \mathbf{j}^{\mathrm{T}}\tilde{\mathbf{w}}=0. \tag{11}$$

That is, $\mathbf{w}$ *achieves the highest UE receive signal power of all vectors that achieve zero eavesdropper receive signal power. The delivered signal power is* $|\mathbf{h}^{\mathrm{T}}\mathbf{w}|^2 = \|(\mathbf{I}_B-\mathbf{j}\mathbf{j}^\dagger)\mathbf{h}\|_2^2 P$.

Note that the mitigation strategy pursued here—maximizing the receive power at the receiver while ensuring that the eavesdropper receives *no* signal—is suboptimal in the sense that it generally does not achieve the secrecy capacity [22]. However, VILLAIN has the following advantages: First, under the conditions of Thm. 1, it achieves perfect secrecy *without* requiring a priori information of the UE's or the eavesdropper's channel. Second, it then achieves perfect secrecy using off-the-shelf MRT precoding, without requiring wiretap coding. The only price to be paid is a randomized pilot phase of length at least two, and computing the VILLAIN channel estimate in Prop. 2. Its computational complexity is dominated by the SVD, whose complexity scales with $\max\{B,T\}\min\{B,T\}^2$.

IV. Simulation Results

A. Line-of-Sight Channel Without Noise at the BS

For illustrative purposes, we start by considering a textbook line-of-sight (LoS) channel, where a BS with $B=8$ antennas arranged as a uniform linear array (ULA) with antennas spaced at half a wavelength is located at the coordinate origin, while the UE and the eavesdropper are located in far-field at degrees of $\theta_{\text{UE}}=70^\circ$ and $\theta_{\text{ED}}=20^\circ$, respectively. We assume that there is no noise at the BS ($\mathsf{N}_{\text{bs}}=0$), that the BS power constraint is $P=1$, and that the channel gains of the UE and the eavesdropper are $\|\mathbf{h}\|_2=\|\mathbf{j}\|_2=1$. A textbook LoS channel vector $\mathbf{g}$ with unit gain can be written in dependence of the angle ϕ (relative the the ULA) as

$$\mathbf{g}(\phi) = \frac{1}{\sqrt{B}}\Big[1, e^{-i\pi\cos(\phi)}, \dots, e^{-i\pi\cos(\phi)(B-1)}\Big]^{\mathrm{T}}, \tag{12}$$

and we have $\mathbf{h}=\mathbf{g}(\theta_{\text{UE}})$ and $\mathbf{j}=\mathbf{g}(\theta_{\text{ED}})$.

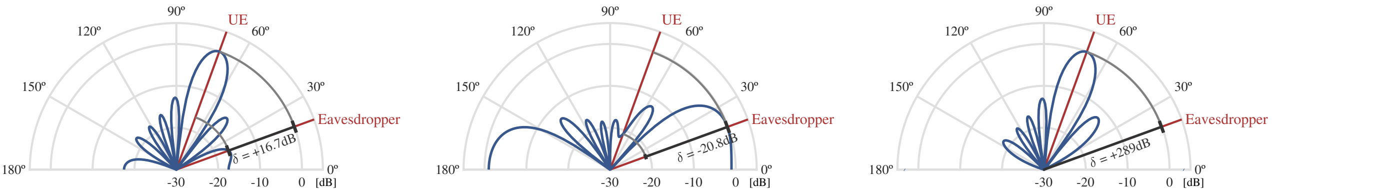
We consider three different scenarios in which the BS computes its precoding vector $\mathbf{w}$. For each of the scenarios, Fig. 1 shows the receive power $\mathbb{E}\big[|\mathbf{g}^{\mathrm{T}}(\phi)\mathbf{x}|^2\big]=|\mathbf{g}^{\mathrm{T}}(\phi)\mathbf{w}|^2$ (in dB) as a function of ϕ. For each of the scenarios, we also compute the *advantage* δ, which we define as the ratio between the power received at the UE and the power received at the eavesdropper,

$$\delta \triangleq \frac{|\mathbf{h}^{\mathrm{T}}\mathbf{w}|^2}{|\mathbf{j}^{\mathrm{T}}\mathbf{w}|^2}. \tag{13}$$

If the noise at the UE and the eavesdropper are equally strong, $\mathsf{N}_{\text{ue}}=\mathsf{N}_{\text{ed}}$, then $\delta>1$ implies that the secrecy capacity *for that precoding vector* $\mathbf{w}$ is positive; $\delta\le1$ implies that it is zero. The three considered scenarios are as follows:

1) Passive Eavesdropper and LS channel estimation: In this scenario, the eavesdropper does not transmit during the pilot phase in which the UE sends a pilot sequence of length $T=8$. The BS uses the LS channel estimator of (5) with the MRT precoder of (6). The signal receive strength (as a function of ϕ) for the resulting precoder is shown in Fig. 1(a). The receive strength is highest at $\phi=\theta_{\text{UE}}$, where a gain of 0 dB is achieved. In contrast, the receive strength at the eavesdropper is significantly lower, and an advantage of $\delta=+16.7$ dB is achieved. This confirms the claim of [8]–[10] that multi-antenna precoding protects naturally against passive eavesdroppers.

2) Active Eavesdropper: This scenario differs from the previous one in that the eavesdropper transmits i.i.d. circularly-symmetric Gaussian samples (independent of $\mathbf{s}_T$, so that the eavesdropper need not know $\mathbf{s}_T$) during the pilot phase, at 25 dB higher expected power than $\mathbf{s}_T$. The BS uses the same LS channel estimator of (5) and MRT-precoder of (6) as in the first scenario. The signal receive strength for the resulting precoding is shown in Fig. 1(b). Since the pilot receive signal is dominated by the eavesdropper, the signal receive strength is highest in the direction of the eavesdropper, where a gain of almost 0 dB is achieved. In contrast, the signal receive strength at the UE is much lower, resulting in a negative gain of $\delta=-20.8$ dB. This result shows a clear need for active eavesdropper mitigation if physical-layer security is desired.



(a) Passive eavesdropper and LS channel est.

(b) Active eavesdropper and LS channel est.

(c) Active eavesdropper and VILLAIN channel est.

Fig. 1. Signal receive power (in dB) as a function of the incidence angle ϕ for three different scenarios. The BS has $B = 8$ antennas arranged as a ULA. We assume a textbook LoS channel with the UE and the eavesdropper being in far-field at $70°$ and $20°$, respectively. The performance of a precoding can be characterized in terms of the ratio δ between the signal receive power at the UE and at the eavesdropper. VILLAIN effectively mitigates the active eavesdropper.

3) Active Eavesdropper with VILLAIN: The third scenario is identical to the second one, except that the BS now uses the VILLAIN channel estimator (together with the MRT precoder of (6)). The signal receive strength of the resulting precoding vector is plotted in Fig. 1(c). The results show that VILLAIN succeeds in mitigating the active eavesdropper: Even though the pilot receive signal is dominated by the eavesdropper's contribution, the signal receive strength is highest at the UE, where a gain of almost 0 dB is achieved. In contrast, the receive strength at the eavesdropper is now so low that it does not even show on the axis, and an advantage of $\delta = +289$ dB is achieved. In theory, the advantage in such a noiseless scenario should be $\delta = +\infty$ dB, but the floating-point accuracy of MATLAB simulations limits the advantage to a finite value.

B. QuaDRiGa UMa Channels With Noise at the BS

We now simulate a more realistic scenario that considers noise at the BS and uses channel vectors that are generated using QuaDRiGa [23] with a 3GPP 38.901 urban macrocellular (UMa) channel model [24]. The carrier frequency is 2 GHz, the BS has $B = 16$ antennas arranged as a ULA spaced at half a wavelength, and the UE and the eavesdropper are uniformly and independently placed at a distance between 10 m and 100 m in a $120°$ sector in front of the BS. We compare the performance of VILLAIN to the performance of a conventional LS channel estimator. In both cases, the UE transmits an i.i.d. $\mathcal{CN}(0, E_s)$ pilot sequence of length $T = 4$. The eavesdropper transmits i.i.d. $\mathcal{CN}(0, Q)$ samples at 30 dB higher transmit power than the UE during the pilot phase (i.e., $10\log_{10}(Q/E_s) = 30$ dB). We quantify the BS noise variance N_{bs} relative to the *transmit* signal power E_s, where we define the signal-to-noise ratio as $\mathit{SNR} = E_s/\mathsf{N}_{\text{bs}}$. We consider three different noise levels: $\mathit{SNR} = 0$ dB, $\mathit{SNR} = 15$ dB, and $\mathit{SNR} = 30$ dB. Fig. 2 shows the cumulative distribution function (CDF) for the different channel estimators and the different SNRs.[5] We see that the SNR is irrelevant when using LS channel estimation—the eavesdropper dominates the receive signal. For each SNR, the eavesdropper achieves a negative advantage (in dB) in around 75% of the cases, and the median advantage is $\delta = -16$ dB. The performance of VILLAIN is far superior and increases with SNR: Already for a 0 dB SNR, VILLAIN achieves a positive advantage (in dB) in 94% of cases and a median advantage of $\delta = +26$ dB. For an SNR of 15 dB and of 30 dB, VILLAIN achieves a positive advantage (in dB) in more than 99% of cases, and a median advantage of $\delta = +51$ dB and $\delta = +68$ dB, respectively. These results show that VILLAIN succeeds in mitigating active eavesdroppers also in the presence of noise.

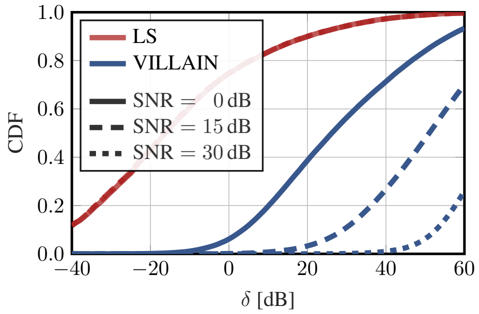


Fig. 2. CDF of VILLAIN and LS channel estimation for different noise levels at the BS. The eavesdropper's transmit signal is 30dB stronger than the UE's.

Acknowledgment

The authors thank A. Stutz for comments and suggestions.

Appendix A
Proofs

A. Proof of Prop. 1

If $\hat{\mathbf{h}} = \mathbf{P}\tilde{\mathbf{h}}$ for some $\tilde{\mathbf{h}} \in \mathbb{C}^B$ and $\mathbf{w} = \sqrt{P}\hat{\mathbf{h}}^*/\|\hat{\mathbf{h}}\|_2$, then we can write

$$\|\hat{\mathbf{h}}\|_2\mathbf{j}^{\mathrm{T}}\mathbf{w} = \sqrt{P}\mathbf{j}^{\mathrm{T}}(\mathbf{P}\tilde{\mathbf{h}})^* = \sqrt{P}\mathbf{j}^{\mathrm{T}}\mathbf{P}^*\tilde{\mathbf{h}}^* \tag{14}$$

$$\overset{(a)}{=} \sqrt{P}\mathbf{j}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\tilde{\mathbf{h}}^* = \sqrt{P}(\mathbf{P}\mathbf{j})^{\mathrm{T}}\tilde{\mathbf{h}}^* \overset{(b)}{=} \mathbf{0}^{\mathrm{T}}\tilde{\mathbf{h}}^* = 0, \tag{15}$$

where (a) follows because $\mathbf{P}$ is an orthogonal projection and so $\mathbf{P}^{\mathrm{H}} = \mathbf{P}$ and (b) follows because $\mathbf{P}\mathbf{j} = \mathbf{0}$. From this, the result follows by dividing both sides by $\|\hat{\mathbf{h}}\|_2 \neq 0$. ■

B. Proof of Prop. 2

For given $\tilde{\mathbf{P}}$, the problem in (9) is quadratic in $\tilde{\mathbf{h}}$, and is minimized by $\tilde{\mathbf{h}} = \tilde{\mathbf{P}}\mathbf{Y}_T(\mathbf{s}_T^{\mathrm{T}})^\dagger$. Plugging this back into (9) gives

$$\min_{\tilde{\mathbf{P}}\in\mathscr{G}_{B-I}(\mathbb{C}^B)} \|\tilde{\mathbf{P}}\mathbf{Y}_T(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger\mathbf{s}_T^{\mathrm{T}})\|_F^2, \tag{16}$$

which is minimized by $\tilde{\mathbf{P}} = \mathbf{I}_B - \mathbf{u}\mathbf{u}^{\mathrm{H}}$, where $\mathbf{u} \in \mathbb{C}^B$ is the left singular vector corresponding to the largest singular value of $\mathbf{Y}_T(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger\mathbf{s}_T^{\mathrm{T}})$ [25]. By plugging this value for $\tilde{\mathbf{P}}$ back into $\tilde{\mathbf{h}} = \tilde{\mathbf{P}}\mathbf{Y}_T(\mathbf{s}_T^{\mathrm{T}})^\dagger$, the result follows. ■

[5] These CDFs take into account the large channel gain variation between the UE and the eavesdropper that results from the random distance to the BS.

C. Proof of Thm. 1

Since we assume $\mathsf{N}_{\mathrm{bs}} = 0$, we have $\mathbf{Y}_T = \mathbf{h}\mathbf{s}_T^{\mathrm{T}} + \mathbf{j}\mathbf{z}^{\mathrm{T}}$. By Prop. 2, the optimal $\tilde{\mathbf{P}}$ equals $\mathbf{I}_B - \mathbf{u}\mathbf{u}^{\mathrm{H}}$, where $\mathbf{u}$ is the left singular vector corresponding to the largest singular value of

$$\mathbf{Y}_T(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger \mathbf{s}_T^{\mathrm{T}}) \tag{17}$$

$$= (\mathbf{h}\mathbf{s}_T^{\mathrm{T}} + \mathbf{j}\mathbf{z}^{\mathrm{T}})(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger \mathbf{s}_T^{\mathrm{T}}) \tag{18}$$

$$= \mathbf{h}\underbrace{\mathbf{s}_T^{\mathrm{T}}(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger \mathbf{s}_T^{\mathrm{T}})}_{=\mathbf{0}} + \mathbf{j}\mathbf{z}^{\mathrm{T}}(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger \mathbf{s}_T^{\mathrm{T}}) \tag{19}$$

$$= \mathbf{j}\mathbf{z}^{\mathrm{T}}(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger \mathbf{s}_T^{\mathrm{T}}). \tag{20}$$

In (20), $\mathbf{z}^{\mathrm{T}}(\mathbf{I}_T - (\mathbf{s}_T^{\mathrm{T}})^\dagger$ is the orthogonal projection of $\mathbf{z}$ onto $\mathrm{span}(\mathbf{s}_T)^\perp$, which is distinct from zero if $\mathbf{z} \notin \mathrm{span}(\mathbf{s}_T)$. Under our assumptions that $T > 1$, that $\mathbf{s}_T \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0,1)$ and $\mathbf{z}\neq\mathbf{0}$ are independent of each other, we have $\mathbf{z} \notin \mathrm{span}(\mathbf{s}_T)$ almost surely. The rest of the proof is conditioned on this almost sure event. It follows that the subspace spanned by $\mathbf{u}$ contains $\mathbf{j}$, in which case the optimal $\tilde{\mathbf{P}}$ can be written as

$$\tilde{\mathbf{P}} = \mathbf{I}_B - \mathbf{u}\mathbf{u}^{\mathrm{H}} = \mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger. \tag{21}$$

By Prop. 2, the channel estimate by VILLAIN is therefore

$$\hat{\mathbf{h}} = (\mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger)\mathbf{Y}_T(\mathbf{s}_T^{\mathrm{T}})^\dagger \tag{22}$$

$$= (\mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger)(\mathbf{h}\mathbf{s}_T^{\mathrm{T}} + \mathbf{j}\mathbf{z}^{\mathrm{T}})(\mathbf{s}_T^{\mathrm{T}})^\dagger \tag{23}$$

$$= (\mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger)\mathbf{h}\underbrace{\mathbf{s}_T^{\mathrm{T}}(\mathbf{s}_T^{\mathrm{T}})^\dagger}_{=1} + \underbrace{(\mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger)\mathbf{j}}_{=\mathbf{0}}\mathbf{z}^{\mathrm{T}}(\mathbf{s}_T^{\mathrm{T}})^\dagger \tag{24}$$

$$= (\mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger)\mathbf{h}. \tag{25}$$

When using MRT precoding, we get the precoding vector

$$\mathbf{w} = (\sqrt{P}/\|\mathbf{P}\mathbf{h}\|_2)(\mathbf{P}\mathbf{h})^*, \tag{26}$$

where we define $\mathbf{P} \triangleq \mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger$ for the remainder of the proof.

We now show that the solution to the optimization problem in (11) coincides with (26). The set of vectors $\tilde{\mathbf{w}}$ that satisfy the constraint $\mathbf{j}^{\mathrm{T}}\tilde{\mathbf{w}} = 0$ is simply the image of $\mathbf{P}^*$ and can be rewritten by substituting $\tilde{\mathbf{w}}$ with $\mathbf{P}^*\bar{\mathbf{w}}$:

$$\{\tilde{\mathbf{w}} : \tilde{\mathbf{w}} \in \mathbb{C}^B, \mathbf{j}^{\mathrm{T}}\tilde{\mathbf{w}} = 0\} = \{\mathbf{P}^*\bar{\mathbf{w}} : \bar{\mathbf{w}} \in \mathbb{C}^B\}. \tag{27}$$

The problem in (11) can therefore be reformulated as

$$\max_{\bar{\mathbf{w}}\in\mathbb{C}^B : \|\mathbf{P}^*\bar{\mathbf{w}}\|_2^2 \le P} |\mathbf{h}^{\mathrm{T}}\mathbf{P}^*\bar{\mathbf{w}}|^2 = \max_{\|\mathbf{P}^*\bar{\mathbf{w}}\|_2^2 \le P} |\mathbf{h}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\bar{\mathbf{w}}|^2 \tag{28}$$

$$= \max_{\|\mathbf{P}^*\bar{\mathbf{w}}\|_2^2 \le P} |(\mathbf{P}\mathbf{h})^{\mathrm{T}}\bar{\mathbf{w}}|^2, \tag{29}$$

which is solved by $\bar{\mathbf{w}} \propto (\hat{\mathbf{P}}\mathbf{h})^*$. By reinserting $\tilde{\mathbf{w}} = \mathbf{P}^*\bar{\mathbf{w}}$, and respecting the power constraint, we get the $\tilde{\mathbf{w}}$ that solves (11):

$$\tilde{\mathbf{w}} = \sqrt{P}\frac{\mathbf{P}^*(\mathbf{P}\mathbf{h})^*}{\|\mathbf{P}^*(\mathbf{P}\mathbf{h})^*\|_2} = \sqrt{P}\frac{(\mathbf{P}\mathbf{h})^*}{\|\mathbf{P}\mathbf{h}\|_2}, \tag{30}$$

which coincides with (26). The signal power at the UE is

$$|\mathbf{h}^{\mathrm{T}}\mathbf{w}|^2 = P\frac{|\mathbf{h}^{\mathrm{T}}(\mathbf{P}\mathbf{h})^*|^2}{\|\mathbf{P}\mathbf{h}\|_2^2} = P\frac{|\mathbf{h}^{\mathrm{H}}\mathbf{P}\mathbf{h}|^2}{\|\mathbf{P}\mathbf{h}\|_2^2} = P\frac{|\mathbf{h}^{\mathrm{H}}\mathbf{P}^{\mathrm{H}}\mathbf{P}\mathbf{h}|^2}{\|\mathbf{P}\mathbf{h}\|_2^2} \tag{31}$$

$$= P\|\mathbf{P}\mathbf{h}\|_2^2 = P\|(\mathbf{I}_B - \mathbf{j}\mathbf{j}^\dagger)\mathbf{h}\|_2^2. \tag{32}$$

This concludes the proof. ■

References

[1] M. E. Whitman and H. J. Mattord, *Principles of Information Security*. Cengage learning, 2021.

[2] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge Univ. Press, 2011.

[3] Y. Wu, A. Khisti, C. Xiao, G. Caire, K.-K. Wong, and X. Gao, "A survey of physical layer security techniques for 5G wireless networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 679–695, 2018.

[4] Z. Li, W. Trappe, and R. Yates, "Secret communication via multi-antenna transmission," in *Proc. 41st Ann. Conf. Inf. Sci. Syst.*, 2007, pp. 905–910.

[5] S. Shafiee and S. Ulukus, "Achievable rates in Gaussian MISO channels with secrecy constraints," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2007, pp. 2466–2470.

[6] Z. Rezki and M.-S. Alouini, "On the finite-SNR diversity-multiplexing tradeoff of zero-forcing transmit scheme under secrecy constraint," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2011, pp. 1–5.

[7] H. Reboredo, J. Xavier, and M. R. Rodrigues, "Filter design with secrecy constraints: The MIMO Gaussian wiretap channel," *IEEE Trans. Signal Process.*, vol. 61, no. 15, pp. 3799–3814, 2013.

[8] D. Kapetanovic, G. Zheng, and F. Rusek, "Physical layer security for massive MIMO: An overview on passive eavesdropping and active attacks," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 21–27, 2015.

[9] A. Bereyhi, S. Asaad, R. R. Müller, R. F. Schaefer, and A. M. Rabiei, "On robustness of massive MIMO systems against passive eavesdropping under antenna selection," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–7.

[10] A. Bereyhi, S. Asaad, R. R. Müller, R. F. Schaefer, G. Fischer, and H. V. Poor, "Robustness of low-complexity massive MIMO architectures against passive eavesdropping," *arXiv preprint arXiv:1912.02444*, 2019.

[11] X. Zhou, B. Maham, and A. Hjørungnes, "Pilot contamination for active eavesdropping," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 903–907, 2012.

[12] Y. Wu, R. Schober, D. W. K. Ng, C. Xiao, and G. Caire, "Secure massive MIMO transmission with an active eavesdropper," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3880–3900, 2016.

[13] A. Bereyhi, S. Asaad, R. R. Müller, R. F. Schaefer, and H. V. Poor, "Secure transmission in IRS-assisted MIMO systems with active eavesdroppers," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2020, pp. 718–725.

[14] L. Li, A. P. Petropulu, and Z. Chen, "MIMO secret communications against an active eavesdropper," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 10, pp. 2387–2401, 2017.

[15] S. Cho, G. Chen, and J. P. Coon, "Zero-forcing beamforming for active and passive eavesdropper mitigation in visible light communication systems," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1495–1505, 2021.

[16] J. Si, Z. Cheng, Z. Li, J. Cheng, H.-M. Wang, and N. Al-Dhahir, "Cooperative jamming for secure transmission with both active and passive eavesdroppers," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5764–5777, 2020.

[17] S. Cho, G. Chen, and J. P. Coon, "Cooperative beamforming and jamming for secure vlc system in the presence of active and passive eavesdroppers," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 4, pp. 1988–1998, 2021.

[18] Y. Zhou, P. L. Yeoh, C. Pan, K. Wang, Z. Ma, B. Vucetic, and Y. Li, "Caching and UAV friendly jamming for secure communications with active eavesdropping attacks," vol. 71, no. 10, pp. 11 251–11 256, 2022.

[19] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, 2020.

[20] H. Jia, L. Ma, and S. Valaee, "STAR-RIS enabled downlink secure NOMA network under imperfect CSI of eavesdroppers," *IEEE Commun. Lett.*, vol. 27, no. 3, pp. 802–806, 2023.

[21] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Nonlinear 1-bit precoding for massive MU-MIMO with higher-order modulation," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, 2016, pp. 763–767.

[22] A. Khisti and G. W. Wornell, "Secure transmission with multiple antennas—Part I: The MISOME wiretap channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3088–3104, 2010.

[23] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, 2014.

[24] 3GPP, "3GPP TR 38.901," Mar. 2022, version 17.0.0.

[25] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

Data-Driven List Polar Decoder for Symmetric and Asymmetric Input Distributions

Ziv Aharoni
Ben-Gurion University
zivah@post.bgu.ac.il

Bashar Huleihel
Ben-Gurion University
basharh@post.bgu.ac.il

Henry D. Pfister
Duke University
henry.pfister@duke.edu

Haim H. Permuter
Ben-Gurion University
haimp@bgu.ac.il

***Abstract*—This paper introduces extensions to data-driven polar decoders, enabling list decoding and accommodating asymmetric input distributions. These are crucial steps to develop data-driven codes that 1) achieve capacity and 2) are competitive in moderate block lengths. We commence by integrating list decoding into the data-driven polar codes, which significantly alleviates the inherent error propagation issues associated with successive cancellation decoding. Secondly, we expand the applicability of these codes to channels with stationary, non-uniform input distributions by incorporating the Honda-Yamamoto scheme. Both modifications are computationally efficient and do not require an explicit channel model. Numerical results validate the efficacy of our contributions, which offer a robust and versatile coding mechanism for various channel conditions.**

I. Introduction

Polar codes allow the construction of capacity-achieving codes for symmetric binary-input memoryless channels [1]. When given N independent copies of a binary discrete memoryless channel (DMC) W, successive cancellation (SC) decoding induces a new set of N binary effective channels $W_N^{(i)}$. Channel polarization is the phenomenon whereby, for N sufficiently large, almost all of the effective bit channels $W_N^{(i)}$ have capacities close to 0 or 1. Specifically, the fraction of channels with capacity close to 1 approaches $\mathsf{I}(W)$ and the fraction of channels with capacity close to 0 approaches $1-\mathsf{I}(W)$, where $\mathsf{I}(W)$ is the channel's symmetric capacity. The construction of polar codes involves choosing which rows to keep from the square generator matrix given by Arikan's transform [1, Section VII]. The encoding and decoding procedures are performed by recursive formulas whose computational complexity is $O(N\log N)$.

Polar codes can also be applied to finite state channels (FSCs). Arikan's transform also polarizes the bit channels $W_N^{(i)}$ in the presence of memory [2], and thus the encoding algorithm is the same as if the channel is memoryless. However, the decoding algorithm needs to be updated since the derivation of the SC decoder in [1] relies on the memoryless property. To account for the channel memory, the channel outputs are represented by a trellis, whose nodes capture the information of the channel's memory. This trellis was embedded into the SC decoding algorithm to yield the successive cancellation trellis (SCT) decoding algorithm [3], [4].

However, the SCT decoder is only applicable when the channel model is known and when the channel's state alphabet size is finite and relatively small. For FSCs, the computational complexity of the SCT decoder is $O(|\mathcal{S}|^3 N\log N)$, where $|\mathcal{S}|$ is the number of channel states. For Markov channels where the set of channel states is not finite, the SCT decoder is not applicable without quantization of its states. With quantization, there may be a strong tension between the computational complexity and the error introduced by quantization. Additionally, the SCT decoder cannot be used for an unknown channel with memory without first estimating the channel as it requires an explicit channel model.

The authors of [5] proposed a novel methodology for data-driven polar decoders. The methodology uses a neural SC (NSC) decoder, which uses four distinct neural networks (NNs) instead of the elementary operations of the SC decoder. Specifically, the NNs approximate the channel's output statistics, the check-node, the bit-node, and the soft decision operations, denoted by E, F, G, H, respectively. The parameters of E, F, G, H are determined in a training phase, in which the mutual information (MI) of the effective channels $W_N^{(i)}$ is estimated. After the training phase, the set of "clean" effective channels are determined by a Monte Carlo (MC) evaluation of the MI of the effective bit channels to complete the code design. The main advantage of this scheme is 1) its computational complexity does not grow cubicly with the channel's state alphabet size, and 2) it does not require an explicit channel model.

However, despite the fact that polar codes are capacity achieving, their performance under SC decoding are inferior to low density parity check (LDPC) and turbo codes at moderate block lengths. One of the reasons for that, as identified in [6], is that in SC decoding, decoding errors at early stages of the decoding procedure propagate to the succeeding bits, which yields in sub-optimal performance. Hence, the authors of [6] design a successive cancellation list (SCL) decoder for polar codes that instead of decoding a single codeword, as in the SC decoder, it decodes L codewords. Then, the decoder chooses one codeword from the list with the highest likelihood[1]. The performance of the SCL decoder improved dramatically towards the performance of the maximum likelihood (ML) decoder, and accordingly it is now part of the 5G standard. Therefore, it is of great interest to examine the performance of data-driven polar codes with list decoding, which is the first goal of this paper.

An additional issue to be addressed when designing capacity achieving codes is to accommodate data-driven polar codes

[1] The authors of [6] also showed the cyclic redundancy check (CRC) bits can be used as side information shared between the decoder and the encoder that allows to choose the correct codeword by checking which word in the list passes the CRC.

with asymmetric input distributions, as the capacity achieving input distribution is not necessarily uniform independently identically distributed (i.i.d.). In that regard, this paper provides an extension of data driven polar codes for stationary input distributions by incorporating the Honda-Yamamoto scheme [7] into the methodology of data-driven polar codes. This is the second goal of the paper.

The paper is organized as follows. Section II defines the notation and gives the necessary background on polar codes. Specifically, it presents polar codes as given in [1], and data-driven polar codes, as given in [5]. Section III extends data-driven polar codes to stationary input distributions. Section IV presents the idea of list decoding and its application to data-driven polar codes. Section V presents the numerical experiments.

II. Notations and Preliminaries

Throughout this paper, we denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space on which all random variables are defined, with $\mathbb{E}$ denoting expectation. Random variables (RVs) are denoted by capital letters and their realizations are denoted by lower-case letters. Calligraphic letters denote sets, e.g. $\mathcal{X}$. We use the notation X^n to denote the RV $(X_1, X_2, \dots, X_n)$ and x^n to denote its realization. The probability $\Pr[X = x]$ is denoted by $P_X(x)$. Stochastic processes are denoted by blackboard bold letters, e.g., $\mathbb{X} := (X_i)_{i\in\mathbb{N}}$. An n-coordinate projection of $\mathbb{P}$ is denoted by $P_{X^nY^n} := \mathbb{P}|_{\sigma(X^n,Y^n)}$, where $\sigma(X^n, Y^n)$ is the σ-algebra generated by (X^n, Y^n). We denote by $[N]$ the set of integers $\{1, \dots, N\}$. The MI between two RVs X, Y is denoted by $\mathsf{I}(X;Y)$.For two distributions P, Q, the cross entropy (CE) is denoted by $h_{\mathsf{CE}}(P, Q)$, the entropy is denoted by $\mathsf{H}(P)$ and the Kullback Leibler (KL) divergence is denoted by $\mathsf{D}_{\mathsf{KL}}(P\|Q)$. The notation $P \ll Q$ indicates that P is absolutely continuous with respect to Q.

The tuple $\left(W_{Y|X}, \mathcal{X}, \mathcal{Y}\right)$ defines a memoryless channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and a transition kernel $W_{Y|X}$. Throughout the paper, we assume that $\mathcal{X} = \{0, 1\}$. For a memoryless channel, we denote its input distribution by $P_X = P_{X_i}$ for all $i \in \mathbb{Z}$. The tuple $\left(W_{Y\|X}, \mathcal{X}, \mathcal{Y}\right)$ defines a time invariant channel with memory, where $W_{Y\|X} = \left\{W_{Y_0|Y^{-1}_{-i+1},X^0_{-i+1}}\right\}_{i\in\mathbb{N}}$. The term $W_{Y^N\|X^N} = \prod_{i=1}^N W_{Y_0|Y^{-1}_{-i+1},X^0_{-i+1}}$ denotes the probability of observing Y^N causally conditioned on X^N [8]. The symmetric capacity of a channel is denoted by $\mathsf{I}(W)$. We denote by $\mathcal{D}_{M,N} = \{x_{j,i}, y_{j,i}\}_{j\in[M],i\in[N]} \sim P_{X^{MN}} \otimes W_{Y^{MN}\|X^{MN}}$ a finite sample of pairs of input-output vectors for M consecutive blocks of N symbols, where $x_{j,i}, y_{j,i}$ denotes the i-th input and output of the j-th block.

A. Polar Codes for Symmetric Channels

Let $G_N = B_N F^{\otimes n}$ be Arikan's polar transform with the generator matrix for block length $N = 2^n$ for $n \in \mathbb{N}$. The matrix B_N is the permutation matrix associated with the bit-reversal permutation. It is defined by the recursive relation $B_N = R_N(I_2 \otimes B_{\frac{N}{2}})$ starting from $B_2 = I_2$. The term I_N denotes the identity matrix of size N and R_N denotes a permutation matrix called reverse-shuffle [1]. The term $A \otimes B$ denotes the Kronecker product of A and B when A, B are matrices, and it denotes a tensor product whenever A, B are distributions. The term $A^{\otimes N} := A \otimes \cdots \otimes A$ denotes an application of the $\otimes$ operator N times.

We define a polar code by the tuple $\left(\mathcal{X}, \mathcal{Y}, W, E^W, F, G, H\right)$ that contains the channel W, the channels embedding E^W and the core components of the SC decoder, F, G, H. We define the effective bit channels by the tuple $\left(W_N^{(i)}, \mathcal{X}, \mathcal{X}^{i-1} \times \mathcal{Y}^N\right)$ for all $i \in [N]$. The term $E^W : \mathcal{Y} \to \mathcal{E}$ denotes the channel embedding, where $\mathcal{E} \subset \mathbb{R}^d$. For example, for a memoryless channel $W := W_{Y|X}$, a valid choice of E^W, as used in the remainder of this paper, is given by the following:

$$E^W(y) = \log\frac{W(y|1)}{W(y|0)} + \log\frac{P_X(1)}{P_X(0)}, \tag{1}$$

where the second term in the right-hand-side (RHS) cancels out in the case where P_X is uniform.

The functions $F : \mathcal{E} \times \mathcal{E} \to \mathcal{E}$, $G : \mathcal{E} \times \mathcal{E} \times \mathcal{X} \to \mathcal{E}$ denote the check-node and bit-node operations, respectively. We denote by $H : \mathcal{E} \to [0, 1]$ a mapping of the embedding into a probability value, i.e. a soft decision. For the choice of E^W in (1), F, G, H are given by

$$\begin{aligned} &F(e_1, e_2) = 2\tanh^{-1}\left(\tanh\frac{e_1}{2}\tanh\frac{e_2}{2}\right), \\ &G(e_1, e_2, u) = e_2 + (-1)^u e_1, \\ &H(e_1) = \sigma(e_1), \end{aligned} \tag{2}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic function and $e_1, e_2 \in \mathcal{E}, u \in \mathcal{X}$. For this choice, the hard decision rule $h : [0, 1] \to \mathcal{X}$ is the round function $h(l) = \mathbb{I}_{l>0.5}$, where $\mathbb{I}$ is the indicator function. Applying SC decoding on the channel outputs yields an estimate of the transmitted bits and their corresponding posterior distribution [1]. Specifically, after observing y^N, SC decoding performs the map $(y^N, f^N) \mapsto \left\{\hat{u}_i, P_{U_i|U^{i-1},Y^N}\left(1|\hat{u}^{i-1}, y^N\right)\right\}_{i\in[N]}$, where f^N are the frozen bits that are shared between the encoder and the decoder. That is, $f_i \in \{0, 1\}$ if $i \in [N]$ is frozen, and $f_i = 0.5$[2] if i is an information bit. This mapping is denoted by

$$\left\{\hat{u}_i, P_{U_i|U^{i-1},Y^N}\left(1|\hat{u}^{i-1}, y^N\right)\right\}_{i\in[N]} = \mathsf{SC}_{\mathsf{decode}}\left(y^N, f^N\right). \tag{3}$$

For more details on SC decoding, the reader may refer to [1, Section VIII].

B. Neural Successive Cancellation Decoder

A NSC decoder [5] is defined by the tuple $\left(\mathcal{X}, \mathcal{Y}, W, E^W_{\theta_1}, F_{\theta_2}, G_{\theta_3}, H_{\theta_4}\right)$, where $E^W_{\theta_1}, F_{\theta_2}, G_{\theta_3}, H_{\theta_4}$ are NNs with parameters $\theta_i \in \Theta \subset \mathbb{R}^p$ in a compact space Θ. For simplicity, we denote $\theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. The parameters θ are estimated in a training phase, in which the MI of the effective bit channels is estimated. The training phase includes the following steps. First, draw $x^N, y^N \sim \mathcal{D}_{M,N}$ and

[2] The value 0.5 is chosen arbitrarily to indicate that the bit needs to be decoded.

compute $u^N = x^N G_N$. Next, the functions $E_\theta^W, F_\theta, G_\theta, H_\theta$, are used to decode u^N with the SC decoding scheme, i.e. by applying $\mathsf{SC}_{\mathsf{decode}}\left(y^N, f^N\right)$ with $f^N = u^N$. This yields in an estimate of $\left\{P_{U_i|U^{i-1},Y^N}\left(1|u^{i-1},y^N\right)\right\}_{i\in[N]}$ denoted by $\left\{P^\theta_{U_i|U^{i-1},Y^N}\left(1|u^{i-1},y^N\right)\right\}_{i\in[N]}$. Finally, θ is optimized by the negative-log-loss (NLL) via stochastic gradient descent (SGD), as given by:

$$\min_{\theta\in\Theta} -\frac{1}{M}\sum_{i=1}^{M}\log P^\theta_{U_i|U^{i-1},Y^N}\left(u_i|u^{i-1},y^N\right). \tag{4}$$

In [5, Algorithm 2], the authors showed a recursive formula for the computation of the NLL. Let

$$\mathsf{L} = \mathsf{NSCTrain}\left(e^N, u^N\right), \tag{5}$$

where $e_i = E_\theta^W\left(y_i\right)$, denote the computation of the NSC loss. Also, the authors of [5] showed that the NSC decoder is a consistent estimator of the theoretical polar decoder whenever W is a FSC.

III. Data-Driven Polar codes for Asymmetric Sources

This section describes how to extend the data-driven polar decoder in Section II-B to the case where the input distribution is not necessarily symmetric. Specifically, it starts with a brief description of the Honda-Yamamoto scheme [7]. Then, it extends the data-driven polar decoder to accommodate asymmetric input distributions by incorporating this scheme in Section III-B.

A. Honda-Yamamoto Scheme for Asymmetric Channels

The Honda-Yamamoto scheme [7] generalizes polar coding for asymmetric input distributions. Here, the polar decoder is applied twice: first, before observing the channel outputs and second, after observing the channel outputs. An equivalent interpretation is that the first application of SC decoding is done on a different channel whose outputs are independent of its inputs. Indeed, in this case, as given in (1), the first term of the RHS cancels out, and it follows that the channel embedding are constant for all $y \in \mathcal{Y}$. Thus, for the first application of SC decoding, we denote the constant *input embedding* by E^X (rather than E^W). The second application of SC decoding follows the same procedure as in the case of symmetric channels.

Accordingly, a polar decoder with non symmetric input distribution is defined by the tuple $\left(\mathcal{X}, \mathcal{Y}, W, E^X, E^W, F, G, H\right)$. Here, we add the input embedding E^X to the definition, where $E_X(y)$ is constant for all $y \in \mathcal{Y}$. An important observation is that the functions F, G, H are independent of the channel, i.e. both applications of SC decoding (before and after observing the channel outputs) share the same functions F, G, H.

B. Honda-Yamamoto Scheme for Data-Driven Polar Decoders

This section considers two issues. The first is the choice of an input distribution. This is addressed by employing algorithms for capacity estimation [9], [10]. The second issue addresses the construction of a NSC decoder that is tailored for stationary input distributions.

Algorithm 1 Data-driven polar code design for channels with memory and non-i.i.d. input distribution

input: Input distribution P_{X^N}, Channel $W_{Y^N\|X^N}$, block length n_{t}, #of info. bits k
output: Optimized $e_\theta^X, E_\theta^W, G_\theta, F_\theta, H_\theta$

```
Initiate the weights of $e_\theta^X, E_\theta^W, G_\theta, F_\theta, H_\theta$
$N = 2^{n_{\mathrm{t}}}$
for $k = 1$ to $\mathsf{N}_{\mathsf{iters}}$ do
    Sample $x^N, y^N \sim P_{X^N} \otimes W_{Y^N\|X^N}$
    $u^N = x^N G_N$
    Duplicate $e_\theta^X$ to $e_X^N$
    Compute $e_Y^N$ by $e_i = E_\theta^W\left(y_i\right)$
    Compute $\mathsf{L}_X$ by applying $\mathsf{NSCTrain}\left(e_X^N, u^N, 0\right)$
    Compute $\mathsf{L}_Y$ by applying $\mathsf{NSCTrain}\left(e_Y^N, u^N, 0\right)$
    Minimize $\mathsf{L}_X + \mathsf{L}_Y$ w.r.t. $\theta$.
end for
return Optimized $\theta$
```

For the choice of the input distribution, we employ a recent method for the optimization of the directed information neural estimator (DINE), as presented in [10]. Therein, the authors provide an reinforcement learning (RL) algorithm that uses DINE to estimate capacity achieving input distributions. The input distribution is approximated with an recurrent neural network (RNN) with parameter space denoted by Π. Let P_X^π be the estimated capacity achieving input distribution. Thus, by application of [10, Algorithm 1], we obtain a model of P_X^π from which we are able to sample the channel inputs.

Extension of the NSC decoder to P_{X^N} (that is not uniform and i.i.d.) involves introducing additional parameters, that we denote by $\theta_5 \in \Theta$. Accordingly, we denote the set of the channel embedding by θ_1, θ_5, where θ_5 denotes the parameters of E^X and θ_1 are the parameters of E^W. We define $E_\theta^X : \mathcal{Y} \to \mathbb{R}^d$ as a constant RV that satisfies $E_\theta^X\left(y\right) = e_X \in \mathbb{R}^d$ for all $y \in \mathcal{Y}$. Accordingly, the NSC in this case is defined by $E_\theta^X, E_\theta^W, F_\theta, G_\theta, H_\theta$. Thus, the NSC decoder needs to be updated in order to optimize E_θ^X as well. This is addressed by first applying the NSC with inputs e_X^N to compute $P^\theta_{U_i|U^{i-1}}$, where $e_X^N \in \mathbb{R}^{d\times N}$ is a matrix whose columns are duplicates of e_X. Second, the NSC is applied with e_Y^N to compute $P^\theta_{U_i|U^{i-1},Y^N}$, where $e_Y^N \in \mathbb{R}^{d\times N}$ is a matrix whose i-th column is $E_\theta^W\left(y_i\right)$.

The training procedure admits the following steps. First, the channel inputs and outputs are sampled by $x^N, y^N \sim P_{X^N}^\pi \otimes W_{Y^N\|X^N}$. Then, the values of $u^N = x^N G_N$ are computed, and form the labels of the algorithm. Next, the channel statistics e_Y^N are computed and the input statistics are duplicated to obtain e_X^N. The next step is to apply the NSC-Train procedure twice, i.e.

$$\mathsf{L}_X = \mathsf{NSCTrain}(e_X^N, u^N) \tag{6}$$

$$\mathsf{L}_Y = \mathsf{NSCTrain}(e_Y^N, u^N), \tag{7}$$

which are minimized via SGD, as shown in Algorithm 1.

IV. List Decoding of Data-Driven Polar Codes

In this section, we delve into the concept of list decoding for polar codes and discuss its integration into our data-driven polar codes. To this end, the NSC is benchmarked against two ground truth decoding methods: the SC decoder and the SCT decoder, depending on the presence or absence of channel memory. Notably, contemporary algorithms predominantly utilize the list decoding technique, known for its improved performance compared to the conventional SC algorithm. Consequently, to enable the NSC decoder to compete with state-of-the-art algorithms, this section incorporates list decoding with the NSC decoder.

A. SC List Decoder

To enhance the error correction performance of polar codes, especially with codes of moderate lengths, the SCL decoding algorithm was introduced in [6]. The fundamental concept behind list decoding lies in leveraging the structured nature of the polar transformation. Instead of relying solely on a single SC decoder, the SCL decoder concurrently decodes multiple codeword candidates. This is achieved by applying multiple SC decoders over the same channel's outputs, with the number of these decoders denoted as the list size L.

The SCL decoder generates a list of potential codewords, each ranked by its likelihood of being the transmitted message. Subsequently, this list undergoes a refining process to identify the most likely original message. To achieve this, the SCL algorithm estimates each bit's value (0 or 1) while considering both possibilities. At each estimation step, the number of codeword candidates (also referred to as paths) doubles. To manage the algorithm's complexity, it employs a memory-saving strategy by retaining only a limited set of L codeword candidates at any given time. Consequently, after each estimation, half of the paths are discarded. To determine which paths to retain, a path metric (PM) is associated with each path. This metric is continuously updated with each new estimation and is computed via the log-likelihood ratios (LLRs). The algorithm maintains the L paths with the lowest path metrics, allowing them to persist and continue the decoding process.

B. NSC List Decoder

Here we highlight that the concept of list decoding can be integrated into our data-driven polar codes. Recall that the NSC decoder uses the same structure as the SC decoder and the SCT decoder, with the only distinction being the replacement of elementary operations with NN. Accordingly, we can seamlessly incorporate the list decoding concept into the NSC decoder. Specifically, since the NSC decoding algorithm can estimate the LLRs at the decision points, we can leverage them to compute the PM and follow the same SCL decoding procedure.

C. Computational Complexity

The standard SC algorithm has a computational complexity of $O\left(N\log(N)\right)$, whereas the SCT algorithm's computational complexity is $O\left(|\mathcal{S}|^3N\log(N)\right)$. In the context of list decoding, a technique based on leveraging the memory sharing structure among the candidate paths was introduced in [6]. This technique demonstrates that the SCL decoder can be implemented with a computational complexity of $O\left(LN\log(N)\right)$. When applying the same technique to the SCT algorithm with list decoding, it follows directly that the computational complexity increases to $O\left(L|\mathcal{S}|^3N\log(N)\right)$.

The following theorem examines the computational complexity of the NSC list decoder for the case where $E_\theta, F_\theta, G_\theta, H_\theta$ are NNs with k hidden units and the embedding space satisfies $\mathcal{E}\subset\mathbb{R}^d$. Due to space limitation, the proof is omitted.

Theorem 1. *Let $E_\theta, F_\theta, G_\theta, H_\theta$ be NNs with k hidden units and let $\mathcal{E}\subset\mathbb{R}^d$. Then, the computational complexity of NSC list decoding is $O\left(LkdN\log_2 N\right)$.*

The main purpose of Theorem 1 is to facilitate a comparison between the NSC list decoder and SCT list decoder. Note that the computational complexity of the SCT list decoder, as previously mentioned, scales with the memory size of the channel $O\left(L|\mathcal{S}|^3N\log N\right)$. This highlights a key advantage of the NSC list decoder since its computational complexity remains independent of the channel's memory size.

V. Experiments

This section presents experiments designed to evaluate the performance of our proposed algorithms. It begins with asymmetric channels in Section V-A and continues with list decoding in Section V-B. In all experiments, the NNs, $F_\theta, G_\theta, H_\theta, E_\theta^X, E_\theta^W$, are implemented by two layered fully-connected NNs with 50 hidden units per layer.

A. Asymmetric Channels

In this section, we conduct experiments to evaluate our methodology for designing polar codes tailored to asymmetric channels. As an example of a memoryless channel, we consider the non-symmetric BEC, as defined in [11]. This channel is defined by two erasures probabilities, ϵ_0, ϵ_1,

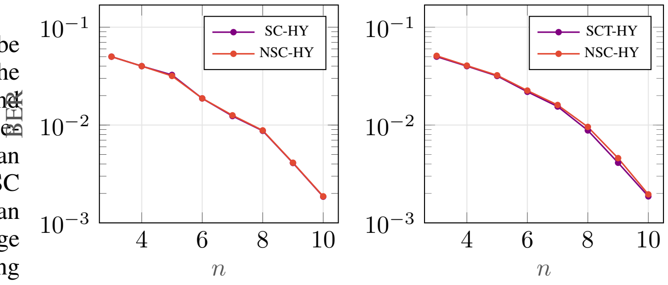


Figure 1: These figures compare the bit error rates (BERs) attained by the Honda-Yamamoto scheme (SC-HY) and its extension to the NSC decoder (NSC-HY). The left and right figures show the results on an asymmetric binary errasure channel (BEC), and the Ising channel, respectively.

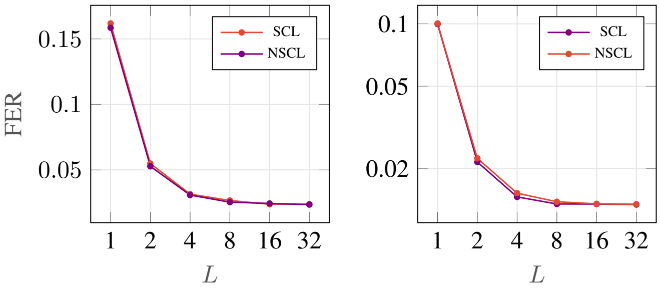


Figure 2: These figures compare the FERs attained by the the SCL decoder and its extension the NSC decoder (NSCL). The left figure shows the results on an binary-input AWGN channel with signal-to-noise ratio (SNR) of 1.5, and the right figure shows the results on the Ising channel.

namely the probabilities for an erasure of the “0” symbol and the “1” symbol, respectively. Accordingly, $W(x|x) = 1 - \epsilon_x$, $W(?|x) = \epsilon_x$ for $x \in \{0, 1\}$. Similar to [11], we choose $\epsilon_0 = 0.4$ and $\epsilon_1 = 0.8159$.

As an instance of a channel with memory, we choose the Ising channel that was introduced in [12], which belongs to the family of FSCs, and therefore, its optimal decoding rule is given by the SCT decoder. This channel is defined by $Y = X$ or $Y = S$ with equal probability, and $S' = X$, where X is the channel input, Y is the channel output, S is the channel states at the beginning of the transmission and S' is the channel's state at the end of the transmission.

Figure 1 compares the BERs obtained via the extension of the Honda-Yamamoto scheme, as described in Section III, and by the optimal decoding rule of the Honda-Yamamoto scheme. The left figure compares the result on the asymmetric BEC, a memoryless channel, and the right figure compares the results on the Ising channel, a FSC.

B. List Decoding

In this Section, we demonstrate the performance of NSC list decoder compared to the SCL decoder. As an example of a memoryless channel, we consider the additive white Gaussian noise (AWGN) channel. The AWGN channel is defined by the following relation $Y = X + N$, where X is the channel input, Y is the channel output, and $N \sim \mathcal{N}(0, \sigma^2)$ is an i.i.d. Gaussian noise. In our experiments $\sigma^2 = 1.5$. Figure 2 illustrates the frame error rates (FERs) obtained via the SCL decoder with the FERs obtained via the NSC list decoder as a function of the list size L. The left figure demonstrates the results for the AWGN channel while the right figure compares the results for the Ising channel. As can be seen in the figures, the NSC list decoder indeed converges to the ground truth SCL decoder for both channels.

VI. Conclusions

This paper presents pivotal extensions to data-driven polar decoder, addressing two critical applications: list decoding and adaptation to asymmetric input distributions. These enhancements are essential steps towards realizing data-driven codes that achieve channel capacity and excel at moderate block lengths. By seamlessly integrating list decoding, we effectively mitigate error propagation issues inherent to SC decoding, improving the practical performance of polar codes. Simultaneously, our incorporation of the Honda-Yamamoto scheme enables these codes to adapt to non-uniform input distributions in a computationally efficient manner, without the need for explicit channel model. Our numerical results validate the effectiveness of these contributions, establishing data-driven polar codes as robust and versatile coding solutions adaptable to diverse channel conditions.

References

[1] E. Arikan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[2] E. Şaşoğlu and I. Tal, “Polar coding for processes with memory,” *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 1994–2003, 2019.

[3] R. Wang, R. Liu, and Y. Hou, “Joint successive cancellation decoding of polar codes over intersymbol interference channels,” *arXiv preprint arXiv:1404.3001*, 2014.

[4] R. Wang, J. Honda, H. Yamamoto, R. Liu, and Y. Hou, “Construction of polar codes for channels with memory,” in *2015 IEEE Information Theory Workshop-Fall (ITW)*, IEEE, 2015, pp. 187–191.

[5] Z. Aharoni, B. Huleihel, H. D. Pfister, and H. H. Permuter, “Data-driven polar codes for unknown channels with and without memory,” submitted to IEEE Int. Symp. Inf. Theory (ISIT), 2023.

[6] I. Tal and A. Vardy, “List decoding of polar codes,” *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2213–2226, 2015.

[7] J. Honda and H. Yamamoto, “Polar coding without alphabet extension for asymmetric models,” *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7829–7838, 2013.

[8] G. Kramer, *Directed Information for Channels with Feedback*. 1998, vol. 11.

[9] D. Tsur, Z. Aharoni, Z. Goldfeld, and H. H. Permuter, “Neural estimation and optimization of directed information over continuous spaces,” *submitted to IEEE Trans. Inf. Theory*,

[10] D. Tsur and Z. Aharoni and Z. Goldfeld and H. H. Permuter, “Optimizing estimated directed information over discrete alphabets,” in *2022 IEEE Int. Symp. Inf. Theory (ISIT)*, IEEE, 2022, pp. 2898–2903.

[11] J. Honda and H. Yamamoto, “Polar coding without alphabet extension for asymmetric channels,” in *2012 IEEE International Symposium on Information Theory Proceedings*, IEEE, 2012, pp. 2147–2151.

[12] T. Berger and F. Bonomi, “Capacity and zero-error capacity of Ising channels,” *IEEE Trans. Inf. Theory*, vol. 36, pp. 173–180, 1990.

M-DAB: An Input-Distribution Optimization Algorithm for Composite DNA Storage by the Multinomial Channel

Adir Kobovich, Eitan Yaakobi and Nir Weinberger
Technion – Israel Institute of Technology, Haifa, Israel
Email: adir.k@campus.technion.ac.il, yaakobi@cs.technion.ac.il, nirwein@technion.ac.il

***Abstract*—Recent experiments have shown that the capacity of DNA storage systems may be significantly increased by synthesizing composite DNA letters. In this work, we model a DNA storage channel with composite inputs as a *multinomial channel*, and propose an optimization algorithm for its capacity achieving input distribution, for an arbitrary number of output reads. The algorithm is termed multidimensional dynamic assignment Blahut-Arimoto (M-DAB), and is a generalized version of the DAB algorithm, proposed by Wesel et al. [1] developed for the binomial channel. We also empirically observe a scaling law behavior of the capacity as a function of the support size of the capacity-achieving input distribution.**

I. Introduction

DNA storage [2], [3] is an emerging technology that involves converting digital information into nucleotide sequences with quaternary encoding, represented by the letters A, C, G, and T. The sequence, known as a *strand*, is written via *synthesis* and is read via *sequencing*. An intriguing aspect of this process is the generation of multiple copies of the same strand during the synthesis process. In this paper, we focus on one approach to harnessing this redundancy, achieved by introducing the concept of *composite DNA letters* [4]–[8]. These composite letters consist of different mixtures of nucleotides, and have been successfully utilized in data encoding experiments [4], [5], [8]. In theory, using composite DNA letters can dramatically increase the capacity of the DNA storage channel, since while the capacity of simple 4-letter DNA encoding is bounded by $\log(4) = 2$ bits per channel use, the capacity of composite DNA encoding is unbounded. Furthermore, the larger capacity enables encoding data in shorter strands, which is particularly effective in DNA storage, due to the high cost of the synthesis process [5], and the nature of the process, in which the error probability increases as the strand gets longer [9].

The process of writing a composite letter and randomly reading n copies can be modeled as the operation of a noisy channel. In this channel, the input is a probability vector of length $k = 4$ (in the case of DNA letters), which represents a mixture of nucleotides. The channel output is distributed as a multinomial random variable, with n event trials and event probabilities given by the input vector. The input represents the expected frequency of occurrences of each of the four nucleotides in the n output copies. We thus refer to this channel as the *multinomial channel*. The maximal storage rate of information possible over this channel is the *capacity* of the channel. As is well-known [10], the capacity is obtained by maximizing the mutual information between the input and output of the channel, over all feasible choices of input distributions, that is, distributions over the $(k-1)$-dimensional probability simplex.

In this paper, we consider the problem of determining the capacity-achieving input distribution (CAID) when the number of reads n is finite. Since the output alphabet of this channel is discrete, there exists a CAID with support of finite cardinality. Thus, the CAID can be parameterized by a finite set of points in the $(k-1)$-dimensional simplex (*locations*) and their respective probabilities (*weights*). If one then further fixes the locations, then the channel is reduced to a discrete memoryless channel (DMC), and the optimal probabilities can be computed by the Blahut-Arimoto algorithm [11]. Consequently, the main computational challenge is to identify the optimal locations in the $(k-1)$-dimensional simplex.

Previous works have addressed this challenge for the special case of $k = 2$, known as the *binomial channel* [12], and for a variation of it called the *particle-intensity channel* [13]. These works introduce an algorithm for finding the CAID, termed the *Dynamic Assignment Blahut-Arimoto* (DAB) algorithm [1]. In this paper, we propose a generalization of the DAB algorithm, which finds the CAID for the multinomial channel ($k > 2$). We focus on the challenges associated with the multidimensionality of the channel symbols, and refer our algorithm as the *multidimensional dynamic assignment Blahut-Arimoto* (M-DAB). Using M-DAB, we compute the CAID for various values of n. These CAIDs can directly be used in coded DNA storage systems to obtain improved coding rates. In addition, we evaluate the cardinality of the support of the CAID as a function of the mutual information. We show that this cardinality matches the scaling law behavior, recently conjectured in [14].

This paper is organized as follows. Section II defines the multinomial channel and the optimization problem. Section III studies properties of the multinomial channel CAID, which be later used in our algorithm. Section IV introduces the DAB algorithm as preliminary work and presents the M-DAB algorithm including the challenges associated with multidimensional inputs. Section V shows the achieved CAID and corresponding channel capacities. Lastly, in Section VI we discuss open problems. Due to limited space, full proofs can be found in the extended version [15].

II. Multinomial Channel Problem Definition

In this section, we formally define the multinomial channel and the resulting CAID optimization problem. The multinomial channel *input* alphabet is the $(k-1)$-dimensional probability simplex, given by $\Delta_k := \{x \in \mathbb{R}_+^k \mid \sum_{i=1}^{k} x_i = 1\}$. The multinomial channel n*-output* alphabet is the set of all multisets with cardinality n over $[k]$, given by $\mathcal{Y}_{n,k} := \{y \in \mathbb{Z}_+^k \mid \sum_{i=1}^{k} y_i = n\}$, whose cardinality satisfies $|\mathcal{Y}_{n,k}| = \binom{n+k-1}{k-1}$. For a given input $x \in \Delta_k$, the output of the *multinomial* channel is given by $Y \sim \text{Multinomial}(n,x)$, that is, given input $x \in \Delta_k$ the output $y \in \mathcal{Y}_{n,k}$ obeys the following transition probability

$$P_{Y|X}^{(n,k)}(y|x) = \frac{n!}{\prod_{j=1}^{k} y_j!} \prod_{j=1}^{k} x_j^{y_j}.$$

The channel is referred to as *binomial* if $k = 2$. Hence, if the input is a composite letter $x \in \Delta_k$, then the expected number of times that the i-th alphabet letter appears in the output strand is nx_i. We remark that this channel only models the randomness in the output due to sampling of the input, but does not model additional noise in the reading process. If the reading process can be modeled as a symmetric DMC, with a total flip probability of ϵ (thus $\frac{\epsilon}{k-1}$ to each of the other $k-1$ letters), then the result is simply a $\text{Multinomial}(n, x * \epsilon)$ channel, where

$$(x*\epsilon)_i := x_i(1-\epsilon) + \epsilon(1-x_i) \quad \text{for all} \quad i \in [k].$$

It is thus straightforward to extend our algorithm and results to this case too. For convenience, we henceforth consider the noiseless channel.

Our main objective is to find the CAID of the multinomial channel, i.e., to find the input distribution that maximizes the mutual information. Let $\mathcal{F}_k$ be the set of all input distributions supported on the input alphabet Δ_k. Specifically, we aim to solve the following optimization problem of the CAID of the multinomial channel:

$$C_{n,k} := \max_{f_X \in \mathcal{F}_k} I(X;Y). \tag{1}$$

Alternatively, the dual problem to (1), also known as the Csiszár minimax capacity theorem [16], is given by

$$C_{n,k} = \min_{P_Y} \max_{x \in \Delta_k} D(P_{Y|X=x} || P_Y), \tag{2}$$

where $D(\cdot||\cdot)$ denotes the Kullback–Leibler (KL) divergence, and P_Y is a distribution over $\mathcal{Y}_{n,k}$.

III. Properties of the Capacity-Achieving Input Distribution

For any finite n, there is no analytical solution for the capacity and CAID of the multinomial channel. Thus, we next derive a few properties of this CAID, which will be useful later on, e.g., in reducing the size of the optimization input space. Interestingly, the same capacity arises in universal coding [17], where it has been demonstrated that the CAID is asymptotically proportional to Jeffrey's prior [18], and the following asymptotic expression holds [19]

$$\lim_{n\to\infty} \left(C_{n,k} - \frac{k-1}{2} \log\left(\frac{n}{2\pi e}\right)\right) = \log\left(\frac{\Gamma^k(1/2)}{\Gamma(k/2)}\right), \tag{3}$$

where $\Gamma(z) := \int_0^\infty e^{-t} t^{z-1} dt$ is the Gamma function.

The first property shows that the CAID can be atomic with finite support. We modify the result of [20], which was proven using Dubins' theorem [21]:

Lemma 1. Consider a channel, with an input X taking values in Δ_k for some $k > 1$, and a discrete finite output alphabet Y. Assume the transition probability distribution function $x \to P_{Y|X}(y|x)$ is continuous for each $y \in Y$. Then, there exists a CAID supported on less than $|Y|$ points in Δ_k.

Corollary 1. There is a CAID with finite support size $m \leq |\mathcal{Y}(k,n)|$. The corresponding input distribution is given by

$$f_X^*(x) = \sum_{i=1}^{m} p_i^* \delta(x - x^{(i)}), \tag{4}$$

and $\delta(x)$ is the Dirac delta function. That is, $f_X^*(x)$ is an atomic distribution.

The second property pertains to the set of input symbols at the vertices of the simplex. In the context of DNA encoding, the vertices symbols are the non-composite letters.

Lemma 2. $C_{n,1}$ is achieved by a uniform distribution on the k vertices of Δ_k.

This property allows us to use the CAID for $n = 1$ as an initialization to our search. The third property pertains to the symmetry of the CAID with respect to (w.r.t.) the input alphabet. For the binomial channel ($k = 2$), a simple symmetry argument combined with the concavity of the mutual information w.r.t. the input distribution implies that the set of distributions satisfying $f_X((x, 1-x)) = f_X((1-x, x))$ includes a CAID. The generalized property for the multinomial channel is more involved, and is given as follows.

Definition 1. Let $\mathcal{S}_k$ be the set of all bijections from $[k]$ to itself (the symmetric group over $[k]$). An input distribution $f_X(x)$ is said to be invariant under input dimension permutation (IDP) if $f_X(x) = f_X(\pi(x))$ for any $x \in \Delta_k$ and any $\pi \in \mathcal{S}_k$.

Lemma 3. There exists a CAID that is invariant under IDP.

The above properties allow us to reduce the search space for a CAID from all possible distributions on Δ_k to the subset of input distributions that are finitely supported with at most m atoms and which are invariant under IDP. Thus, we will search a CAID over distributions supported on the $(k-1)$-dimensional ordered simplex $\Delta_k^{\geq} := \{x \in \Delta_k \mid x_0 \geq x_1 \geq \cdots \geq x_{k-1} \geq 0\}$, such that the input distribution f_X corresponding to a distribution $\tilde{f}_X$ supported on $\Delta_k^{\geq}$ is given by

$$f_X = \frac{1}{k!} \sum_{\pi \in \mathcal{S}_k} \tilde{f}_{\pi(X)}. \tag{5}$$

IV. Multidimensional Dynamic Assignment Blahut-Arimoto

Our proposed algorithm M-DAB generalizes the DAB algorithm, which finds the CAID for the multinomial channel, rather than for the binomial channel. The main novelty of M-DAB is the handling of the multiple dimensions of the multinomial channel. To present M-DAB, we first briefly review algorithms for the binomial channel [12], and specifically the DAB algorithm.

A. Preliminaries: Algorithms for the Binomial Channel

In [12], it was suggested to solve the dual problem (1) using the ellipsoid method [22]. However, [1] reported that this method converges slowly, even with well-chosen initial conditions. The DAB algorithm was developed to overcome this. Subsequently, [13] further improved and utilized it to find a CAID of the particle-intensity channel.

The DAB algorithm can be thought of as a primal-dual alternating optimization algorithm, in which the weights and locations of the CAID (4) are alternatively updated via the primal and dual problems. Specifically, the dual problem is used to update the locations, and the main idea is that given P_Y, adding a point in the location of the maximizer $x_{\max} \in \Delta_k$ tends to reduce the dual objective in (2). The DAB algorithm for the CAID of $C_{n,k=2}$ is as follows:

0) Initialize the locations of f_X as the locations of the CAID of $C_{n-1,k=2}$.
1) Run the Blahut-Arimoto algorithm on the current locations to optimize the weights, obtain an input distribution f_X, and compute the value of the primal objective (1). Also compute the output distribution P_Y.
2) Using P_Y, find the maximizer $x_{\max} \in \Delta_k$ of the KL divergence in (2).
3) Find the nearest mass point x_{closest} in f_X to $x_{\max}$.
4) Determine whether to add a point to the support of f_X.
5) Move x_{closest} in the direction of $x_{\max}$ and compute the value of the dual objective function (2).
6) Stop if the primal and dual are ϵ-close. Otherwise Jump to step 1.

B. Overview of the M-DAB Algorithm

The M-DAB algorithm (Algorithm 1) is based on the steps of the DAB algorithm. The CAID optimization for $C_{n,k}$ is initialized with the mass points from the CAID of $C_{n-1,k}$, and for $C_{1,k}$ the vertices of the simplex are used, as suggested by Lemma 2. An important operation of M-DAB is *ReduceToOrderedSimplex* (row 1), which reduces the search space to the ordered simplex by simply removing all the mass points outside the ordered simplex. Whenever a computation requires the full input distribution (rows 3, 4, 5 and 15), we use the expansion operation *Expand* which creates the full simplex from the ordered simplex by inserting all the permutations as suggested in (5).

At each iteration we use the Blahut-Arimoto algorithm (row 3) over the current locations, in order to get the corresponding weights, which together with the locations uniquely determine the input distribution f_X. The input distribution, together with the channel transition probabilities, are then used to calculate the mutual information (row 4) and the output distribution P_Y. Next, we search for the symbol $x_{\max}$, which maximizes the KL divergence between its output distribution and P_Y (row 5). The maximum divergence D is the value of the dual function, and thus it is an upper bound for the capacity. The algorithm will continue until the mutual information is ϵ close to the upper bound (row 6).

Algorithm 1 M-DAB

```
Input: $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(m)}), n, \epsilon$.
Output: $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(m')}), \mathbf{p} = (p_1, p_2, \ldots, p_m)$ st.
    $I(\mathbf{x}, \mathbf{p}) \geq C - \epsilon$.
1:  $\mathbf{x} \leftarrow ReduceToOrderedSimplex(\mathbf{x})$
2:  while True do
3:      $\mathbf{p} \leftarrow BlahutArimoto(Expand(\mathbf{x}))$
4:      $I \leftarrow I(Expand(\mathbf{x}), \mathbf{p})$
5:      $D, x_{\max} \leftarrow \max_{x \in \Delta_k^{\geq}} D(P_{Y|X=x} || P_Y(Expand(\mathbf{x}), \mathbf{p}))$
6:      if $D - I \leq \epsilon$ then
7:          return $Expand(\mathbf{x}), \mathbf{p}$.
8:      end if
9:      $d_x, x_{\text{closest}} \leftarrow \min_x D(x || x_{\max})$
10:     $d_v, v_{\text{closest}} \leftarrow \min_{v \in Vertices} D(v || x_{\max})$
11:     if $d_x > d_v$ then
12:         $\mathbf{x} \leftarrow AddPoint(\mathbf{x}, v_{\text{closest}})$
13:     else
14:         $\mathbf{g} \leftarrow CreateDirectionVector(\mathbf{x}, x_{\text{closest}}, x_{\max})$
15:         $\delta_{\max} \leftarrow \max_\delta I\left(Expand(\mathbf{x} + \delta\mathbf{g})\right)$
16:         $\mathbf{x} \leftarrow \mathbf{x} + \delta_{\max}\mathbf{g}$
17:     end if
18: end while
```

The next steps involve adjusting our guess based on $x_{\max}$, by checking which mass point is closest to $x_{\max}$ using the KL divergence as a distance measure (row 9). In case that one of the vertices of the ordered simplex is closer (row 10), this point is added (row 12), using the *AddPoint* operation, which simply inserts the point to $\mathbf{x}$. Otherwise, we need to adjust the closest mass point $x_{closest}$. For simplicity, we represent the set $\mathbf{x}$ as a vector. The operation *CreateDirectionVector* (row 14) is used to create the direction vector g, which is $\mathbf{0}$ for any x that is not $x_{closest}$, and there it equals $x_{\max} - x_{closest}$. Then, the algorithm uses a line search (row 15) to find the optimal adjustment step δ, and uses it to adjust the locations (row 16).

C. A Generalization to Higher Input Dimensions

We next emphasize how the multidimensionality $k > 2$ affects the steps made in the DAB algorithm, and what modifications were needed.

Initialization and Step 0: We first harness the symmetry assured by Lemma 3 to reduce the search space. In the binomial channel case, the DAB algorithm utilizes the apparent symmetry between inputs $x \in [0, 1]$ and $1 - x \in [0, 1]$ and adjusts the locations in pairs. M-DAB further expands this symmetry, to make use of all the permutations, and reduce the search space from Δ_k to $\Delta_k^{\geq}$ (row 1), thus reducing the number of the optimized location parameters. The distribution supported on $\Delta_k^{\geq}$ is symmetrically expanded to an input distribution on Δ_k as in (5), and the latter is used to compute the mutual information, for running the Blahut-Arimoto algorithm, for computing the output distribution, and for finding the maximizer of the KL divergence (rows 3, 4 and 5). This process is equivalent to changing the initialization of Step 0 to use only the CAID mass points supported on $\Delta_k^{\geq}$. We also mention that the dependence of each run of the algorithm on an initialization based on previous runs causes the error to accumulate. Step 6 includes a threshold, ϵ, such that the algorithm continues until the mutual information is ϵ-close to the upper bound (row 6). This threshold affects whether the initialization will be sufficient or not for larger values of n.

Step 1: This step is similar to the DAB algorithm.

Step 2: This step is challenging in M-DAB since it requires a more complicated multidimensional maximization, compared to a simple line search for $k = 2$. Moreover, the bounds of the maximization problem, i.e., the edges of the ordered simplex, tend to behave as a small scale of the original problem. In order to address these challenges we use known the simplicial homology global optimization (SHGO) algorithm [23], with sampling using the Sobol sequence [24].

Step 3: The DAB algorithm finds the closest mass point to the global maximum, and this point is later used in Step 4 and Step 5. This method works well for the binomial case, but in the multinomial case, we experimentally observed iterations in which adjusting the closest point in the direction of the maximum does not increase the value of the mutual information. Moreover, DAB does not adjust the vertices $\{0, 1\}$, even if they are the closest ones [1]. Later, [13] suggested to constrain the search of the closest points to the interval bounded by $x_{\max}$ and 0.5. We thus conclude that the Euclidean distance is not necessarily the most efficient distance measure, and modify the M-DAB algorithm to use the KL divergence as the distance measure (row 9). The KL divergence is empirically better, and appears to be a natural measure in this scenario, as we would like to find which input is more likely to be interpreted as the global maximum. Figure 1 illustrates such a case. During the first iteration of the algorithm for $C_{n=7,k=3}$ we find the maximum in $(0.616, 0.192, 0.192)$ while the nearest point using the Euclidean distance is $(0.682, 0.318, 0)$. This point is on the edge of the simplex, and trying to adjust it is not beneficial, so the algorithm will not converge.

Step 4: The M-DAB algorithm decides whether to add a mass point or not. For $k = 2$, DAB adds points either at $x = 0.5$, or by splitting this point. For the multidimensional case, a new mass point may be required in any of the vertices of $\Delta_k^{\geq}$ (rows 10, 11 and 12), whereas in the binomial case, the vertices $\{0, 1\}$ are always occupied. Moreover, M-DAB algorithm adds a new mass point to the input distribution implicitly whenever a point moves from one of the symmetry axes.

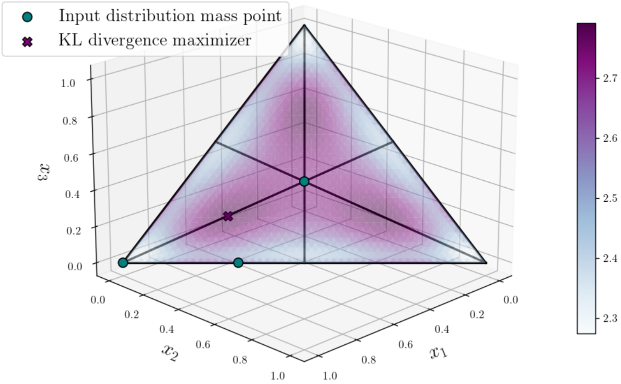


Figure 1: The first iteration of the M-DAB algorithm for $C_{n=7,k=3}$. The simplex is an equilateral triangle, and the ordered simplex is a right-angled triangle on the bottom-left. There are 3 mass points in the ordered simplex (blue 'o'). The color bar represents the KL divergence value in each point in the simplex, and the maximizer is marked with purple 'x'.

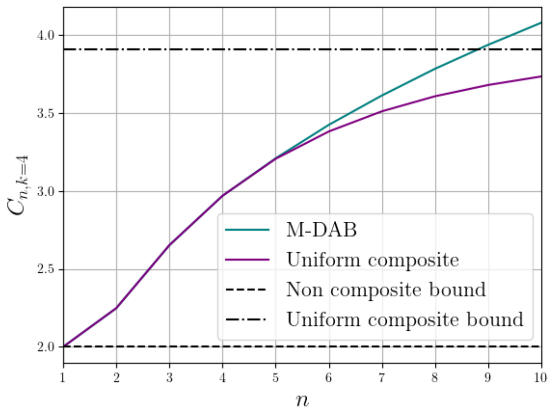


Figure 2: The capacity achieved for $C_{n,k=4}$. In blue, our method M-DAB, and in purple, the method used in [5]. In black are the upper bounds on any base 4 and 15 encoding.

V. Results

In this section, we present empirical results of the M-DAB algorithm. Specifically, for the case $k = 4$ suitable for our DNA storage motivation, the mass points locations and weights are available in the extended version [15]. These CAIDs can be used in future experiments and systems of composite DNA storage. In Figure 2, the capacity is achieved by running M-DAB for $k = 4$. We compare the mutual information achieved by M-DAB to the naive method of using only the *uniform composite* $(1, 0, 0, 0)$, $(0.5, 0.5, 0, 0)$, $(0.25, 0.25, 0.25, 0.25)$ and $(0.333, 0.333, 0.333, 0)$, proposed in [5]. Considering the critical number of copies where our result surpasses other methods, using composite letters is not beneficial only for $n = 1$. Thus, composite letters are strictly better than using ordinary DNA encoding whenever the output strands are redundant, i.e., $n \geq 2$. The uniform composite is

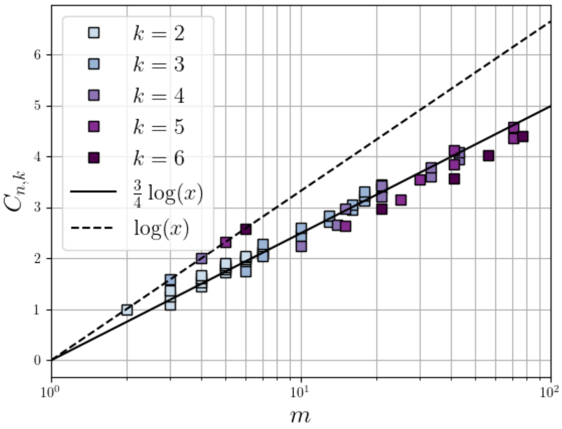


Figure 3: The capacity as function of the number of mass points in the minimal support size CAID. The dimensions are represented in different colors. A scaling law can be observed, where the capacity behaves as a logarithm with a factor of $\frac{3}{4}$.

a CAID for small values of copies, the CAID obtained by M-DAB is better for any $n \geq 5$. Furthermore for any $n \geq 9$ copies, the capacity of M-DAB surpasses the limit of any base 15 method ($\log(15) \sim 3.907$).

The capacity of the multinomial channel computed by the M-DAB algorithm for finite n allows to compare it with a scaling law recently found in [25] for the binomial and Gaussian channels, which was conjectured to be universal in a subsequent study [14]. The scaling law claims that the mutual information $I(X;Y)$ for CAID supported on m atoms scales as $\frac{3}{4}\log m$. Our method allows us to plot the capacity and support size for different values of n, as a parametric curve (Figure 3). Doing so we numerically validate the conjecture of [14], and testify that M-DAB finds the CAID with the minimum number of mass points.

VI. Future Research

We have seen that the dimensionality of the input complicates the design of the M-DAB algorithm compared to the DAB algorithm. While the experimental results demonstrate the effectiveness of M-DAB, theoretical convergence guarantees are lacking. This is challenging since the M-DAB algorithm optimizes the location of just one mass point at each iteration, and so can be viewed as a *coordinate descent* algorithm. The convergence analysis of such algorithms is not always obvious. We notice that even for small values of n (such as $n = 9$) the CAID requires several dozens of mass points, and as suggested by the scaling law, the support size scales exponentially as a function of the capacity. This might be impractical to implement in DNA storage systems, since any mass point requires a specific mixture of nucleotides, it is desirable to use a minimal number of such mixtures. This raises a natural and important follow-up problem, which is to determine the CAID of the multinomial channel under a constraint on the support size.

Acknowledgments

The research was Funded by the European Union (ERC, DNAStorage, 865630). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The research of N. W. was supported by the Israel Science Foundation (ISF), grant no. 1782/22.

References

[1] R. D. Wesel *et al.*, "Efficient binomial channel capacity computation with an application to molecular communication," in *ITA*. IEEE, 2018, pp. 1–5.

[2] G. M. Church *et al.*, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[3] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[4] L. Anavy *et al.*, "Data storage in DNA with fewer synthesis cycles using composite DNA letters," *Nat. Biotechnol.*, vol. 37, no. 10, pp. 1229–1236, 2019.

[5] Y. Choi *et al.*, "High information capacity DNA-based data storage with augmented encoding characters using degenerate bases," *Sci. Rep.*, vol. 9, no. 1, p. 6582, 2019.

[6] I. Preuss, Z. Yakhini, and L. Anavy, "Data storage based on combinatorial synthesis of DNA shortmers," *bioRxiv*, pp. 2021–08, 2021.

[7] W. Zhang *et al.*, "Limited-magnitude error correction for probability vectors in DNA storage," in *ICC*. IEEE, 2022, pp. 3460–3465.

[8] Y. Yan *et al.*, "Scaling logical density of DNA storage with enzymatically-ligated composite motifs," *bioRxiv*, 2023.

[9] J. Bornholt *et al.*, "Toward a DNA-based archival storage system," *Ieee Micro*, vol. 37, no. 3, pp. 98–104, 2017.

[10] T. M. Cover, *Elements of information theory*. JWS, 1999.

[11] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[12] C. Komninakis, L. Vandenberghe, and R. D. Wesel, "Capacity of the binomial channel, or minimax redundancy for memoryless sources," in *IEEE ISIT*, 2001, pp. 127–127.

[13] N. Farsad *et al.*, "Capacities and optimal input distributions for particle-intensity channels," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 6, no. 3, pp. 220–232, 2020.

[14] M. C. Abbott and B. B. Machta, "A scaling law from discrete to continuous solutions of channel capacity problems in the low-noise limit," *J. Stat. Phys.*, vol. 176, pp. 214–227, 2019.

[15] A. Kobovich, E. Yaakobi, and N. Weinberger, "M-DAB: An input-distribution optimization algorithm for composite DNA storage by the multinomial channel," *researchgate*, 2023.

[16] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. CUP, 2011.

[17] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783–795, 1973.

[18] B. S. Clarke *et al.*, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Stat. Plan. Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[19] Q. Xie *et al.*, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[20] H. Witsenhausen, "Some aspects of convexity useful in information theory," *IEEE Trans. Inf. Theory*, vol. 26, no. 3, pp. 265–271, 1980.

[21] L. E. Dubins, "On extreme points of convex sets," *J. Math. Anal. Appl.*, vol. 5, no. 2, pp. 237–244, 1962.

[22] S. P. Boyd and L. Vandenberghe, *Convex optimization*. CUP, 2004.

[23] S. C. Endres *et al.*, "A simplicial homology algorithm for lipschitz optimisation," *J. Glob. Optim.*, vol. 72, pp. 181–217, 2018.

[24] S. Joe and F. Y. Kuo, "Constructing sobol sequences with better two-dimensional projections," *SIAM J. Sci. Comput.*, vol. 30, no. 5, pp. 2635–2654, 2008.

[25] H. H. a. Mattingly, "Maximizing the information learned from finite data selects a simple model," *PNAS*, vol. 115, no. 8, pp. 1760–1765, 2018.

Mismatched Decoding: Generalized Mutual Information under Small Mismatch

Priyanka Patel
University of Cambridge
pp490@cantab.ac.uk

Francesc Molina
University of Cambridge
Universitat Politècnica de Catalunya
fm585@cam.ac.uk

Albert Guillén i Fàbregas
University of Cambridge
Universitat Pompeu Fabra
guillen@ieee.org

***Abstract*—This paper investigates achievable information rates in mismatched decoding when the channel is close to the decoding rule in terms of relative entropy. We derive an approximation of the worst-case generalized mutual information as a function of the radius of a small relative entropy ball centered at the decoding metric, allowing to characterize the loss incurred due to good, yet imperfect channel estimation.**

I. Introduction and Problem Setup

Mismatched decoding is the problem that studies reliable communication employing a fixed and possibly sub-optimal metric for decoding. Mismatched decoding encompasses a number of important problems such as channel uncertainty, bit-interleaved coded modulation, finite-precision arithmetic and zero-error communication [1]. The problem is described as follows. Consider reliable transmission of M messages over a discrete memoryless channel with input X and output Y, taking values from discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The input distribution is denoted by $Q_X(x) = \Pr[X = x]$ for all $x \in \mathcal{X}$ and the channel transition distribution is defined as $W(y|x) = \Pr[Y = y|X = x]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For transmission, the encoder transmits the n-symbol codeword $\boldsymbol{x}^{(m)} = (x_1^{(m)}, \dots, x_n^{(m)})$ corresponding to message $m = 1, \dots, M$ from the codebook $\mathcal{C}_n = \{\boldsymbol{x}^{(i)}\}_{1\le i\le M}$. The decoder receives $\boldsymbol{y}$ and estimates the transmitted message as

$$\widehat{m} = \operatorname*{argmax}_{1\le \bar{m}\le M} \prod_{j=1}^{n} q\big(x_j^{(\bar{m})}, y_j\big). \tag{1}$$

When $q(x, y) = W(y|x)$, the decoder is said to be matched and coincides with maximum-likelihood decoding; in any other case, the decoder is referred to as mismatched. An error is declared when $\widehat{m} \neq m$, and the probability of error for $\mathcal{C}_n$ is defined as $p_e(\mathcal{C}_n) = \Pr[\widehat{m} \neq m]$.

A number of achievable rates for mismatched decoding have been derived in the literature [1]. When standard i.i.d. random coding is employed, the corresponding rate is the generalized mutual information (GMI) [2] given by

$$I_{\text{GMI}}(Q_X) = \sup_{s\ge 0}\ \mathbb{E}_{Q_X\times W}\left[\log \frac{q(X,Y)^s}{\mathbb{E}_{Q_X}[q(\bar{X},Y)^s|Y]}\right]. \tag{2}$$

The GMI is known to be tight with respect to the ensemble of i.i.d. codes [1]. In general, we have that $I_{\text{GMI}}(Q_X) \le C_{\text{M}}$, where C_{M} is the mismatch capacity. Although the GMI is an achievable rate for arbitrary decoding metrics $q(x, y)$, we consider the case where the decoder metric is a channel estimate $\widehat{W}(y|x)$ corresponding to the output of a channel estimator. We analyze the GMI for a mismatched decoder that uses the channel estimate $\widehat{W}(y|x)$ as if it were perfect. We impose a constraint on the level of mismatch between estimated and true channels by defining an appropriate distance measure, and find the worst-case achievable rate for small mismatch. Similarly to [3], for small mismatch between the channel estimate $\widehat{W}$ and the true channel W we require that

$$W \in \mathcal{B}(Q_X, \widehat{W}, r) = \big\{W \colon \mathrm{D}(\widehat{W}\|W|Q_X) \le r\big\}, \tag{3}$$

where $\mathcal{B}(Q_X, \widehat{W}, r)$ is a relative entropy ball centered at $\widehat{W}$ of radius r, assumed to be small. This definition adopts a decoder-centric perspective in which the ball is centered around the known quantity, i.e., the channel estimate employed to decode.

One of the advantages of this formulation for sufficiently small r is that we can resort to [4, eq. (1)–(4)] to express the relative entropy as function of $\theta(y|x) \triangleq W(y|x) - \widehat{W}(y|x)$ minus a non-negative term of minor relevance, as

$$\mathrm{D}(\widehat{W}\|W|Q_X) = \frac{1}{2}\sum_{x,y} Q_X(x)\frac{\theta^2(y|x)}{\widehat{W}(y|x)} - o\bigg(\sum_{x,y} Q_X(x)\frac{\theta^2(y|x)}{\widehat{W}(y|x)}\bigg). \tag{4}$$

Without loss of generality, we adopt throughout the paper natural logarithms and information units in nats.

II. Worst-Case GMI

In this section, we derive the worst-case GMI for small mismatch. We begin by defining the mismatched information density as

$$i_s(x,y) = \log \frac{\widehat{W}(y|x)^s}{\mathbb{E}_{Q_X}[\widehat{W}(y|X)^s]}, \tag{5}$$

where $s \ge 0$, for which the GMI can therefore be written as

$$I_{\text{GMI}}(Q_X) = \sup_{s\ge 0}\ \mathbb{E}_{Q_X\times W}[i_s(X,Y)]. \tag{6}$$

This work was supported in part by the European Research Council under Grant 725411 and by the Spanish Ministry of Economy and Competitiveness under Grant PID2020-116683GB-C22. Francesc Molina is also supported by the Spanish Ministry of Universities through Margarita Salas Fellowship.

The worst-case GMI is defined as

$$\underline{I}_{\text{GMI}}(Q_X, \widehat{W}, r) = \min_{W \in \mathcal{B}(Q_X, \widehat{W}, r)} \sup_{s \geq 0} \mathbb{E}_{Q_X \times W}[i_s(X, Y)] \quad (7)$$

where the minimization is over all valid conditional probability distributions W in the relative entropy ball $\mathcal{B}(Q_X, \widehat{W}, r)$. Since the true channel is unknown, the worst-case GMI problem (7) finds the channel that gives the worst possible GMI. This gives an indication of the loss incurred by good (but not perfect) channel estimation.

Theorem 1. *Consider a channel estimate $\widehat{W}$ and fixed input distribution Q_X. Then, for sufficiently small $r \geq 0$, the worst-case GMI is*

$$\begin{aligned} &\underline{I}_{\text{GMI}}(Q_X, \widehat{W}, r) \\ &\quad = \sup_{s \geq 0} I_s^{\text{ML}}(Q_X, \widehat{W}) - \sqrt{2r \cdot V_s(Q_X, \widehat{W})} - o(r) \quad (8) \end{aligned}$$

where the term $o(r)$ is non-negative,

$$I_s^{\text{ML}}(Q_X, \widehat{W}) = \mathbb{E}_{Q_X \times \widehat{W}}\big[i_s(X, Y)\big], \quad (9)$$

and

$$V_s(Q_X, \widehat{W}) = \mathbb{E}_{Q_X}\big[\text{Var}_{\widehat{W}}[i_s(X, Y)|X]\big]. \quad (10)$$

Proof. The proof of Theorem 1 is provided in Appendix A; only the main steps are outlined here. We minimize the dual expression for GMI (7) dropping the $o(\cdot)$ term in (4) as [4], thus obtaining an accurate upper bound on $\underline{I}_{\text{GMI}}$ as $r \to 0$. The convex minimization problem is vectorized and then solved using the standard Lagrangian method. □

In addition, observe that for a fixed $\widehat{W}$ the worst-case GMI is upper bounded by the mutual information between input and output achieved through estimated channel $\widehat{W}$ with input Q_X:

$$\begin{aligned} \underline{I}_{\text{GMI}}(Q_X, \widehat{W}, r) &\leq \sup_{s \geq 0} I_s^{\text{ML}}(Q_X, \widehat{W}) \quad (11) \\ &= I_{\text{MI}}(Q_X, \widehat{W}). \quad (12) \end{aligned}$$

Rates above this cannot be achieved. The bound is tight at $r = 0$, in which case $W = \widehat{W}$ and $I_{\text{MI}}(Q_X, W) = I_{\text{MI}}(Q_X, \widehat{W})$.

Corollary 1. *Let the approximate worst-case GMI be*

$$\tilde{\underline{I}}_{\text{GMI}}(Q_X, \widehat{W}, r) = \sup_{s \geq 0} I_s^{\text{ML}}(Q_X, \widehat{W}) - \sqrt{2r \cdot V_s(Q_X, \widehat{W})}. \quad (13)$$

Then, the minimizing channel transition distribution is

$$\tilde{W}^*_{\text{GMI}}(y|x) = \widehat{W}(y|x)\Big(1 - \sqrt{2r} \cdot \varphi(x, y, i_s)\Big) \quad (14)$$

with

$$\varphi(x, y, i_s) \triangleq \frac{i_s(x, y) - \mathbb{E}_{\widehat{W}}[i_s(x, Y)]}{\sqrt{V_s(Q_X, \widehat{W})}}. \quad (15)$$

Observe that $\tilde{W}^*_{\text{GMI}}$ is only a valid conditional probability distribution provided it is non-negative, for which the following condition on the radius of the divergence ball must hold for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, everywhere $\widehat{W}(y|x) > 0$:

$$r < \frac{1}{2\varphi^2(x, y, i_s)}. \quad (16)$$

The condition is not restrictive for sufficiently small r and values of s near the optimal one. Specific examples are shown in the simulations.

Corollary 2. *The approximate worst-case GMI can be lower-bounded by setting $s = 1$ as*

$$\tilde{\underline{I}}_{\text{GMI}} \geq I_{\text{MI}}(Q_X, \widehat{W}) - \sqrt{2r \cdot V_1(Q_X, \widehat{W})}\ ; \quad (17)$$

a tight approximation to the worst-case GMI for sufficiently small values of r. As $r \to 0$, the penalty term shrinks until the estimated channel mutual information $I_{\text{MI}}(Q_X, \widehat{W})$ is achieved, above which rates cannot be achieved.

A. Example: Symmetric $\widehat{W}$ and Equiprobable Q_X

We derive the worst-case GMI for discrete and symmetric estimated channels $\widehat{W}$ and an equiprobable input distribution $Q_X(x) = |\mathcal{X}|^{-1}$ (where $|\mathcal{X}|$ is the cardinality of the input set). Due to the symmetry of $\widehat{W}$, previous expressions can be further simplified and expressed using one of its rows that we denote $\widehat{W}_{\text{sym}}$. The approximate worst-case GMI is given by

$$\begin{aligned} \tilde{\underline{I}}_{\text{GMI}}(Q_X, \widehat{W}, r) = \sup_{s \geq 0} \Bigg\{ &\log \frac{|\mathcal{X}|}{\sum_y \widehat{W}_{\text{sym}}(y)^s} - sH(\widehat{W}_{\text{sym}}) \\ &+ \sqrt{2r \cdot V_s(Q_X, \widehat{W})} \Bigg\} \quad (18) \end{aligned}$$

with

$$\begin{aligned} V_s(Q_X, \widehat{W}) &= s^2 \text{Var}_{\widehat{W}_{\text{sym}}}[\log \widehat{W}_{\text{sym}}] \\ &= s^2 \Big(\mathbb{E}_{\widehat{W}_{\text{sym}}}[\log^2 \widehat{W}_{\text{sym}}] - H^2(\widehat{W}_{\text{sym}})\Big) \end{aligned} \quad (19)$$

and where

$$H(\widehat{W}_{\text{sym}}) = -\sum_y \widehat{W}_{\text{sym}}(y) \log \widehat{W}_{\text{sym}}(y) \quad (20)$$

is the entropy of the probability mass function $\widehat{W}_{\text{sym}}$. Equation (18) can be lower bounded by setting $s = 1$ to yield

$$\tilde{\underline{I}}_{\text{GMI}}(Q_X, \widehat{W}, r) \geq C(\widehat{W}) - \sqrt{2r \cdot \text{Var}_{\widehat{W}_{\text{sym}}}[\log \widehat{W}_{\text{sym}}]} \quad (21)$$

where $C(\widehat{W}) \triangleq \log|\mathcal{X}| - H(\widehat{W}_{\text{sym}})$ is the matched capacity of (symmetric) DMC $\widehat{W}$.

B. Example: Ternary-Input Ternary-Output $\widehat{W}$

We compute the *approximate worst-case* GMI $\tilde{\underline{I}}_{\text{GMI}}$ from (13) for input distribution channel estimate $\widehat{W}$ given by

$$Q_X = \begin{bmatrix} 0.3 & 0.3 & 0.4 \end{bmatrix} \quad (22)$$

$$\widehat{W} = \begin{bmatrix} 0.85 & 0.05 & 0.1 \\ 0.15 & 0.825 & 0.025 \\ 0.025 & 0.1 & 0.875 \end{bmatrix}. \quad (23)$$

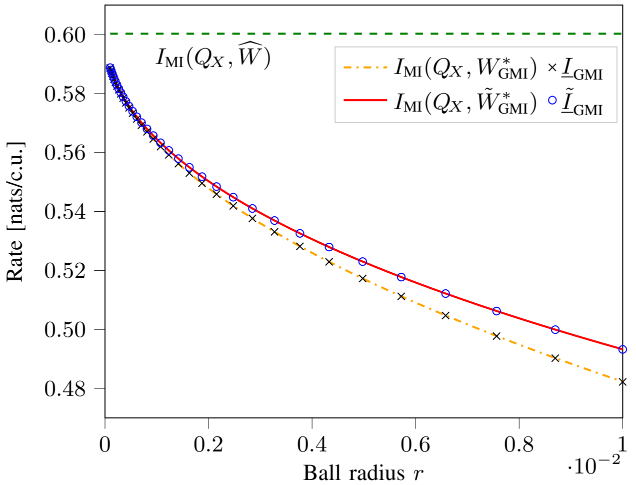


Fig. 1. Achievable rates and approximations computed for fixed Q_X in (22) and estimated channel $\widehat{W}$ in (23).

We plot the approximation in Figure 1 along with the worst-case GMI $\underline{I}_{\text{GMI}}$, numerically computed from (7) using an off-the-shelf solver. To clarify notation, asterisks in superscripts indicate optimal variables.

$I_{\text{MI}}(Q_X, \widehat{W})$ is shown by dashed line in Figure 1 for reference; it is achievable as $r \to 0$. We also plot the mutual information $I_{\text{MI}}(Q_X, \tilde{W}^*_{\text{GMI}})$ of the channel from Corollary 1, as well as that of the optimal channel W^*_{GMI} for the worst-case GMI. For $r > 0$, the curves decrease rapidly from the reference $I_{\text{MI}}(Q_X, \widehat{W})$. In particular, both the true and approximated worst-case GMI decrease with an infinitely negative gradient at $r = 0$. The mutual information for the worst-case channels also exhibits similar behavior. This is because the most harmful channel in the relative entropy ball is such that it causes the GMI rate to decrease with an infinite slope. This shows that even a small mismatch can have a significant impact on the achievable transmission rates.

Our final comment is related to the validity of the approximation. In the example reported in Figure 1, $r = 0.52$ is the maximum radius limit at which $\tilde{W}^*_{\text{GMI}}$ and the corresponding GMI remain positive. For all other channel estimates $\widehat{W}$ we considered, the limit is not restrictive for the range of validity of the approximation, i.e., $r < 0.01$.

Appendix A
Proof of Theorem 1

We formulate the problem based on the dual expression as

$$\begin{aligned}\underline{I}_{\text{GMI}}(Q_X, \widehat{W}, r) &= \min_{W \in \mathcal{B}(Q_X, \widehat{W}, r)} \sup_{s \geq 0} \mathbb{E}_{Q_X \times W}[i_s(X,Y)] \qquad (24)\\ &= \sup_{s \geq 0} \min_{W \in \mathcal{B}(Q_X, \widehat{W}, r)} \mathbb{E}_{Q_X \times W}[i_s(X,Y)] \qquad (25)\end{aligned}$$

The minimax theorem [5] is applied to switch the order of the optimizations from (24) to (25) since $\mathbb{E}_{Q_X \times W}[i_s(X,Y)]$ is convex with respect to W and concave with respect to s [1, Ch. 2.3], and constraints are convex in W.

The inner optimization problem can be vectorized and rewritten in terms of the auxiliary vector

$$\boldsymbol{\theta} = \left[\theta(y_1|x_1), \ldots, \theta(y_{|\mathcal{Y}|}|x_1), \theta(y_1|x_2), \ldots, \theta(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|})\right]^T \qquad (26)$$

where $\theta(y|x) = W(y|x) - \widehat{W}(y|x)$. It follows that for sufficiently small r

$$\begin{aligned}&\underline{I}_s(Q_X, \widehat{W}, r)\\ &= \min_{\substack{\frac{1}{2}\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta} - o(\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta}) \leq r \\ \underline{\mathbf{1}}_j^T \boldsymbol{\theta} = 0,\ 1 \leq j \leq |\mathcal{X}|}} \left\{ I_s^{\text{ML}}(Q_X, \widehat{W}) + \boldsymbol{\theta}^T \nabla I_s \right\} \qquad (27)\\ &= \min_{\substack{\frac{1}{2}\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta} \leq r \\ \underline{\mathbf{1}}_j^T \boldsymbol{\theta} = 0,\ 1 \leq j \leq |\mathcal{X}|}} \left\{ I_s^{\text{ML}}(Q_X, \widehat{W}) + \boldsymbol{\theta}^T \nabla I_s \right\} - o(r) \qquad (28)\end{aligned}$$

with

$$I_s^{\text{ML}}(Q_X, \widehat{W}) = \mathbb{E}_{Q_X \times \widehat{W}}[i_s(X,Y)], \qquad (29)$$

$$K(\widehat{W}) = \operatorname{diag}\left(\frac{Q_X(x_1)}{\widehat{W}(y_1|x_1)}, \ \ldots, \frac{Q_X(x_{|\mathcal{X}|})}{\widehat{W}(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|})} \right), \qquad (30)$$

$$\nabla I_s = [Q_X(x_1) i_s(x_1, y_1), \ \ldots, Q_X(x_{|\mathcal{X}|}) i_s(x_{|\mathcal{X}|}, y_{|\mathcal{Y}|})]^T, \qquad (31)$$

$$\underline{\mathbf{1}}_j = [0 \ \ldots \ 0 \ 1_{(1,j)} \ \ldots \ 1_{(|\mathcal{Y}|,j)} \ 0 \ \ldots \ 0]^T. \qquad (32)$$

In the optimization problem, the $\underline{\mathbf{1}}_j^T \boldsymbol{\theta} = 0$ constraints ensure that for every $x_j \in \mathcal{X}$, $\sum_y W(y|x_j) = 1$. To handle the error terms in the inequality constraint, it is easy to see that the constraint is dominated by the first term as $r \to 0$. Then, the problem can be equivalently written by translating the lowest-order term of the constraint to the cost function as $o(\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta})$, which turns into $o(r)$ after applying the constraint. We do not explicitly impose a positivity constraint on W since a sufficiently small $r \geq 0$ exists such that the positivity of the resulting conditional distribution is guaranteed. The resulting optimization problem is convex, so the KKT conditions are necessary and sufficient [6]. The standard Lagrangian method is used to solve it.

References

[1] J. Scarlett, A. Guillén i Fàbregas, A. Somekh-Baruch, and A. Martinez, "Information-Theoretic Foundations of Mismatched Decoding," *Foundations and Trends® in Communications and Information Theory*, vol. 17, no. 2–3, pp. 149–401, 2020.

[2] G. Kaplan and S. Shamai, "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEU. Archiv für Elektronik und Übertragungstechnik*, vol. 47, no. 4, pp. 228–239, 1993.

[3] P. Boroumand and A. Guillén i Fàbregas, "Mismatched Hypothesis Testing: Error Exponent Sensitivity," *IEEE Transactions on Information Theory*, vol. 68, pp. 6738–6761, 10 2022.

[4] S. Borade and L. Zheng, "Euclidean information theory," in *2008 IEEE International Zurich Seminar on Communications*, 2008, pp. 14–17.

[5] K. Fan, "Minimax Theorems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 39, no. 1, pp. 42–47, 1953.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

Coding for the Gilbert-Elliott Channel: A Mismatched Decoding Perspective

Yutong Han
Technical University of Munich
yutong.han@tum.de

Albert Guillén i Fàbregas
University of Cambridge
Universitat Pompeu Fabra
guillen@ieee.org

***Abstract*—We derive a lower bound to the error exponent of the Gilbert-Elliott channel by means of mismatched decoding. The corresponding achievable rate, the generalized mutual information, is shown to coincide with the lower bound of Mushkin and Bar-David.**

I. Introduction

The Gilbert-Elliott channel [1], [2] is an elementary binary-input binary-output finite-state channel (FSC) described by the two-state Markov chain in Fig. 1. When the channel is in the 'good' state, transmission occurs over a 'good' binary-symmetric channel (BSC) with crossover probability δ_{g}. Similarly, when the channel is in the 'bad' state, transmission occurs over a 'bad' BSC with crossover probability δ_{b}. In other words, the channel transition law $W(y|x,s)$ is determined by the BSC corresponding to the state. Reference [3] defined the channel memory as $\mu \triangleq 1-g-b$. For $\mu > 0$, the channel has a persistent memory, whereas for $\mu < 0$ it has an oscillatory memory. When $\mu = 0$ the channel is said to be memoryless, i.e. the current state is independent of all previous states. The Gilbert-Elliott channel is known to be indecomposable, i.e., the effect of the initial state dies away with time [4, Sec. 4.6]. We denote the stochastic Markov transition matrix by $\Gamma \triangleq \begin{bmatrix} p_{GG} & p_{GB} \\ p_{BG} & p_{BB} \end{bmatrix} = \begin{bmatrix} 1-b & b \\ g & 1-g \end{bmatrix}$ and denote by $[\pi_G, \pi_B] = \left[\frac{g}{g+b}, \frac{b}{g+b}\right]$ the stationary distribution of the Markov chain that defines the channel (see Fig. 1).

We define the channel input and output sequences $x^n, y^n \in \{0,1\}^n$, where n is the length of the sequences. We consider reliable transmission of M messages over the Gilbert-Elliott channel described above. Each message is assigned a codeword from a codebook $\mathcal{C} = \{x_1^n, \dots, x_M^n\}$. The rate of the code is defined as $R = \frac{1}{n}\log M$. The channel capacity of the Gilbert-Elliott channel has been studied in a number of works but no single-letter closed-form expression has yet been found. Reference [3] derived upper and lower bounds to the capacity, which was numerically evaluated in [5]. Since the underlying channels are BSCs, the capacity is attained by an equiprobable input distribution $Q(0) = Q(1) = \frac{1}{2}$.

In this paper, we develop a mismatched decoding (see e.g. [6] and references therein) approach to coding over the Gilbert-Elliott channel. Specifically, we derive a lower bound to the error exponent by means of mismatched decoding, employing a memoryless decoding metric corresponding to a single BSC. We show that the corresponding achievable rate, the generalized mutual information (GMI) [7], coincides with the bound derived by Mushkin and Bar-David [3].

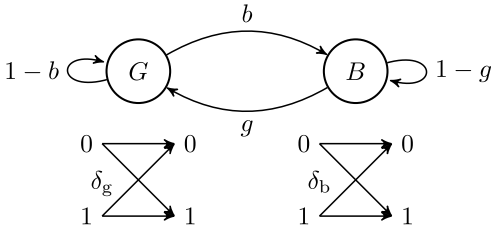


Fig. 1. Gilbert-Elliott channel model.

II. Mismatched Decoding

Mismatched decoding arises in situations where the decoder does not employ a maximum-likelihood decoder, but instead uses a maximum-metric decoder with a sub-optimal decoding metric $q^n(x^n, y^n)$ [6]. This occurs in a number of cases of relevance such as channel uncertainty, reduced-complexity decoding, bit-interleaved coded modulation, and zero-error communication [6]. In addition, mismatched decoding is employed to derive achievable information rates in situations where the channel capacity does not admit simple expressions. In these instances, a decoding metric that somehow simplifies the derivation is chosen. In this paper, although the channel has memory, we will assume a decoding metric that ignores this memory, i.e.,

$$q^n(x^n, y^n) = \prod_{i=1}^{n} q(x_i, y_i). \tag{1}$$

Specifically, we will assume that $q(x,y)$ is the channel transition probability of a single BSC with a crossover probability that depends on the Gilbert-Elliott channel parameters.

Following the footsteps of Gallager [4, Sec. 5.9], it can be shown that there exists a code of rate R and length n such that the error probability for a given message m, given the initial state s_0, can be bounded by

$$P_{e,m}(s_0) \le \mathrm{e}^{-n(E_r(R)-\epsilon)} \tag{2}$$

This work has been funded in part by the European Research Council under ERC grant agreement 725411 and by the Spanish Ministry of Economy and Competitiveness under grant PID2020-116683GB-C22.

for any $\epsilon > 0$, and sufficiently large n, where

$$E_r(R) = \max_{0\le\rho\le 1}\sup_{\tau\ge 0} F_\infty(\rho,\tau) - \rho R \tag{3}$$

with $F_\infty(\rho,\tau) = \lim_{n\to\infty} F_n(\rho,\tau)$,

$$F_n(\rho,\tau) = \max_{Q_{X^n}}\min_{s_0} E_{0,n}(\rho,\tau,Q_{X^n},s_0) \tag{4}$$

$$\begin{aligned}&E_{0,n}(\rho,\tau,Q_{X^n},s_0)\\ &\quad= -\frac{1}{n}\log\mathbb{E}\left[\left(\sum_{\bar{x}^n} Q(\bar{x}^n)\frac{q^n(\bar{x}^n,Y^n)^\tau}{q^n(X^n,Y^n)^\tau}\right)^\rho\right]\end{aligned} \tag{5}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over the joint distribution given the initial state $P(x^n,y^n|s_0)$. The E_0 function will be denoted by $E_{0,n}(\rho,\tau)$ by leaving the dependencies on input distribution and initial state implicit. This exponent naturally leads to the following generalized mutual information rate

$$I_{\text{gmi}} = \sup_{\tau\ge 0}\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\log\frac{q^n(X^n,Y^n)^\tau}{\sum_{\bar{x}^n} Q(\bar{x}^n)q(\bar{x}^n,Y^n)^\tau}\right]. \tag{6}$$

III. Gilbert-Elliott Error Exponent

Let s^n be a binary state sequence of length n and let n_G be the number of good states in s^n. Define the set of all sequences with n_G good states as $\mathcal{T}^n_{n_G}$; this is the binary type class of type $\frac{n_G}{n}$. Also, let $E_0^{\mathrm{g}}(\rho,\tau)$ and $E_0^{\mathrm{b}}(\rho,\tau)$ be the mismatched E_0 functions corresponding to the good and bad BSCs with decoding crossover probabilities δ_q, and define $\Delta E_0(\rho,\tau) \triangleq E_0^{\mathrm{b}}(\rho,\tau) - E_0^{\mathrm{g}}(\rho,\tau)$. By spelling out the expectation in (5) and marginalizing over state sequences s^n, we have that

$$\sum_{s^n} P(s^n)\sum_{x^n,y^n} P(x^n,y^n|s^n)\left(\frac{\sum_{\bar{x}^n} Q(\bar{x}^n)q^n(\bar{x}^n,y^n)^\tau}{q^n(x^n,y^n)^\tau}\right)^\rho \tag{7}$$

$$= \sum_{s^n} P(s^n)\prod_{i=1}^n\sum_{x_i,y_i} P(x_i,y_i|s_i)\left(\frac{\sum_{\bar{x}} Q(\bar{x})q(\bar{x},y_i)^\tau}{q(x_i,y_i)^\tau}\right)^\rho \tag{8}$$

$$= \sum_{n_G=0}^{n}\sum_{\bar{s}^n\in\mathcal{T}^n_{n_G}} P(\bar{s}^n)e^{-n_G E_0^{\mathrm{g}}(\rho,\tau)}e^{-(n-n_G)E_0^{\mathrm{b}}(\rho,\tau)} \tag{9}$$

$$= e^{-nE_0^{\mathrm{b}}(\rho,\tau)}\sum_{n_G=0}^{n}\sum_{\bar{s}^n\in\mathcal{T}^n_{n_G}} P(\bar{s}^n)e^{n_G\Delta E_0(\rho,\tau)} \tag{10}$$

where (8) holds since we assume a product input distribution, a memoryless decoding metric, and the fact that $P(x^n,y^n|s^n) = \prod_{i=1}^n P(x_i,y_i|s_i)$, and (9) follows from re-writing as a function of n_G. We rewrite (10) as the expectation over the random variable N_G with $\Pr\{N_G = n_G\} = \sum_{\bar{s}^n\in\mathcal{T}^n_{n_G}} P(\bar{s}^n)$ as

$$F_\infty(\rho,\tau) = E_0^{\mathrm{b}}(\rho,\tau) - \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[e^{N_G\Delta E_0(\rho,\tau)}\right]. \tag{11}$$

In order to calculate the error exponent of the Gilbert-Elliott channel we need to find $\lim_{n\to\infty} E_{0,n}(\rho,\tau)$. To this end, define the Markov composition of a given sequence s^n as $A(s^n) \triangleq \begin{bmatrix} n_{GG} & n_{GB}\\ n_{BG} & n_{BB}\end{bmatrix}$, where n_{jk} stands for the number of transitions from state j to state k, $j,k\in\{G,B\}$. By normalizing this matrix we get $\Phi(s^n) \triangleq \begin{bmatrix} f_{GG} & f_{GB}\\ f_{BG} & f_{BB}\end{bmatrix}$ where $f_{jk} = n_{jk}/n_j$ with n_j representing the number of state j in sequence s^n. Similarly we define the empirical distribution as $F(s^n) \triangleq [f_G, f_B]$ where $f_j = n_j/n$, and by construction $F(s^n)$ is the stationary distribution for $\Phi(s^n)$.

The probability of a given sequence s^n with Markov composition $A(s^n)$ can be expressed as [8]

$$P(s^n) = p_{GG}^{n_{GG}}p_{GB}^{n_{GB}}p_{BG}^{n_{BG}}p_{BB}^{n_{BB}} \tag{12}$$

$$= \exp\left[\sum_{j,k\in\{G,B\}} n_{jk}\log p_{jk}\right] \tag{13}$$

$$= \exp\left[n\sum_{j\in\{G,B\}} f_j\sum_{k\in\{G,B\}} f_{jk}\log p_{jk}\right] \tag{14}$$

$$= \exp\left[n\sum_{j\in\{G,B\}} f_j\sum_{k\in\{G,B\}} f_{jk}\log\frac{p_{jk}}{f_{jk}} + f_{jk}\log f_{jk}\right] \tag{15}$$

$$= \exp\left[-n\sum_{j\in\{G,B\}} f_j\Big(D(\Phi^{(j)}\|\Gamma^{(j)}) + H(\Phi^{(j)})\Big)\right] \tag{16}$$

$$= \exp\big[-n\big(D(\Phi\|\Gamma|F) + H(\Phi|F)\big)\big] \tag{17}$$

where (14) follows directly from the previous definition; (17) holds since $D(\Phi\|\Gamma|F)$ is the conditional relative entropy between the rows of Φ and those of Γ, that is

$$\begin{aligned}D(\Phi\|\Gamma|F) = f_G&\left(f_{GG}\log\frac{f_{GG}}{p_{GG}} + f_{GB}\log\frac{f_{GB}}{p_{GB}}\right)\\ &+ f_B\left(f_{BG}\log\frac{f_{BG}}{p_{BG}} + f_{BB}\log\frac{f_{BB}}{p_{BB}}\right).\end{aligned} \tag{18}$$

Since $n_j = \sum_k n_{jk} = \sum_k n_{kj}$, symmetry property $n_{GB} = n_{BG}$ holds. Thus, the Markov type of a length-n sequence is determined if n_G and n_{GG} are known. Thus,

$$\sum_{\bar{s}^n\in\mathcal{T}^n_{n_G}} P(\bar{s}^n) = \sum_{n_{GG}=0}^{n_G}\sum_{\bar{s}^n\in\mathcal{A}^n_{n_G,n_{GG}}} p_{GG}^{n_{GG}}p_{GB}^{n_{GB}}p_{BG}^{n_{BG}}p_{BB}^{n_{BB}} \tag{19}$$

$$= \sum_{n_{GG}=0}^{n_G}\left|\mathcal{A}^n_{n_G,n_{GG}}\right|e^{-n[D(\Phi\|\Gamma|F)+H(\Phi|F)]} \tag{20}$$

where $\mathcal{A}^n_{n_G,n_{GG}}$ is the set of sequences with Markov type described by n_G and n_{GG}. Davisson *et al.* [8] showed that for a two-state Markov transition,

$$\left|\mathcal{A}^n_{n_G,n_{GG}}\right| \doteq e^{nH(\Phi|F)} \tag{21}$$

where the notation $a_n \doteq b_n$ means that $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Substituting this into (20), we get

$$\sum_{\bar{s}^n \in \mathcal{T}^n_{n_G}} P(\bar{s}^n) \doteq \sum_{n_{GG}=0}^{n_G} e^{-nD(\Phi\|\Gamma|F)} \tag{22}$$

$$\doteq \max_{n_{GG}\in[0,n_G]} e^{-nD(\Phi\|\Gamma|F)} \tag{23}$$

$$= e^{-nD^*(\Phi\|\Gamma|F)} \tag{24}$$

where we denote $D^*(\Phi\|\Gamma|F) \triangleq \min_{f_{GG}\in[0,f_G]} D(\Phi\|\Gamma|F)$.

Lemma 1. *The minimum relative entropy in* (24) *is given by*

$$D^*(\Phi\|\Gamma|F) = \begin{cases} f_G \log \frac{bg-\beta+2\mu f_G}{2(1-b)\mu f_G} + f_B \log \frac{bg-\beta+2\mu f_B}{2(1-g)\mu f_B} & (\mu \neq 0) \\ f_G \log \frac{f_G}{g} + f_B \log \frac{f_B}{b} & (\mu = 0) \end{cases} \tag{25}$$

and it is achieved when

$$f^*_{GG} = \begin{cases} \frac{bg+2\mu f_G-\beta}{2\mu f_G} & \mu \neq 0, \\ f_G & \mu = 0. \end{cases} \tag{26}$$

where $\beta = \sqrt{b^2g^2 + 4bg\mu f_G f_B}$.

Observe that $D^*(\Phi\|\Gamma|F)$ is a function of f_G and the channel parameters. The divergence becomes zero if and only if $\Phi = \Gamma$. In other words, the empirical distribution is exactly equal to the stationary distribution, namely

$$\min_{f_G\in[0,1]} D^*(\Phi\|\Gamma|F) = D^*(\Phi\|\Gamma|F)\big|_{f_G=\pi_G} = 0. \tag{27}$$

To see this, we differentiate w.r.t. f_G, which gives

$$\frac{\partial D^*(\Phi\|\Gamma|F)}{\partial f_G} = \log \frac{(1-g)f_B f^*_{GG}}{(1-b)(f_B - f_G + f_G f^*_{GG})} = 0 \tag{28}$$

for $\mu \neq 0$. We find that $f_G = \pi_G$ and we write $f^*_{GG} = 1-b$, which by substituting back to (25), we get $D^*(\phi\|\Gamma|F) = 0$, which is achieved uniquely at $f_G = \pi_G$.

Applying the LogSumExp(LSE) inequality $\max_i a_i \leq \log \sum_{i=1}^n \exp(a_i) \leq \log n + \max_i a_i$ and substituting (24) into (11), we obtain

$$F_\infty(\rho,\tau) = E_0^{\mathrm{b}}(\rho,\tau) - \lim_{n\to\infty} \frac{1}{n} \max_{n_G\in[0,n]} \left[n_G \Delta E_0(\rho,\tau) - nD^*(\Phi\|\Gamma|F)\right] \tag{29}$$

$$= E_0^{\mathrm{b}}(\rho,\tau) - \max_{\lambda\in[0,1]} \left[\lambda \Delta E_0(\rho,\tau) - D^*(\Phi\|\Gamma|F)\big|_{f_G=\lambda}\right] \tag{30}$$

where the last equation holds by interchanging the maximum and limit as a result of the function inside the square bracket of (29) being uniformly continuous, and we denote $\lambda \triangleq \lim_{n\to\infty} \frac{n_G}{n}$.

It can be shown that the function $\lambda\Delta E_0(\rho,\tau) - D^*(\Phi\|\Gamma|F)\big|_{f_G=\lambda}$ is concave in λ. Thus, the maximum value can be found by equating the partial derivative to zero. which leads to the optimizing λ^*

$$\lambda^* = \frac{\sqrt{\alpha} - 1 + g + (1-b)e^{\Delta E_0(\rho,\tau)}}{2\sqrt{\alpha}} \tag{31}$$

with $\alpha = (1-g)^2 + 2(bg-\mu)e^{\Delta E_0(\rho,\tau)} + (1-b)^2 e^{2\Delta E_0(\rho,\tau)}$. By construction λ^* is independent of the blocklength, and substituting λ^* into (30) gives rise to the following theorem.

Theorem 1. *The mismatched Gilbert-Elliott* F_∞ *function is equal to*

$$F_\infty(\rho,\tau) = \lambda^* E_0^{\mathrm{g}}(\rho,\tau) + (1-\lambda^*)E_0^{\mathrm{b}}(\rho,\tau) + D^*(\Phi\|\Gamma|F)|_{f_G=\lambda^*} \tag{32}$$

with λ^* *given in* (31).

For memoryless channels with $\mu = 0$, i.e., the current state is independent of all previous states, we have

$$\lambda^* = \frac{ge^{\Delta E_0(\rho,\tau)}}{b + ge^{\Delta E_0(\rho,\tau)}}. \tag{33}$$

Together with (25), Theorem 1 can be written in the form

$$F_\infty(\rho,\tau) = E_0^{\mathrm{b}}(\rho,\tau) - \log\left(b + ge^{\Delta E_0(\rho,\tau)}\right). \tag{34}$$

Observe that applying Jensen's inequality to (11) yields

$$E_{0,n}(\rho,\tau) \leq E_0^{\mathrm{b}}(\rho,\tau) - \frac{\mathbb{E}[N_G]}{n}\Delta E_0(\rho,\tau). \tag{35}$$

Since the definition of stationarity implies

$$\lim_{n\to\infty} \frac{\mathbb{E}[N_G]}{n} = \pi_G \tag{36}$$

this gives the simple upper bound

$$F_\infty(\rho,\tau) \leq \pi_G E_0^{\mathrm{g}}(\rho,\tau) + \pi_B E_0^{\mathrm{b}}(\rho,\tau). \tag{37}$$

IV. Generalized Mutual Information

In this section, we study the GMI of the Gilbert-Elliott channel. We write the GMI as

$$I_{\mathrm{gmi}} = \sup_{\tau\geq 0} I_{\mathrm{gmi}}(\tau). \tag{38}$$

We rewrite (6) using the assumption of memoryless decoding metric as stated in (1)

$$I_{\mathrm{gmi}}(\tau) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \log \frac{q(X_i,Y_i)^\tau}{\sum_{\bar{x}} Q(\bar{x})q(\bar{x},Y_i)^\tau}\right] \tag{39}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{s^n} P(s^n) \sum_{x^n,y^n} P(x^n,y^n|s^n,s_0) \times \sum_{i=1}^n \log \frac{q(x_i,y_i)^\tau}{\sum_{\bar{x}} Q(\bar{x})q(\bar{x},y_i)^\tau} \tag{40}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{s^n} P(s^n) \prod_{i=1}^n \sum_{x_i,y_i} P(x_i,y_i|s_i) \times \sum_{j=1}^n \log \frac{q(x_j,y_j)^\tau}{\sum_{\bar{x}} Q(\bar{x})q(\bar{x},y_j)^\tau} \tag{41}$$

where (40) follows by marginalizing over the state sequence s^n and (41) uses the fact that the state sequence is independent of the input sequence. Using the distributive law of multiplication and the fact that the term inside the logarithm only selects the corresponding joint probability while the rest will sum up to one, we can express the GMI in terms of conditional expectation as

$$I_{\text{gmi}}(\tau) = \lim_{n\to\infty} \frac{1}{n} \sum_{s^n} P(s^n) \sum_{i=1}^{n} \mathbb{E}\left[\log \frac{q(X_i, Y_i)^\tau}{\sum_{\bar{x}} Q(\bar{x}) q(\bar{x}, Y_i)^\tau} \middle| S_i \right] \tag{42}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{s^n} P(s^n) \sum_{i=1}^{n} I_{\text{gmi}}^{s_i}(\tau) \tag{43}$$

which is a general expression for FSCs with state transition being independent of the input sequence under any memoryless decoding metric. For the Gilbert-Elliott channel using the same argument as before, we can write

$$I_{\text{gmi}}(\tau)$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{n_G=0}^{n} \sum_{\bar{s}^n \in \mathcal{T}_{n_G}^n} P(\bar{s}^n) \left[n_G I_{\text{gmi}}^{\text{g}}(\tau) + n_B I_{\text{gmi}}^{\text{b}}(\tau) \right] \tag{44}$$

$$= I_{\text{gmi}}^{\text{b}}(\tau) + \left[I_{\text{gmi}}^{\text{g}}(\tau) - I_{\text{gmi}}^{\text{b}}(\tau) \right] \lim_{n\to\infty} \frac{\mathbb{E}(N_G)}{n} \tag{45}$$

which using (36) yields

$$I_{\text{gmi}} = \sup_{\tau \geq 0} \pi_G I_{\text{gmi}}^{\text{g}}(\tau) + \pi_B I_{\text{gmi}}^{\text{b}}(\tau), \tag{46}$$

the weighted sum of the GMIs per channel, weighted by the stationary distribution.

Given that the memoryless decoding metric $q(x, y)$ can be chosen arbitrarily, we select it as a BSC with crossover probability δ_q with $0 < \delta_q < 0.5$. In this case, we have that

$$I_{\text{gmi}}^{\text{g}}(\tau) = \log \frac{2}{\delta_q^\tau + (1-\delta_q)^\tau} + (1-\delta_{\text{g}}) \log(1-\delta_q)^\tau + \delta_{\text{g}} \log \delta_q^\tau \tag{47}$$

and $I_{\text{gmi}}^{\text{b}}(\tau)$ has a similar form with δ_{g} replaced by δ_{b}. It was shown in [6, Sec. 2] that for a given $q(x, y)$, the GMI is a concave maximization problem (38).

For the optimal τ, the I_{gmi} can be shown to be concave in δ_q. Then we can determine the optimal value of τ and δ_q that maximize the GMI by setting the partial derivatives to zero, yielding $\tau = 1$ and $\delta_q = \pi_G \delta_{\text{g}} + \pi_B \delta_{\text{b}}$.

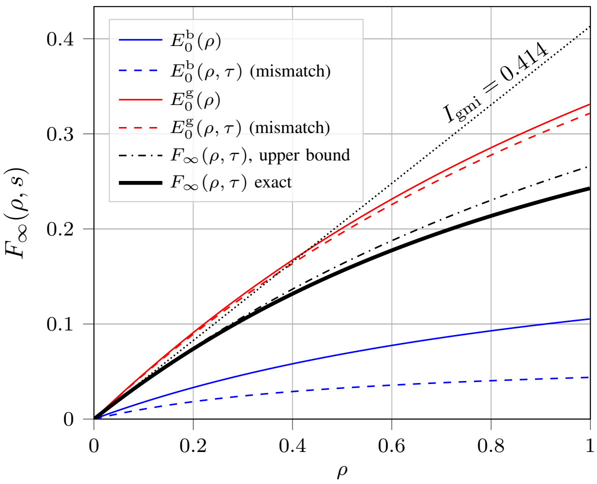


Fig. 2. Function F_∞ for a persistent Gilbert-Elliott channel with parameters $b = 0.1,\ g = 0.4,\ \delta_{\text{g}} = 0.05$ and $\delta_{\text{b}} = 0.2$.

Theorem 2. *The GMI of the Gilbert-Elliott channel using a mismatched BSC with crossover probability $\delta_q = \pi_G \delta_{\text{g}} + \pi_B \delta_{\text{b}}$ for decoding is*

$$I_{\text{gmi}} = \log 2 - h_2(\pi_G \delta_{\text{g}} + \pi_B \delta_{\text{b}}) \tag{48}$$

where $h_2(p) \triangleq -p \log p - (1-p) \log(1-p)$ denotes the binary entropy function. Equality in (48) *is attained if and only if $\mu = 0$, i.e., in the memoryless case.*

The capacity lower bound (48) coincides with the lower bound Mushkin and Bar-David [3, eq. (2.31)].

We illustrate Theorems 1 and 2 by means of an example for a persistent Gilbert-Elliott channel with parameters $b = 0.1,\ g = 0.4,\ \delta_{\text{g}} = 0.05$ and $\delta_{\text{b}} = 0.2$. The F_∞ function given in (32) and the Jensen's inequality upper bound in (37) are depicted in Fig. 2 together with those for the good and bad states using a mismatched BSC with crossover probability $\delta_q = \pi_G \delta_{\text{g}} + \pi_B \delta_{\text{b}}$. The parameter τ has been optimized in all curves. We use (48) to compute

$$I_{\text{gmi}} = \log 2 - h_2\left(\frac{0.4}{0.4 + 0.1} \times 0.05 + \frac{0.1}{0.4 + 0.1} \times 0.2 \right) \tag{49}$$

$$= 0.414 \text{ nat/channel use} \tag{50}$$

which coincides with the gradient of $F_\infty(\rho, \tau)$ at $\rho = 0$. As we observe, the upper bound is very close to F_∞ for small values of ρ.

References

[1] E. N. Gilbert, "Capacity of a burst-noise channel," *The Bell System Technical Journal*, vol. 39, no. 5, pp. 1253–1265, 1960.

[2] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *The Bell System Technical Journal*, vol. 42, no. 5, pp. 1977–1997, 1963.

[3] M. Mushkin and I. Bar-David, "Capacity and coding for the gilbert-elliott channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1277–1290, 1989.

[4] R. Gallager, *Information Theory and Reliable Communication.* USA: John Wiley & Sons, Inc., 1968.

[5] M. Rezaeian, "Computation of capacity for gilbert-elliott channels, using a statistical method," in *2005 Australian Communications Theory Workshop*, 2005, pp. 56–61.

[6] J. Scarlett, A. Guillén i Fàbregas, A. Somekh-Baruch, and A. Martinez, "Information-theoretic foundations of mismatched decoding," *Found. Trends Commun. Inf. Theory*, vol. 17, no. 2-3, pp. 149–401, 2020. [Online]. Available: https://doi.org/10.1561/0100000101

[7] G. Kaplan and S. Shamai, "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEU. Archiv für Elektronik und Übertragungstechnik*, vol. 47, no. 4, pp. 228–239, 1993.

[8] L. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. Inf. Theory*, vol. 27, no. 4, pp. 431–438, 1981.

Error Exponents of Block Codes for Finite-State Sources with Mismatch

Mehdi Dabirnia
Universitat Pompeu Fabra
CTTC
mdabirnia@cttc.es

Albert Guillén i Fàbregas
University of Cambridge
Universitat Pompeu Fabra
guillen@ieee.org

***Abstract*—We derive random-coding/binning error exponents for block source coding for finite-state sources. Specifically, our derivation accounts for a mismatch in the finite-state source model, recovers known special cases and provides an achievable rate, the generalized entropy rate, that quantifies the loss in rate with respect to the entropy rate induced by mismatch.**

I. Introduction

Consider a finite-state source with source alphabet $\mathcal{V}$, a finite set of states $\mathcal{S} = \{1, 2, \dots, A\}$, A conditional probability measures $\{p(\cdot|s)\}_{s\in\mathcal{S}}$, and a next-state function $f : \mathcal{S} \times \mathcal{V} \to \mathcal{S}$. Given an initial state s_0, the conditional probability of a source sequence $v^n = v_1, \dots, v_n \in \mathcal{V}^n$ is defined as

$$P_{V^n|S}(v^n|s_0) = \prod_{i=1}^{n} P_{V|S}(v_i|s_{i-1}) \tag{1}$$

where $s_i = f(s_{i-1}, v_i)$ for all $1 \le i \le n$. We define,

$$\bar{P}_{V^n}(v^n) = \sum_{s_0} \frac{1}{A} P_{V^n|S}(v^n|s_0). \tag{2}$$

With some abuse of notation we will use $\mathcal{S}$ to refer to this source model. A finite-state model $\mathcal{S}$ is said to be irreducible if and only if it is possible, with nonzero probability, to reach each state from any other state in a finite number of states.

A block source code $\mathcal{C}(n, R)$ is defined as a mapping $g : \mathcal{V}^n \to \mathcal{X}$ of source n-tuples $v^n \in \mathcal{V}^n$ to a set of indices/bins/codewords $\mathcal{X} = \{1, \dots, M\}$ where $R = \frac{1}{n}\log M$ is the code rate. A decoder $\phi : \mathcal{X} \to \mathcal{V}^n$ maps each index/bin/codeword back into a source n-tuple $\hat{v}^n$. Typically the number of codewords is smaller than the number of n-tuples and the decoder makes an error whenever $\phi(x(v^n)) \neq v^n$.

Given the source model $\mathcal{S}$, the initial state s_0 and a block source code $\mathcal{C}(n, R)$, the decoder that minimizes the average probability of the error is the maximum-likelihood (ML) decoder which maps each codeword x to the most likely source sequence encoded into x, i.e.,

$$\hat{v}^n = \phi(x) = \underset{v^n:\, g(v^n)=x}{\arg\max}\; P_{V^n}(v^n|s_0). \tag{3}$$

Block source codes have been considered in a number of works in different contexts, see e.g. [1]–[4] and references therein.

In practical systems, the exact source model is almost never known exactly and we are bound to fit some model to the source data and use the extracted model for designing an efficient code. Therefore it is meaningful to assume that in a block source coding setup, the decoder always uses a mismatched source model for decoding. Such a model can be generated during the encoding process and shared with the decoder to allow for correct decoding.

II. Main Results

Assume that instead of the real source model $\mathcal{S}$, we describe the source with a mismatched model with a finite set of states $\hat{\mathcal{S}} = \{1, 2, \dots, \hat{A}\}$, $\hat{A}$ conditional probability measures $\{Q(\cdot|\hat{s})\}_{\hat{s}\in\hat{\mathcal{S}}}$, and a next-state function $\hat{f} : \hat{\mathcal{S}} \times \mathcal{V} \to \hat{\mathcal{S}}$.

We consider a maximum metric decoding based on mismatched model without the knowledge of initial state as follows,

$$\hat{v}^n = \hat{\phi}(x) = \arg\max_{v^n} q(v^n, x), \tag{4}$$

where

$$q(v^n, x) = \bar{Q}_{V^n}(v^n)\mathbb{1}\{g(v^n) = x\}, \tag{5}$$

and $\bar{Q}_{V^n}(v^n) = \frac{1}{\hat{A}}\sum_{\hat{s}_0} Q_{V^n|S}(v^n|\hat{s}_0)$.

In the following, we consider the random ensemble of (n, R) codes for alphabet $\mathcal{V}$ as the set of all (n, R) block codes where each source n-tuple is mapped randomly, independently and with equal probability $\frac{1}{M}$ into one of the M indices or codewords independent from the initial state s_0. In this paper, we study a random-coding error exponent for finite-state sources $\mathcal{S}$ with decoding based on a mismatched source model $\hat{\mathcal{S}}$.

Theorem 1. *For a finite-state source with irreducible model $\mathcal{S}$, there exists a block code with $M = \lceil e^{nR} \rceil$ codewords such that using a decoder based on mismatched source model $\hat{\mathcal{S}}$ for any initial state $\bar{s}_0$ we have*

$$p_e(\bar{s}_0) \le e^{-n\, e_r(R)} \tag{6}$$

where

$$e_r(R) = \sup_{\rho\in[0,1],\tau\ge 0} \rho R - E_{\mathrm{s}}(\rho, \tau), \tag{7}$$

$$E_{\mathrm{s}}(\rho, \tau) = \log \lambda(\rho, \tau) + \rho \log \hat{\lambda}(\tau) + o(n) \tag{8}$$

This work has been funded in part by the European Research Council under ERC grant agreement 725411 and by the Spanish Ministry of Economy and Competitiveness under grant PID2020-116683GB-C22 and by the Universities and Research Department of Generalitat de Catalunya under grant 2021 SGR 00772.

and $\lambda(\rho,\tau), \hat{\lambda}(\tau)$ are respectively the largest magnitude eigenvalues of the matrices $\Gamma_{\rho,\tau} \in \mathbb{R}^{A\hat{A}\times A\hat{A}}, \hat{\Gamma}_\tau \in \mathbb{R}^{\hat{A}\times\hat{A}}$ with entries

$$\gamma_{j\hat{j}k\hat{k}}(\rho,\tau) = \sum_v \frac{P_{V,S|S}(v,j|k)}{Q_{V,S|S}(v,\hat{j}|\hat{k})^{\tau\rho}} \tag{9}$$

$$\hat{\gamma}_{\hat{j}\hat{k}}(\tau) = \sum_v Q_{V,S|S}(v,\hat{j}|\hat{k})^{\tau}, \tag{10}$$

where $\gamma_{j\hat{j}k\hat{k}}(\rho,\tau)$ is the entry in row $(j-1)\hat{A}+\hat{\jmath}$ and column $(k-1)\hat{A}+\hat{k}$ of matrix $\Gamma_{\rho,\tau}$.

Proof: Since the decoding metric is independent of the initial state we bound the random-coding error probability as

$$\bar{p}_e \leq \sum_{v^n} \bar{P}_{V^n}(v^n)\mathbb{P}\left[\bigcup_{\bar{v}^n \neq v^n} \{q(\bar{v}^n,X) \geq q(v^n,X)\}\right] \tag{11}$$

For events $\{\mathcal{B}_i\}$ it can be shown that for any $0 \leq \rho \leq 1$ we have $\mathbb{P}[\bigcup_i \mathcal{B}_i] \leq (\sum_i \mathbb{P}[\mathcal{B}_i])^\rho$ [5, Ch. 5]. Using the random ensemble definition, for any sequences $v^n, \bar{v}^n$ we have

$$\mathbb{P}[q(\bar{v}^n,X) \geq q(v^n,X)] = \frac{1}{M}\mathbb{1}\left[\bar{Q}_{V^n}(\bar{v}^n) \geq \bar{Q}_{V^n}(v^n)\right] \tag{12}$$

$$\leq \frac{1}{M}\frac{\bar{Q}_{V^n}(\bar{v}^n)^\tau}{\bar{Q}_{V^n}(v^n)^\tau} \tag{13}$$

for any $\tau \geq 0$. Therefore, the average error probability (averaged also over the initial state)

$$\bar{p}_e \leq \frac{1}{M^\rho}\sum_{v^n} \bar{P}_{V^n}(v^n)\Bigg(\sum_{\bar{v}^n \neq v^n} \frac{\bar{Q}_{V^n}(\bar{v}^n)^\tau}{\bar{Q}_{V^n}(v^n)^\tau}\Bigg)^\rho. \tag{14}$$

Since the average error probability over the ensemble is upper bounded as in (14), there is at least one code in the ensemble that satisfies the above bound. Also, since the error probability for such a code is an average over A equally likely states, the conditional error probability given any particular initial state, can be no more than A times the average. This gives a bound on error probability which is valid for any initial state and no longer depends on the assumption of the equally likely states as per (2). Therefore, conditional on any initial state $\bar{s}_0 \in \mathcal{S}$ the average error probability is bounded as

$$\bar{p}_e(\bar{s}_0) \leq \frac{A}{M^\rho}\sum_{v^n\in\mathcal{V}^n} \bar{P}_{V^n}(v^n)\Bigg(\sum_{\bar{v}^n \neq v^n} \frac{\bar{Q}_{V^n}(\bar{v}^n)^\tau}{\bar{Q}_{V^n}(v^n)^\tau}\Bigg)^\rho. \tag{15}$$

For any sequences $v^n, \bar{v}^n$ we have

$$\frac{\bar{Q}_{V^n}(\bar{v}^n)^\tau}{\bar{Q}_{V^n}(v^n)^\tau} \leq \hat{A}^{|\tau-1|}\frac{\sum_{\hat{s}_0} Q_{V^n|S}(\bar{v}^n|\hat{s}_0)^\tau}{\sum_{\hat{s}'_0} Q_{V^n|S}(v^n|\hat{s}'_0)^\tau}. \tag{16}$$

This can be seen by considering separately the cases for $\tau \leq 1$ and $\tau \geq 1$. For $\tau \leq 1$ we use the inequality $(\sum a_i)^r \leq \sum a_i^r$ for $0 < r \leq 1$ to upper bound the numerator and $(\sum P_i a_i)^r \geq \sum P_i a_i^r$ for $r \leq 1$ to lower bound the denominator. For $\tau \geq 1$ we use $(\sum P_i a_i)^r \leq \sum P_i a_i^r$ for $r \geq 1$ to upper bound the numerator and $(\sum a_i)^r \geq \sum a_i^r$ for $r \geq 1$ to lower bound the denominator. Substituting (16) in (15) we get

$$\bar{p}_e(\bar{s}_0) \leq \frac{A\hat{A}^{\rho|\tau-1|}}{M^\rho} \times \sum_{v^n\in\mathcal{V}^n} \bar{P}_{V^n}(v^n)\left(\sum_{\bar{v}^n}\frac{\sum_{\hat{s}_0} Q_{V^n|S}(\bar{v}^n|\hat{s}_0)^\tau}{\sum_{\hat{s}'_0} Q_{V^n|S}(v^n|\hat{s}'_0)^\tau}\right)^\rho. \tag{17}$$

We further bound the term in brackets by swapping the sums over $\bar{v}^n$ and $\hat{s}_0$ and upper bounding the numerator using $(\sum a_i)^r \leq \sum a_i^r$ for $0 < r \leq 1$ and lower bounding the denominator using $(\sum P_i a_i)^r \geq \sum P_i a_i^r$ for $r \leq 1$. This gives

$$\bar{p}_e(\bar{s}_0) \leq \frac{A\hat{A}^{\rho|\tau-1|}}{M^\rho} \times \sum_{v^n\in\mathcal{V}^n} \bar{P}_{V^n}(v^n)\frac{\sum_{\hat{s}_0}\left(\sum_{\bar{v}^n} Q_{V^n|S}(\bar{v}^n|\hat{s}_0)^\tau\right)^\rho}{\hat{A}^{\rho-1}\sum_{\hat{s}'_0} Q_{V^n|S}(v^n|\hat{s}'_0)^{\tau\rho}}. \tag{18}$$

Using (2), changing the order of sums over v^n and s_0 and upper bounding the sum over s_0 by A times maximum over s_0, and upper bounding the sum over $\hat{s}_0$ by $\hat{A}$ times maximum over $\hat{s}_0$ and upper bounding the sum over $\hat{s}'_0$ as

$$\frac{1}{\sum_{\hat{s}'_0} Q_{V^n|S}(v^n|\hat{s}'_0)^{\tau\rho}} \leq \max_{\hat{s}'_0}\frac{1}{\hat{A}Q_{V^n|S}(v^n|\hat{s}'_0)^{\tau\rho}}$$

we obtain

$$\bar{p}_e(\bar{s}_0) \leq \frac{A\hat{A}^{\rho|\tau-1|-\rho+1}}{M^\rho}\max_{s_0}\max_{\hat{s}'_0}\max_{\hat{s}_0} e^{nE_{\mathrm{s}}(\rho,\tau,s_0,\hat{s}_0)} \tag{19}$$

where

$$E_{\mathrm{s}}(\rho,\tau,s_0,\hat{s}'_0,\hat{s}_0) = \frac{1}{n}\log\sum_{v^n\in\mathcal{V}^n} P_{V^n|S}(v^n|s_0)\Bigg(\sum_{\bar{v}^n}\frac{Q_{V^n|S}(\bar{v}^n|\hat{s}_0)^\tau}{Q_{V^n|S}(v^n|\hat{s}'_0)^\tau}\Bigg)^\rho \tag{20}$$

We notice that the bound in (19) is valid for the general finite-state source without the deterministic state transition assumption. Similarly to the channel coding case with ML decoding [5, Sec. 5.9] it can be shown that for any $s_0, \hat{s}'_0, \hat{s}_0$ the function $E_{\mathrm{s}}(\rho,\tau,s_0,\hat{s}'_0,\hat{s}_0)$ is continuous, increasing and convex in ρ with $E_{\mathrm{s}}(0,s,s_0,\hat{s}'_0,\hat{s}_0) = 0$. This will prove important to derive the corresponding achievable rate. In the following, we proceed to simplifying (19).

In the following, we split the maximization argument in (19) into two terms and work out the two terms separately

$$\bar{p}_e(\bar{s}_0) \leq \frac{A\hat{A}^{\rho|\tau-1|-\rho+1}}{M^\rho}\left(\max_{s_0}\max_{\hat{s}_0}\sum_{v^n\in\mathcal{V}^n}\frac{P_{V^n|S}(v^n|s_0)}{Q_{V^n|S}(v^n|\hat{s}_0)^{\tau\rho}}\right. \times \left.\max_{\hat{s}_0}\left(\sum_{\bar{v}^n} Q_{V^n|S}(\bar{v}^n|\hat{s}_0)^\tau\right)^\rho\right), \tag{21}$$

where in the first term we change the notation from $\hat{s}'_0$ to $\hat{s}_0$, since after splitting there is no more confusion between those.

Based on the state transition mechanism of source model $\mathcal{S}$ and mismatched model $\hat{\mathcal{S}}$ given the initial states s_0 and $\hat{s}_0$, the source sequence $v^n = (v_1,\ldots,v_n)$ uniquely determines state sequences $\boldsymbol{s} = \boldsymbol{s}(v^n,s_0)$ and $\hat{\boldsymbol{s}} = \hat{\boldsymbol{s}}(v^n,\hat{s}_0)$. Therefore, similarly to [5, Eq. (5.9.31)] we can define

$$\frac{P_{V^n,\boldsymbol{S}|S}(v^n,\boldsymbol{s}|s_0)}{Q_{V^n,\boldsymbol{S}|S}(v^n,\hat{\boldsymbol{s}}|\hat{s}_0)^{\tau\rho}} = \begin{cases} \frac{P_{V^n|S}(v^n|s_0)}{Q_{V^n|S}(v^n|\hat{s}_0)^{\tau\rho}} & \text{for } \boldsymbol{s}, \hat{\boldsymbol{s}} \\ 0 & \text{othewise,} \end{cases} \tag{22}$$

and

$$\frac{P_{V^n,\boldsymbol{S}|S}(v^n,\boldsymbol{s}|s_0)}{Q_{V^n,\boldsymbol{S}|S}(v^n,\hat{\boldsymbol{s}}|\hat{s}_0)^{\tau\rho}} = \prod_{i=1}^{n} \frac{P_{V,S|S}(v_i,s_i|s_{i-1})}{Q_{V,S|S}(v_i,\hat{s}_i|\hat{s}_{i-1})^{\tau\rho}} \tag{23}$$

where

$$P_{V,S|S}(v_i,s_i|s_{i-1}) = \begin{cases} P_{V|S}(v_i|s_{i-1}) & \text{for } s_i = f(s_{i-1},v_i) \\ 0 & \text{othewise.} \end{cases} \tag{24}$$

We thus write the first term in brackets in (21) as

$$\max_{s_0}\max_{\hat{s}_0}\sum_{\boldsymbol{s},\hat{\boldsymbol{s}}}\sum_{v^n}\prod_{i=1}^{n}\frac{P_{V,S|S}(v_i,s_i|s_{i-1})}{Q_{V,S|S}(v_i,\hat{s}_i|\hat{s}_{i-1})^{\tau\rho}} \tag{25}$$

$$= \max_{s_0}\max_{\hat{s}_0}\sum_{\boldsymbol{s},\hat{\boldsymbol{s}}}\prod_{i=1}^{n}\sum_{v_i}\frac{P_{V,S|S}(v_i,s_i|s_{i-1})}{Q_{V,S|S}(v_i,\hat{s}_i|\hat{s}_{i-1})^{\tau\rho}} \tag{26}$$

Similarly, we write the second term in brackets in (21) as

$$\max_{\hat{s}_0}\left(\sum_{\hat{\boldsymbol{s}}}\prod_{i=1}^{n}\sum_{\bar{v}_i}Q_{V,S|S}(\bar{v}_i,\hat{s}_i|\hat{s}_{i-1})^{\tau}\right)^{\rho}. \tag{27}$$

Now we define an $A\hat{A}\times A\hat{A}$ matrix $\Gamma_{\rho,\tau}$ with elements

$$\gamma_{j\hat{j}k\hat{k}}(\rho,\tau) = \sum_{v}\frac{P_{V,S|S}(v,j|k)}{Q_{V,S|S}(v,\hat{j}|\hat{k})^{\tau\rho}} \tag{28}$$

for $j = f(k,v)$, $\hat{j} = \hat{f}(\hat{k},v)$ and $\gamma_{j\hat{j}k\hat{k}}(\rho,\tau) = 0$ otherwise. We also define an $\hat{A}\times\hat{A}$ matrix $\hat{\Gamma}_{\tau}$ with elements

$$\hat{\gamma}_{\hat{j}\hat{k}}(\tau) = \sum_{v}Q_{V,S|S}(v,\hat{j}|\hat{k})^{\tau}, \tag{29}$$

for $\hat{j} = \hat{f}(\hat{k},v)$ and $\hat{\gamma}_{\hat{j}\hat{k}}(\tau) = 0$ otherwise. Observe that the matrices $\Gamma_{\rho,\tau}$ and $\hat{\Gamma}_{\tau}$, are not always stochastic matrices. We denote by $\mathbf{1}$ and $\hat{\mathbf{1}}$ respectively column vectors of length $A\hat{A}$ and $\hat{A}$ with all 1's and by $e(s_0\hat{s}_0)$ and $e(\hat{s}_0)$ respectively row vectors with a 1 in position corresponding to $(s_0-1)\hat{A}+\hat{s}_0$ and $\hat{s}_0$, and 0 in all other components. We rewrite the bound (21) as

$$\bar{p}_e(\bar{s}_0) \le \frac{A\hat{A}^{\rho|\tau-1|-\rho+1}}{M^{\rho}}\left(\max_{s_0}\max_{\hat{s}_0}e(s_0\hat{s}_0)\Gamma_{\rho,\tau}^{n}\mathbf{1}\right)$$
$$\times\left(\max_{\hat{s}_0}\left(e(\hat{s}_0)\hat{\Gamma}_{\tau}^{n}\mathbf{1}\right)^{\rho}\right) \tag{30}$$

We note that if both actual and mismatched models are irreducible, the product model corresponding to the matrix $\Gamma_{\rho,\tau}$ will have a single irreducible subset and the rest of the product states will be transient states, namely their stationary probability will be zero. Therefore we can omit rows and columns corresponding to those transient states from $\Gamma_{\rho,\tau}$ matrix and obtain an irreducible matrix. Assuming that the matrices $\Gamma_{\rho,\tau}$ and $\hat{\Gamma}_{\tau}$ are irreducible, using Perron-Frobenius theorem we know that they have largest magnitude eigenvalues with real positive values. We denote these dominant eigenvalues by $\lambda(\rho,\tau)$ and $\hat{\lambda}(\tau)$ and their corresponding positive right eigenvectors by $\boldsymbol{u}(\rho,\tau) = \big(u_1(\rho,\tau),\ldots,u_{A\hat{A}}(\rho,\tau)\big)$, $u_{j\hat{j}}(\rho,\tau) > 0$ and $\hat{\boldsymbol{u}}(\tau) = \big(u_1(\tau),\ldots,u_{\hat{A}}(\tau)\big)$, $u_{\hat{j}}(\tau) > 0$ respectively, such that

$$\Gamma_{\rho,\tau}\boldsymbol{u}(\rho,\tau) \;=\; \lambda(\rho,\tau)\boldsymbol{u}(\rho,\tau) \tag{31}$$

$$\hat{\Gamma}_{\tau}\hat{\boldsymbol{u}}(\tau) \;=\; \hat{\lambda}(\tau)\hat{\boldsymbol{u}}(\tau). \tag{32}$$

The positive right eigenvectors $\boldsymbol{u}(\rho,\tau)$ and $\hat{\boldsymbol{u}}(\tau)$ are unique except for a multiplicative factor. To make them unique we assume that

$$\sum_{j\hat{j}}u_{j\hat{j}}(\rho,\tau) = 1, \qquad \sum_{\hat{j}}\hat{u}_{\hat{j}}(\tau) = 1. \tag{33}$$

If we denote by $u_{\max}(\rho,\tau)$ and $u_{\min}(\rho,\tau)$ the largest and smallest component of the positive right eigenvector $\boldsymbol{u}(\rho,\tau)$, then for any s_0 and $\hat{s}_0$ we have

$$\frac{u_{\min}(\rho,\tau)}{u_{\max}(\rho,\tau)}\lambda^{n}(\rho,\tau) \le e(s_0\hat{s}_0)\Gamma_{\rho,\tau}^{n}\mathbf{1} \le \frac{u_{\max}(\rho,\tau)}{u_{\min}(\rho,\tau)}\lambda^{n}(\rho,\tau). \tag{34}$$

Using a similar bound on the second term in (30) we obtain

$$\bar{p}_e(\bar{s}_0) \le \frac{A\hat{A}^{\rho|\tau-1|-\rho+1}}{M^{\rho}}\cdot\frac{u_{\max}(\rho,\tau)}{u_{\min}(\rho,\tau)}\lambda^{n}(\rho,\tau)$$
$$\times\left(\frac{\hat{u}_{\max}(\tau)}{\hat{u}_{\min}(\tau)}\hat{\lambda}^{n}(\tau)\right)^{\rho} \tag{35}$$

$$= e^{-n(\rho R - E_{\mathrm{s}}(\rho,\tau))}, \tag{36}$$

where

$$E_{\mathrm{s}}(\rho,\tau) = \log\lambda(\rho,\tau) + \rho\log\hat{\lambda}(\tau) + \delta_n \tag{37}$$

and

$$\delta_n = \frac{1}{n}\log\left(\frac{u_{\max}(\rho,\tau)}{u_{\min}(\rho,\tau)}\cdot\frac{\hat{u}_{\max}(\tau)^{\rho}}{\hat{u}_{\min}(\tau)^{\rho}}A\hat{A}^{\rho|\tau-1|-\rho+1}\right). \tag{38}$$

Finally, observe that for any $s_0, \hat{s}_0$ using (34) we obtain

$$\left|E_{\mathrm{s}}(\rho,\tau,s_0,\hat{s}_0) - \log\lambda(\rho,\tau) - \rho\log\hat{\lambda}(\tau)\right|$$
$$\le \frac{1}{n}\log\left(\frac{u_{\max}(\rho,\tau)}{u_{\min}(\rho,\tau)}\cdot\frac{\hat{u}_{\max}(\tau)^{\rho}}{\hat{u}_{\min}(\tau)^{\rho}}\right). \tag{39}$$

■

We observe that the term (38) is the only term depending on the block length n is decreasing with n since the argument of the log function is greater than or equal to 1. This in turn shows that the corresponding achievable rate, termed the generalized entropy rate, is non-increasing in n and thus, it is attained in the limit for $n \to \infty$. The generalized entropy rate is defined as

$$H_{\mathrm{ger}}(\mathcal{V}) = \inf_{\tau\ge 0}\sum_{k\hat{k}}u_{k\hat{k}}(0)\sum_{v}-P_{V|S}(v|k)$$
$$\times\log\left(\frac{Q_{V|S}(v|\hat{k})^{\tau}}{\sum_{\hat{k}'}\hat{u}_{\hat{k}'}(\tau)\sum_{\bar{v}}Q_{V|S}(\bar{v}|\hat{k}')^{\tau}}\right), \tag{40}$$

where $u_{k\hat{k}}(0)$ is the stationary probability of the product state $k\hat{k}$ and $\hat{u}_{\hat{k}'}(\tau)$ is the $\hat{k}'$-th element of the eigenvector $\hat{\boldsymbol{u}}(\tau)$.

Theorem 2. *For a finite-state source with irreducible model $\mathcal{S}$ using block code and a decoder based on a mismatched source model $\hat{\mathcal{S}}$, the generalized entropy rate $H_{\text{ger}}(\mathcal{V})$ is an achievable rate.*

Proof: For any $\tau \geq 0$ the generalized entropy rate is obtained as $n \to \infty$ and thus it is given by

$$\begin{aligned} H_{\text{ger}}(\mathcal{V}, \tau) &= \lim_{n\to\infty} \frac{\partial}{\partial\rho} E_{\text{s}}(\rho, \tau, s_0, \hat{s}_0)|_{\rho=0} \\ &= \frac{\partial}{\partial\rho} \log\lambda(\rho,\tau)|_{\rho=0} + \log\hat{\lambda}(\tau). \end{aligned} \tag{41}$$

The next steps to obtain the generalized entropy rate follow similar lines as [6] where Vašek derived the error exponent and entropy rate of an ergodic Markov source. Using (33), from (31) and (32) we have

$$\sum_{j\hat{\jmath}}\sum_{k\hat{k}} \gamma_{j\hat{\jmath}k\hat{k}}(\rho,\tau)u_{k\hat{k}}(\rho,\tau) = \sum_{j\hat{\jmath}} \lambda(\rho,\tau)u_{j\hat{\jmath}}(\rho,\tau) = \lambda(\rho,\tau), \tag{42}$$

and

$$\sum_{\hat{\jmath}}\sum_{\hat{k}} \hat{\gamma}_{\hat{\jmath}\hat{k}}(\tau)\hat{u}_k(\tau) = \sum_{\hat{\jmath}} \hat{\lambda}(\tau)\hat{u}_{\hat{\jmath}}(\tau) = \hat{\lambda}(\tau). \tag{43}$$

In the following, in order to simplify the notation, we define $p_{jk}(v) = P_{V,S|S}(v, j|k)$ and $q_{\hat{\jmath}\hat{k}}(v) = Q_{V,S|S}(v, \hat{\jmath}|\hat{k})$. Taking the derivative of $\log\lambda(\rho,\tau)$ with respect to ρ using (42) and simplifying it we obtain

$$\begin{aligned} \frac{\partial}{\partial\rho}\log\lambda(\rho,\tau) &= -\sum_{j\hat{\jmath}}\sum_{k\hat{k}}\sum_{v} \frac{p_{jk}(v)}{q_{\hat{\jmath}\hat{k}}(v)^{\tau\rho}\lambda(\rho,\tau)} \\ &\times \log(q_{\hat{\jmath}\hat{k}}(v)^{\tau})u_{k\hat{k}}(\rho,\tau) + \sum_{j\hat{\jmath}}\sum_{k\hat{k}} \frac{\gamma_{j\hat{\jmath}k\hat{k}}(\rho,\tau)}{\lambda(\rho,\tau)} u'_{k\hat{k}}(\rho,\tau). \end{aligned} \tag{44}$$

Since for $\rho = 0$ from (28) the entries of the matrix $\Gamma_{0,\tau}$ do not depend on τ, we omit the dependence on τ. We observe that the resulting matrix denoted by Γ_0 is a stochastic matrix with column sums equal to 1, i.e., $\sum_{j\hat{\jmath}}\gamma_{j\hat{\jmath}k\hat{k}}(0) = \sum_j p_{jk}(v) = 1$. Therefore it has a largest magnitude eigenvalue $\lambda(0) = 1$ with positive right eigenvector $\boldsymbol{u}(0)$ which is the stationary state distribution of the product finite-state model.

Evaluating (44) at $\rho = 0$ we obtain

$$\begin{aligned} \frac{\partial}{\partial\rho}\log\lambda(\rho,\tau)|_{\rho=0} &= -\sum_{j\hat{\jmath}}\sum_{k\hat{k}}\sum_{v} p_{jk}(v)\log(q_{\hat{\jmath}\hat{k}}(v)^{\tau})u_{k\hat{k}}(0) \\ &+ \sum_{k\hat{k}} u'_{k\hat{k}}(0). \end{aligned} \tag{45}$$

Taking the derivative of both sides in (42) with respect to ρ we obtain

$$\sum_{j\hat{\jmath}} \left(\lambda'(\rho,\tau)u_{j\hat{\jmath}}(\rho,\tau) + \lambda(\rho,\tau)u'_{j\hat{\jmath}}(\rho,\tau)\right) = \lambda'(\rho,\tau). \tag{46}$$

Simplifying the left hand side and canceling $\lambda'(\rho,\tau)$ from both sides we have

$$\lambda(\rho,\tau)\sum_{j\hat{\jmath}} u'_{j\hat{\jmath}}(\rho,\tau) = 0. \tag{47}$$

Since $\lambda(\rho,\tau)$ is strictly positive, we get $\sum_{j\hat{\jmath}} u'_{j\hat{\jmath}}(\rho,\tau) = 0$ and therefore, the second term in (45) is cancelled. Introducing (45) and (43) in (41) we obtain

$$\begin{aligned} H_{\text{ger}}(\mathcal{V},\tau) &= -\sum_{j\hat{\jmath}}\sum_{k\hat{k}}\sum_{v} p_{jk}(v)\log(q_{\hat{\jmath}\hat{k}}(v)^{\tau})u_{k\hat{k}}(0) \\ &+ \log\left(\sum_{\hat{\jmath}}\sum_{\hat{k}}\sum_{v} q_{\hat{\jmath}\hat{k}}(v)^{\tau}\hat{u}_{\hat{k}}(\tau)\right). \end{aligned} \tag{48}$$

Noting that $p_{jk}(v) = 0$ if $j \neq f(k, v)$ and similarly $q_{\hat{\jmath}\hat{k}}(v) = 0$ if $\hat{\jmath} \neq \hat{f}(\hat{k}, v)$, we merge the sums over $j\hat{\jmath}$ and $k\hat{k}$ and also sums over $\hat{\jmath}$ and $\hat{k}$ in (48) obtaining

$$\begin{aligned} H_{\text{ger}}(\mathcal{V},\tau) &= -\sum_{k\hat{k}} u_{k\hat{k}}(0)\sum_{v} P_{V|S}(v|k)\log(Q_{V|S}(v|\hat{k})^{\tau}) \\ &+ \log\left(\sum_{\hat{k}} \hat{u}_{\hat{k}}(\tau)\sum_{v} Q_{V|S}(v|\hat{k})^{\tau}\right). \end{aligned} \tag{49}$$

Noticing that $\sum_{k\hat{k}} u_{k\hat{k}}(0)\sum_v P_{V|S}(v|k) = 1$, combining the two terms in (49) we obtain (40). ■

III. Special cases

Using our general result from Section II, we recover special cases of a memoryless mismatched model and a matched finite-state model.

Theorem 3. *For a finite-state source with irreducible model $\mathcal{S}$, there exists a block code with $M = \lceil e^{nR}\rceil$ codewords such that using a decoder based on a memoryless mismatched source model for any initial state $\bar{s}_0$ we have*

$$\bar{p}_e(\bar{s}_0) \leq e^{-n\, e_r(R)}$$

where

$$e_r(R) = \sup_{\rho\in[0,1],\tau\geq 0} \rho R - E_{\text{s}}(\rho,\tau),$$

and

$$E_{\text{s}}(\rho,\tau) = \log\lambda(\rho,\tau) + \rho\log\sum_{v} Q_V(v)^{\tau} \tag{50}$$

$$+ \frac{1}{n}\log\left(\frac{u_{\max}(\rho,\tau)}{u_{\min}(\rho,\tau)}A\right), \tag{51}$$

with achievable rate

$$H_{\text{ger}}(V) = \inf_{\tau\geq 0} -\sum_{v} P_V(v)\log\frac{Q_V(v)^{\tau}}{\sum_{\bar{v}} Q_V(\bar{v})^{\tau}}, \tag{52}$$

where $P_V(v) = \sum_k u_k(0)P_{V|S}(v|s = k)$.

In the case where the source is also memoryless, (51) reduces to

$$E_{\text{s}}(\rho, s) = \log\sum_{v} P_V(v)\left(\frac{\sum_{\bar{v}} Q_V(\bar{v})^{\tau}}{Q_V(v)^{\tau}}\right)^{\rho}. \tag{53}$$

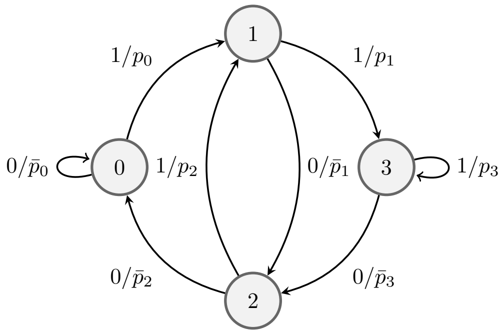


Fig. 1: Binary source with 4 states.

with achievable rate given by (52).

Theorem 4. *For a finite-state source with irreducible model $\mathcal{S}$, there exists a block code with $M = \lceil e^{nR} \rceil$ codewords such that using a matched decoder for any initial state $\bar{s}_0$ we have*

$$\bar{p}_e(\bar{s}_0) \le e^{-n\, e_r(R)}$$

where

$$e_r(R) = \max_{\rho \in [0,1]} \rho R - E_{\mathrm{s}}(\rho),$$

and

$$E_{\mathrm{s}}(\rho) = (1+\rho)\log\lambda(\rho) + \frac{1+\rho}{n}\log\left(\frac{u_{\max}(\rho)}{u_{\min}(\rho)}A\right), \quad (54)$$

with achievable rate

$$\begin{aligned} H_{\mathrm{ger}}(\tau) &= \inf_{\tau \ge 0} -\sum_k u_k(0)\sum_v P_{V|S}(v|k)\log P_{V|S}(v|k), \\ &= \inf_{\tau \ge 0} \sum_k u_k(0) H(V|k). \qquad (55) \end{aligned}$$

Example 1. *Consider a binary source with 4 states given in Fig. 1. The entropy rate of this source is given by $H(\mathcal{V}) = \sum_i \pi_i H(V|s_i)$ where π_i is the stationary probability of being in state s_i. Assuming the conditional distributions of the source as $\{p_0, p_1, p_2, p_3\} = \{0.3, 0.6, 0.2, 0.7\}$ we can calculate the stationary probabilities as $\{\pi_0, \pi_1, \pi_2, \pi_3\} = \{0.4, 0.15, 0.15, 0.3\}$ and the entropy rate of the source as $H(\mathcal{V}) = 0.8708$ bits. Assume that at the decoder we attempt to describe this source with three different models: i) a matched model, ii) a mismatched model with 2 states as shown in Fig. 2 with conditional probabilities $\{p_a, p_b\}$ as $p_a = \sum_{i \in \{0,2\}} \frac{\pi_i}{\pi_0 + \pi_2} p_i = 0.2727$ and $p_b = \sum_{i \in \{1,3\}} \frac{\pi_i}{\pi_1 + \pi_3} p_i = 0.6667$, iii) a memoryless model with distribution $\{1-p, p\}$ $p = \sum_i \pi_i p_i = 0.45$. The generalized entropy rate of the models are $H_{\mathrm{ger}}^{(ii)}(\mathcal{V}) = 0.8782$ and $H_{\mathrm{ger}}^{(iii)}(\mathcal{V}) = 0.9928$ bits, respectively. Fig. 3 illustrates the error exponent and entropy rate losses due mismatch.*

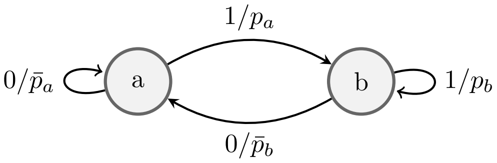


Fig. 2: Mismatched model with 2 states.

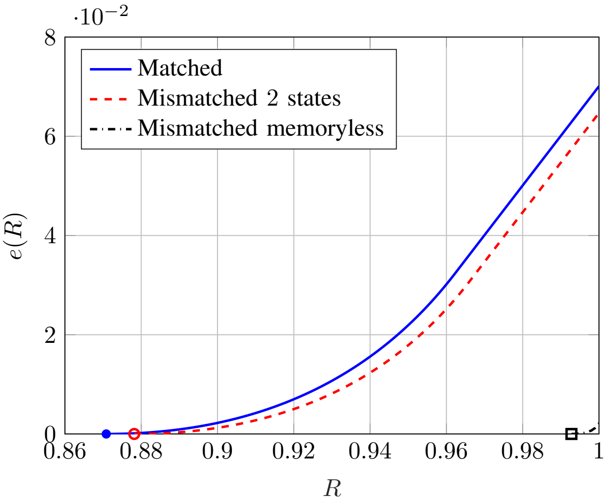


Fig. 3: Error exponents for Example 1.

References

[1] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783–795, 1973.

[2] T. Ancheta, "Syndrome-source-coding and its universal generalization," *IEEE Trans. Inf. Theory*, vol. 22, no. 4, pp. 432–436, 1976.

[3] R. G. Gallager, "Source coding with side information and universal coding," M.I.T. Technical Report LIDS-P-937, Tech. Rep., 1979.

[4] G. Caire, S. Shamai, and S. Verdú, "Noiseless data compression with low-density parity-check codes," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 66, pp. 263–284, 2004.

[5] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc. New York, NY, USA, 1968.

[6] K. Vašek, "On the error exponent for ergodic Markov source," *Kybernetika*, vol. 16, no. 4, pp. 318–329, 1980.

Error exponents for source coding under logarithmic loss

Hamdi Joudeh and Han Wu
ICT Lab, Eindhoven University of Technology, The Netherlands

***Abstract*—In source coding under the logarithmic loss distortion measure, a source is compressed into a message, which is then decompressed into a soft reconstruction (i.e. probability distribution). The distortion is measured by the remaining uncertainty about the source given the message. Shkel and Verdú showed that this lossy source coding setting is intimately related to almost lossless source coding with list decoding, and used this insight to characterize the single-shot excess distortion error probability. In this work, we build upon this connection to list decoding and derive error exponents for source coding under logarithmic loss, without and with side information. The error exponents are closely related to their almost lossless counterparts.**

I. Introduction

In this paper we study the problem of fixed-length lossy source coding of a discrete memoryless source (DMS) under the logarithmic loss (log-loss) distortion measure. While the log-loss is most commonly used in prediction and learning theory, its adoption as a distortion measure in lossy source coding is also natural, specifically in settings where the decoder produces a soft reconstruction (i.e. probability distribution) of the source instead of a point estimate [1]–[3].

The log-loss distortion measure enjoys some mathematical properties that enable elegant characterizations in a number of settings. For instance, under an average distortion criterion, the rate-distortion function is given by [1, Example 2]

$$R(\Delta) = H(X) - \Delta \tag{1}$$

where $H(X)$ is the source entropy and Δ is the average log-loss distortion (assume $0 \leq \Delta \leq H(X)$). The converse for the corresponding coding theorem is obtained by bounding the average log-loss distortion using the conditional entropy of the source given its reconstruction. By building upon this property, Courtade and Weissman [2] derived tight outer bounds in various multi-terminal source coding settings under average log-loss distortion (see also Courtade and Wesel [1]).

More recently, Shkel and Verdú [3] derived single-shot bounds under both excess and average log-loss distortion criteria, without and with decoder side information (see [4], [5] for universal extensions). Key to their approach is a close connection between the log-loss setting and the almost lossless setting with list decoding. As we shall see, this connection to list decoding also plays a central role in our current work.

In this paper, instead of single-shot bounds, we are interested in error exponents under an excess log-loss distortion criterion. We derive error exponents without and with side information, while mainly focusing on universal schemes. We also demonstrate close connections to results in almost lossless settings.

Notation: We use standard notation, briefly explained here. $\mathcal{P}(\mathcal{X})$ denotes the probability simplex on a finite alphabet $\mathcal{X}$. For a probability mass function (pmf) $P_X \in \mathcal{P}(\mathcal{X})$, we denote its support by $\mathcal{S}(P_X)$ and its entropy by $H(P_X)$. The relative entropy between two pmfs Q_X and P_X is denoted by $D(Q_X \| P_X)$. For (X, Y) with joint pmf $P_{XY} = P_{X|Y}P_Y$, the conditional entropy of X given Y is denoted by $H(P_{X|Y}|P_Y)$. The type of a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is denoted by $P_{\boldsymbol{x}}$, and $\mathcal{P}_n(\mathcal{X})$ is the set of all types of sequences in $\mathcal{X}^n$. For $Q \in \mathcal{P}_n(\mathcal{X})$, the corresponding type class is denoted by $\mathcal{T}_n(Q)$. Given a second sequence $\boldsymbol{y} \in \mathcal{Y}^n$, $P_{\boldsymbol{xy}}$ and $P_{\boldsymbol{x}|\boldsymbol{y}}$ are the joint and conditional types. $\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ and $\mathcal{P}_n(\mathcal{X}|\mathcal{Y})$ are the sets of all joint and conditional types. $\mathcal{T}_n(Q_{X|Y}|\boldsymbol{y})$ is the conditional type class of $Q_{X|Y} \in \mathcal{P}_n(\mathcal{X}|\mathcal{Y})$ given $\boldsymbol{y}$. We will make use of types and type classes and their basic properties, such as cardinality and probability bounds (see, e.g., [6, Ch.2]).

II. Source Coding Under Log-Loss

Consider a DMS with finite alphabet $\mathcal{X}$ that randomly generates i.i.d. source sequences $\boldsymbol{X} \triangleq (X_1, X_2, \ldots, X_n)$ according to a pmf $P_X \in \mathcal{P}(\mathcal{X})$. We use $\boldsymbol{x} \triangleq (x_1, x_2, \ldots, x_n)$ to denote a realization of $\boldsymbol{X}$. A *soft reconstruction* of $\boldsymbol{x}$ is a member of $\mathcal{P}(\mathcal{X}^n)$, i.e. a distribution on $\mathcal{X}^n$, denoted by $\hat{P}_n$. The log-loss distortion between $\boldsymbol{x}$ and $\hat{P}_n$ is defined as

$$\mathsf{d}(\boldsymbol{x}, \hat{P}_n) \triangleq \log \frac{1}{\hat{P}_n(\boldsymbol{x})}. \tag{2}$$

The log-loss, also referred to as the self-information loss, can be understood as the remaining uncertainty about $\boldsymbol{x}$ given its reconstruction $\hat{P}_n$ [1]–[3]. For instance, $\mathsf{d}(\boldsymbol{x}, \hat{P}_n)$ is zero if and only if $\hat{P}_n$ has a single mass point at $\boldsymbol{x}$, i.e. an exact *hard reconstruction*; and infinite whenever $\boldsymbol{x}$ has zero probability under $\hat{P}_n$. For convenience, we work with the normalized (per-symbol) log-loss defined as $\mathsf{d}_n(\boldsymbol{x}, \hat{P}_n) \triangleq \frac{1}{n}\mathsf{d}(\boldsymbol{x}, \hat{P}_n)$.

In the lossy source coding setting considered in this work, the sequence $\boldsymbol{X}$ is encoded into a message index from the finite set $\mathcal{M}_n$, which is then decoded into a soft reconstruction from $\mathcal{P}(\mathcal{X}^n)$. A lossy source code of block-length n is thus a pair of mappings $\phi_n : \mathcal{X}^n \to \mathcal{M}_n$ and $\varphi_n : \mathcal{M}_n \to \mathcal{P}(\mathcal{X}^n)$, referred to as the encoder and decoder respectively.

For a lossy source code (ϕ_n, φ_n), the code *rate* is given by $\frac{1}{n}\log|\mathcal{M}_n|$, while $\mathbb{P}\left[\mathsf{d}_n\big(\boldsymbol{X}, \varphi_n(\phi_n(\boldsymbol{X}))\big) > \Delta\right]$ is the excess distortion *error probability* for some distortion level $\Delta \geq 0$. We say that (ϕ_n, φ_n) is an (n, R, Δ, ϵ)-code if

$$\frac{1}{n}\log|\mathcal{M}_n| \leq R \quad \text{and} \quad \mathbb{P}\left[\mathsf{d}_n\big(\boldsymbol{X}, \varphi_n(\phi_n(\boldsymbol{X}))\big) > \Delta\right] \leq \epsilon.$$

This work was partially supported by the European Research Council (ERC) under the ERC Starting Grant N. 101116550 (IT-JCAS).

The minimal error probability for fixed (n, R, Δ) is defined as

$$\varepsilon(n, R, \Delta) \triangleq \inf \{\epsilon : \text{there exists an } (n, R, \Delta, \epsilon)\text{-code}\} .$$

We are interested in characterizing the asymptotic behaviour of $\varepsilon(n, R, \Delta)$, captured through the error exponent defined as

$$E(R, \Delta) \triangleq \varliminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\varepsilon(n, R, \Delta)}. \tag{3}$$

Remark 1. The above problem does not fall under the umbrella of standard lossy source coding in discrete memoryless settings [6], [7]. In the standard paradigm, the reconstruction is a sequence drawn from a discrete product alphabet; and the distortion is additive, i.e. a normalized sum of single-letter distortions. In the setting considered here, the reconstruction alphabet $\mathcal{P}(\mathcal{X}^n)$ is not a product alphabet and is not discrete; and the distortion measure is not additive. For $\mathsf{d}_n(\boldsymbol{x}, \hat{P}_n)$ to be additive, the soft reconstruction $\hat{P}_n$ must be a product distribution, as in earlier works on log-loss source coding [1], [2]. This need not be the case in general, and as we shall see, the codes we propose employ non-product soft reconstructions.

A. Connection to list decoding

In [3], Shkel and Verdú established a fundamental connection between lossy source coding under log-loss and almost lossless source coding with list decoding, leading to an exact characterization of $\varepsilon(n, R, \Delta)$. This connection is central to the approach we take here, therefore, we review it in some detail. We start with a key lemma linking the log-loss of a soft reconstruction to the list size in list decoding.

To this end, fix a soft reconstruction $\hat{P}_n \in \mathcal{P}(\mathcal{X}^n)$ and a distortion level $\Delta \geq 0$. We say that a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is Δ-covered by $\hat{P}_n$ if $\mathsf{d}_n(\boldsymbol{x}, \hat{P}_n) \leq \Delta$. If $\hat{P}_n$ Δ-covers every element of a set (or list) $\mathcal{L}_n \subseteq \mathcal{X}^n$, then the set $\mathcal{L}_n$ is also said to be Δ-covered by the soft reconstruction $\hat{P}_n$.

Lemma 1. *Let $\mathcal{L}_n \subseteq \mathcal{X}^n$. There exists a soft-reconstruction $\hat{P}_n \in \mathcal{P}(\mathcal{X}^n)$ that Δ-covers $\mathcal{L}_n$ if and only if*

$$|\mathcal{L}_n| \leq \lfloor \exp(n\Delta) \rfloor. \tag{4}$$

Proof. The direct part holds by taking $\hat{P}_n$ to be uniform on $\mathcal{L}_n$ and zero elsewhere. The converse part follows from [3, Lemma 1], reproduced here for completeness. Let $\mathcal{B}_n(\Delta, \hat{P}_n)$ be the set of all source sequences Δ-covered by $\hat{P}_n$, i.e.

$$\mathcal{B}_n(\Delta, \hat{P}_n) \triangleq \left\{ \boldsymbol{x} \in \mathcal{X}^n : \mathsf{d}_n(\boldsymbol{x}, \hat{P}_n) \leq \Delta \right\}. \tag{5}$$

It is sufficient to show $|\mathcal{B}_n(\Delta, \hat{P}_n)| \leq \lfloor \exp(n\Delta) \rfloor$. Note that $\boldsymbol{x} \in \mathcal{B}_n(\Delta, \hat{P}_n)$ implies $\hat{P}_n(\boldsymbol{x}) \geq \exp(-n\Delta)$, and therefore

$$1 = \sum_{\boldsymbol{x}\in\mathcal{X}^n} \hat{P}_n(\boldsymbol{x}) \geq \sum_{\boldsymbol{x}\in\mathcal{B}_n(\Delta,\hat{P}_n)} \hat{P}_n(\boldsymbol{x}) \geq \left|\mathcal{B}_n(\Delta, \hat{P}_n)\right| \exp(-n\Delta).$$

The bound is tightened by including the floor function. □

Now consider an almost lossless list source code: here a source sequence is encoded into one of $\lfloor \exp(nR) \rfloor$ message indices, and a message index is decoded into a list of $\lfloor \exp(n\Delta) \rfloor$ source sequences. A decoding error occurs if the generated source sequence is not in the decoded list. From Lemma 1, we see that such a code with error probability ϵ can be converted into a (n, R, Δ, ϵ) log-loss source code. Conversely, a log-loss source code can be converted into an almost lossless list source code. This connection leads to the following result.

Theorem. (Shkel-Verdú [3, Theorem 5-6]). *Let $G : \mathcal{X}^n \to \{1, 2, \ldots, |\mathcal{X}^n|\}$ be a a probability rank function that ranks source sequences in decreasing order of their probability. Then*

$$\varepsilon(n, R, \Delta) = \mathbb{P}\big[G(\boldsymbol{X}) > \lfloor \exp(nR) \rfloor \lfloor \exp(n\Delta) \rfloor\big]. \tag{6}$$

III. Error Exponent

We now present the first result of this paper. To this end, we first define the function $F(R)$ on $0 \leq R < \log |\mathcal{S}(P_X)|$ as

$$F(R) \triangleq \min_{Q_X : H(Q_X) \geq R} D(Q_X \| P_X). \tag{7}$$

Note that $F(R) = E(R, 0)$, which is the error exponent in the almost lossless case [6], [8], [9]. $F(R)$ is continuous, convex and increasing on its domain, with $F(R) = 0$ on $0 \leq R \leq H(P_X)$. The expression in (7) is known as a *primal form.* $F(R)$ admits an equivalent *dual form* given as [6, Prob.2.15]

$$F(R) = \sup_{\rho\geq 0} \rho \left(R - H_{\frac{1}{1+\rho}}(X) \right) \tag{8}$$

where $H_{\frac{1}{1+\rho}}(X)$ is the Rényi entropy of order $1/(1 + \rho)$.

Recall that the log-loss rate-distortion function is given by $R(\Delta) = H(P_X) - \Delta$, and to ensure that $\varepsilon(n, R, \Delta)$ goes to zero, we must have $R > H(P_X) - \Delta$. On the other hand, for rates satisfying $R \geq \log |\mathcal{S}(P_X)| - \Delta$, the whole support of P_X^n can be covered by lists of size $\mathrm{e}^{n\Delta}$. Therefore, the relevant range is $H(P_X) < R + \Delta < \log |\mathcal{S}(P_X)|$.

Theorem 1. *Let (R, Δ) be a rate-distortion pair such that $H(P_X) < R + \Delta < \log |\mathcal{S}(P_X)|$. Then*

$$E(R, \Delta) = F(R + \Delta). \tag{9}$$

We observe that $E(R, \Delta) = E(R + \Delta, 0)$, i.e. the log-loss error exponent as a function of R is merely a translation of the almost lossless error exponent by Δ bits to the left. This can be understood in light of the optimal source code that achieves (6) as follows. For a log-loss source code of rate R and distortion Δ, an excess distortion error occurs when the source generates a sequence with probability rank greater than $\lfloor \mathrm{e}^{nR} \rfloor \lfloor \mathrm{e}^{n\Delta} \rfloor$, that is the number of sequences covered by $\lfloor \mathrm{e}^{nR} \rfloor$ lists of size $\lfloor \mathrm{e}^{n\Delta} \rfloor$ each. On the other hand, an almost lossless source code of rate $R + \Delta$ makes an error when the source generates a sequence with probability rank greater than $\lfloor \mathrm{e}^{n(R+\Delta)} \rfloor$. Asymptotically, the two error events have almost the same probability, yielding in the same error exponent.

The above argument is sufficient for proving Theorem 1, yet it employs a code that depends on the source pmf (or probability rank function). Further on we present an alternative proof using the method of types, extending the Longo-Sgarro approach [9] (see also Csiszár-Körner [6]) from the almost lossless case to the log-loss case. As is often the case with types-based proofs, a universal scheme emerges.

It is worthwhile mentioning that the log-loss error exponent expression in (9) can be obtained from Marton's error exponent [7] by replacing the general rate-distortion function with its log-loss counterpart. While this is perhaps expected, the result in Theorem 1 does not follow directly from Marton's proof, at least not without modification, as the log-loss setting considered here is not a special case of the classical rate-distortion setting (see Remark 1). We shall see next that Theorem 1 is proved directly using the connection to list decoding.

A. Proof of Theorem 1

1) Achievability: Fix $R > 0$ and $n \in \mathbb{N}$, and let $J_n = |\mathcal{P}_n(\mathcal{X})|$ and $M_n = \lfloor (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nR} \rfloor$. For every type $Q \in \mathcal{P}_n(\mathcal{X})$, partition $\mathcal{T}_n(Q)$ into M_n lists all roughly of the same size. $\mathcal{L}_n(\boldsymbol{x})$ denotes the list containing $\boldsymbol{x}$. By construction,

$$|\mathcal{L}_n(\boldsymbol{x})| \leq \left\lceil \frac{|\mathcal{T}_n(P_{\boldsymbol{x}})|}{M_n} \right\rceil \tag{10}$$

for every $\boldsymbol{x} \in \mathcal{X}^n$. An emitted source sequence $\boldsymbol{x}$ is encoded into $\phi_n(\boldsymbol{x}) = \big(t_n(\boldsymbol{x}), l_n(\boldsymbol{x})\big)$, where $t_n(\boldsymbol{x}) \in \{1, 2, \ldots, J_n\}$ is the type index and $l_n(\boldsymbol{x}) \in \{1, 2, \ldots, M_n\}$ is the list index. Upon receiving $\phi_n(\boldsymbol{x})$, the decoder reproduces the list $\mathcal{L}_n(\boldsymbol{x})$ containing $\boldsymbol{x}$. The corresponding soft reconstruction $\varphi_n(\phi_n(\boldsymbol{x})) = \hat{P}_n\big(\,\cdot\,|\phi_n(\boldsymbol{x})\big)$ is set as

$$\hat{P}_n\big(\hat{\boldsymbol{x}}|\phi_n(\boldsymbol{x})\big) = \begin{cases} \frac{1}{|\mathcal{L}_n(\boldsymbol{x})|}, & \hat{\boldsymbol{x}} \in \mathcal{L}_n(\boldsymbol{x}) \\ 0, & \text{otherwise.} \end{cases}$$

The rate of this code satisfies $\frac{1}{n}\log(J_n M_n) \leq R$. For any sequence $\boldsymbol{x} \in \mathcal{X}^n$, the log-loss incurred by the corresponding reconstruction $\varphi_n(\phi_n(\boldsymbol{x}))$ is bounded above as

$$\begin{aligned} \mathsf{d}_n\big(\boldsymbol{x}, \varphi_n(\phi_n(\boldsymbol{x}))\big) &= \frac{1}{n}\log|\mathcal{L}_n(\boldsymbol{x})| \\ &\leq \frac{1}{n}\log\left\lceil \frac{\mathrm{e}^{nH(P_{\boldsymbol{x}})}}{\lfloor \mathrm{e}^{nR - |\mathcal{X}|\log(1+n)} \rfloor} \right\rceil && (11) \\ &\leq H(P_{\boldsymbol{x}}) - R + \delta_n && (12) \end{aligned}$$

where $\delta_n \geq 0$ goes to zero as n grows large.

We now analyze the error probability. To this end, fix $\Delta \geq 0$ such that $H(P_X) < R + \Delta < \log|\mathcal{S}(P_X)|$. We can see from (12) that all source sequences in the set $\mathcal{B}_n$, defined as

$$\mathcal{B}_n = \bigcup_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) \leq R + \Delta - \delta_n} \mathcal{T}_n(Q), \tag{13}$$

are reconstructed with a log-loss not exceeding Δ, and thus the excess distortion error event is included in $\mathcal{X}^n \setminus \mathcal{B}_n$. The error probability under source pmf P_X is bounded above as

$$\begin{aligned} P_X^n(\mathcal{X}^n \setminus \mathcal{B}_n) &= \sum_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) > R + \Delta - \delta_n} P_X^n\big(\mathcal{T}_n(Q)\big) \\ &\leq \sum_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) > R + \Delta - \delta_n} \mathrm{e}^{-nD(Q\|P_X)} && (14) \\ &\leq (n+1)^{|\mathcal{X}|} \max_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) > R + \Delta - \delta_n} \mathrm{e}^{-nD(Q\|P_X)} && (15) \\ &\leq (n+1)^{|\mathcal{X}|} \mathrm{e}^{-nF(R + \Delta - \delta_n)}. && (16) \end{aligned}$$

The above steps are standard and use the properties of types and type classes. Achievability follows from (16) and the continuity of $F(R')$ on $H(P_X) < R' < \log|\mathcal{S}(P_X)|$.

Remark 2. As mentioned earlier, the above source code is universal and does not depend on P_X. The code is also universal with respect to the distortion level Δ, and only depends on the rate R (and block-length n). For fixed R, the same sequence of codes achieves a positive exponent for every P_X and Δ satisfying $R < H(P_X) - \Delta$. The key to the universality with respect to Δ is the *variable list partitioning* of type classes, where list sizes depend on the type and code rate but not on the distortion level (see (10)). The same partitioning is used by Bunte and Lapidoth in [10] in the context of strictly lossless list source coding (also known as task encoding), where the focus is on analyzing list size moments.

Remark 3. In the achievability proof of Marton's error exponent [7], a key ingredient is the type covering lemma which states that for any type $Q \in \mathcal{P}_n(\mathcal{X})$ with a rate-distortion function satisfying $R(Q, \Delta) \leq R - \delta$, the corresponding type class $\mathcal{T}_n(Q)$ can be Δ-covered by e^{nR} reconstruction sequences. The type covering lemma is often proved using random selection (i.e. random coding). In the log-loss setting considered here, type covering is accomplished through simple partitioning, and the proof does not rely on random coding.

2) Converse: Fix a pair (R, Δ) and a source pmf P_X such that $H(P_X) < R + \Delta < \log|\mathcal{S}(P_X)|$. For every block-length n, let $(\phi_n^\star, \varphi_n^\star)$ be an optimal code achieving the minimal error probability $\varepsilon(n, R, \Delta)$. Moreover, define the set $\mathcal{B}_n^\star \triangleq \big\{\boldsymbol{x} \in \mathcal{X}^n : \mathsf{d}_n\left(\boldsymbol{x}, \varphi_n^\star\left(\phi_n^\star(\boldsymbol{x})\right)\right) \leq \Delta\big\}$. An error occurs whenever the source produces a sequence in $\mathcal{X}^n \setminus \mathcal{B}_n^\star$.

Let $M(\Delta, \mathcal{B}_n^\star)$ be the minimum number of soft reconstruction required to Δ-cover $\mathcal{B}_n^\star$. From Lemma 1, we know that any soft reconstruction can Δ-cover at most $\lfloor \mathrm{e}^{n\Delta} \rfloor$ source sequence. Therefore, we must have

$$M(\Delta, \mathcal{B}_n^\star) \geq \left\lceil \frac{|\mathcal{B}_n^\star|}{\lfloor \mathrm{e}^{n\Delta} \rfloor} \right\rceil. \tag{17}$$

It immediately follows that the rate R of $(\phi_n^\star, \varphi_n^\star)$ must satisfy

$$\mathrm{e}^{nR} \geq M(\Delta, \mathcal{B}_n^\star) \geq |\mathcal{B}_n^\star| \mathrm{e}^{-n\Delta}. \tag{18}$$

Now let $\delta_n = \frac{|\mathcal{X}|}{n}\log(1+n) + \frac{1}{n}\log 2$ and let $Q \in \mathcal{P}_n(\mathcal{X})$ be a type such that $H(Q) \geq R + \Delta + \delta_n$. The cardinality of the corresponding type class $\mathcal{T}_n(Q)$ is bounded below as

$$|\mathcal{T}_n(Q)| \geq (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nH(Q)} \geq 2\mathrm{e}^{n(R+\Delta)} \geq 2|\mathcal{B}_n^\star|$$

from which we conclude that at least half of the sequences in $\mathcal{T}_n(Q)$ are not contained in $\mathcal{B}_n^\star$. Therefore

$$\begin{aligned} \varepsilon(n, R, \Delta) &= P_X^n\left(\mathcal{X}^n \setminus \mathcal{B}_n^\star\right) \geq \frac{1}{2} P_X^n\left(\mathcal{T}_n(Q)\right) \\ &\geq \frac{1}{2}(n+1)^{-|\mathcal{X}|} \mathrm{e}^{-nD(Q\|P_X)} = \mathrm{e}^{-nD(Q\|P_X) - n\delta_n} && (19) \end{aligned}$$

obtained from the standard type class probability lower bound. This holds for all types satisfying $H(Q) \geq R + \Delta + \delta_n$, hence

$$-\frac{1}{n}\log\varepsilon(n, R, \Delta) \leq \delta_n + \min_{Q \in \mathcal{P}_n(\mathcal{X}) : H(Q) \geq R + \Delta + \delta_n} D(Q\|P_X)$$

$$= \delta_n + F_n(R + \Delta + \delta_n) \tag{20}$$

where $F_n(R')$ is defined as $F(R')$ in (7) except that the minimization is over types in $\mathcal{P}_n(\mathcal{X})$ instead of all pmfs in $\mathcal{P}(\mathcal{X})$. By definition, we know that $F(R') \leq F_n(R')$. In addition, it can be shown that $F_n(R') \leq F(R') + \delta'_n$ for some $\delta'_n > 0$ that goes to zero as n grows large.[1] By combining this with (20) and taking the limit, the converse result follows.

IV. Side Information

In this section we consider settings with side information. Here we have a pair of DMSs with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. The sources randomly generate an i.i.d. sequence of pairs $(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \big((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\big)$ according to a joint pmf $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The goal remains to compress the sequence $\boldsymbol{X}$ and then decompress it into a soft reconstruction in $\mathcal{P}(\mathcal{X}^n)$, but now $\boldsymbol{Y}$ is available either at both the encoder and decoder sides, or at the decoder side only.

A. Encoder-decoder side information

For the case where the side information sequence is available at both the encoder and decoder sides, a lossy source code of block-length n is given by the pair of mappings $\phi_n : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{M}_n$ and $\varphi_n : \mathcal{M}_n \times \mathcal{Y}^n \to \mathcal{P}(\mathcal{X}^n)$. A (n, R, Δ, ϵ)-code, minimal error probability and error exponent are defined in a standard manner. The latter two are denoted by $\varepsilon_{X|Y}(n, R, \Delta)$ and $E_{X|Y}(R, \Delta)$ respectively.

This problem is very similar to its counterpart with no side information, i.e. given $\boldsymbol{Y} = \boldsymbol{y}$, the problem reduces to encoding and decoding a memoryless source (not necessarily i.i.d.) with distribution $P_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^n P_{X|Y}(x_i|y_i)$. Nevertheless, it is still useful to characterize the error exponent in this case, as it provides an upper bound for the more interesting case with decoder side information only. To this end, we define $F_{X|Y}(R)$ on $0 \leq R < \log|\mathcal{S}(P_X)|$ as

$$F_{X|Y}(R) \triangleq \min_{Q_{XY} : H(Q_{X|Y}|Q_Y) \geq R} D(Q_{XY} \| P_{XY}). \tag{21}$$

$F_{X|Y}(R)$ is continuous, convex and increasing on its domain; and is zero for $0 \leq R \leq H(P_{X|Y}|P_Y)$. Moreover, we have $F_{X|Y}(R) = E_{X|Y}(R, 0)$, that is the error exponent in the almost lossless case. $F_{X|Y}(R)$ also admits the following dual form in terms of Arimoto's conditional Rényi entropy [11]

$$F_{X|Y}(R) = \sup_{\rho \geq 0} \rho \left(R - H_{\frac{1}{1+\rho}}(X|Y)\right) \tag{22}$$

obtained using Lagrangian duality techniques (see, e.g., the proof of [10, Equation (32)] by Bunte and Lapidoth).

Theorem 2. *Let (R, Δ) be a rate-distortion pair such that $H(P_{X|Y}|P_Y) < R + \Delta < \log|\mathcal{S}(P_X)|$. Then*

$$E_{X|Y}(R, \Delta) = F_{X|Y}(R + \Delta). \tag{23}$$

The proof of Theorem 2 (omitted for brevity) is very similar to that of Theorem 1, but relies on conditional types.

B. Decoder side information: Wyner-Ziv

We now turn our attention to the case where the side information sequence $\boldsymbol{Y}$ is only available at the decoder. This is the Wyner-Ziv setting, specialized to the log-loss distortion measure. A lossy source code of block-length n here is given by the pair $\phi_n : \mathcal{X}^n \to \mathcal{M}_n$ and $\varphi_n : \mathcal{M}_n \times \mathcal{Y}^n \to \mathcal{P}(\mathcal{X}^n)$. The minimal error probability and error exponent are denoted by $\varepsilon^{\mathrm{WZ}}_{X|Y}(n, R, \Delta)$ and $E^{\mathrm{WZ}}_{X|Y}(R, \Delta)$ respectively.

Next, we observe that the encoder-decoder side information result in Theorem 2 provides the following converse bound

$$E^{\mathrm{WZ}}_{X|Y}(R, \Delta) \leq E^{\mathrm{sp}}_{X|Y}(R, \Delta) \triangleq F_{X|Y}(R + \Delta). \tag{24}$$

For $\Delta = 0$, the setting reduces to the Slepian-Wolf problem, and we denote the error exponent by $E^{\mathrm{SW}}_{X|Y}(R)$. The bound $E^{\mathrm{SW}}_{X|Y}(R) \leq F_{X|Y}(R)$, a special case of (24), was obtained by Gallager in [12] (see also Csiszár and Körner [13, Theorem 3]). This bound on the Slepian-Wolf error exponent is sometimes referred to as the sphere-packing exponent [14], due to close resemblance to the sphere-packing exponent in channel coding. Similarly, the bound in (24) can be thought of as a sphere-packing exponent for the log-loss Wyner-Ziv setting.

Next, we derive a lower bound for $E^{\mathrm{WZ}}_{X|Y}(R, \Delta)$. Define the function $\tilde{F}_{X|Y}(R)$ on $0 \leq R < \log|\mathcal{S}(P_X)|$ as

$$\tilde{F}_{X|Y}(R) \triangleq \min_{Q_{XY}} \left\{ D(Q_{XY} \| P_{XY}) + \left|R - H(Q_{X|Y}|Q_Y)\right|^+ \right\}$$

where $|a|^+ \triangleq \max\{0, a\}$. Note that $\tilde{F}_{X|Y}(R) \leq F_{X|Y}(R)$. Moreover, $\tilde{F}_{X|Y}(R)$ admits a dual form

$$\tilde{F}_{X|Y}(R) = \max_{\rho \in [0,1]} \rho \left(R - H_{\frac{1}{1+\rho}}(X|Y)\right). \tag{25}$$

$\tilde{F}_{X|Y}(R)$ is an achievable error exponent in the Slepian-Wolf setting [12], [13], referred to as the random-coding error exponent, as it is achieved through random coding (or binning) in close resemblance to the random-coding exponent in channel coding. A corresponding random-coding error exponent for the log-loss Wyner-Ziv problem is presented next.

Theorem 3. *Let (R, Δ) be a rate-distortion pair such that $H(P_{X|Y}|P_Y) < R + \Delta < \log|\mathcal{S}(P_X)|$. Then*

$$E^{\mathrm{WZ}}_{X|Y}(R, \Delta) \geq E^{\mathrm{r}}_{X|Y}(R, \Delta) \triangleq \tilde{F}_{X|Y}(R + \Delta).$$

For fixed Δ, the exponents $E^{\mathrm{r}}_{X|Y}(R, \Delta)$ and $E^{\mathrm{sp}}_{X|Y}(R, \Delta)$ coincide on $H(P_{X|Y}|P_Y) - \Delta < R \leq R_{\mathrm{cr}}$, where R_{cr} is the largest rate at which the convex curve $E^{\mathrm{sp}}_{X|Y}(R, \Delta)$, as a function of R, meets it supporting line of slope 1. Note that R_{cr} is reminiscent of the *critical rate* in channel coding. Above this rate, the two exponents differ in general.

C. Proof of Theorem 3

The proof is based on random binning and a list decoding variant of the universal minimum entropy decoding rule [13], [15]. In the analysis of this scheme, we use $H(\boldsymbol{x}|\boldsymbol{y})$ as a shorthand notation for the conditional entropy $H(P_{\boldsymbol{x}|\boldsymbol{y}}|P_{\boldsymbol{y}})$ calculated from the joint type $P_{\boldsymbol{x}\boldsymbol{y}} = P_{\boldsymbol{x}|\boldsymbol{y}} P_{\boldsymbol{y}}$.

[1] By continuity and since $\bigcup_{n \in \mathbb{N}} \mathcal{P}_n(\mathcal{X})$ is dense in $\mathcal{P}(\mathcal{X})$ [9, Rem. 2].

Let $J_n = |\mathcal{P}_n(\mathcal{X})|$ and $M_n = \lfloor (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nR} \rfloor$ for fixed R and n. A binning function $b_n : \mathcal{X}^n \to \{1, 2, \ldots, M_n\}$ is a mapping that assigns an index $b_n(\boldsymbol{x}) \in \{1, 2, \ldots, M_n\}$ to every source sequence $\boldsymbol{x} \in \mathcal{X}^n$. For a fixed bin assignment, determined by a given binning function b_n, the set of all source sequences with the same bin index as $\boldsymbol{x}$ is denoted by

$$\mathcal{B}_n(\boldsymbol{x}|b_n) \triangleq \{\hat{\boldsymbol{x}} \in \mathcal{X}^n : b_n(\hat{\boldsymbol{x}}) = b_n(\boldsymbol{x})\}.$$

Further on, we will analyze the error probability averaged over an ensemble of binning functions. To that end, we denote a random binning function by B_n, where b_n is a realization of B_n. We use Gallager's ensemble [12]: every source sequence is assigned a bin index uniformly at random; and bin assignments are pairwise independent across sequences.

Encoding: The generated source sequence $\boldsymbol{x}$ is encoded into $\phi_n(\boldsymbol{x}) = \big(t_n(\boldsymbol{x}), b_n(\boldsymbol{x})\big)$, where $t_n(\boldsymbol{x})$ is the type index and $b_n(\boldsymbol{x})$ is the bin index. The rate satisfies $\frac{1}{n} \log (J_n M_n) \leq R$.

Decoding: Upon receiving $\phi_n(\boldsymbol{x})$, the decoder knows that $\boldsymbol{x}$ is in the set $\mathcal{T}_n(P_{\boldsymbol{x}}) \cap \mathcal{B}_n(\boldsymbol{x}|b_n)$. Now suppose that the side information sequence is equal to $\boldsymbol{y}$. For every sequence $\hat{\boldsymbol{x}} \in \mathcal{T}_n(P_{\boldsymbol{x}}) \cap \mathcal{B}_n(\boldsymbol{x}|b_n)$, the decoder computes the conditional entropy $H(\hat{\boldsymbol{x}}|\boldsymbol{y})$ and produces a list $\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})$ of size

$$|\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})| = \min \left\{ \lfloor \mathrm{e}^{n\Delta} \rfloor , |\mathcal{T}_n(P_{\boldsymbol{x}}) \cap \mathcal{B}_n(\boldsymbol{x}|b_n)| \right\}$$

comprising source sequences with the lowest conditional entropy. The soft reconstruction is taken to be uniformly supported on $\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})$. It is clear that an excess distortion error occurs if $\mathcal{L}_n(\phi_n(\boldsymbol{x}), \boldsymbol{y})$ does not include the encoded source sequence $\boldsymbol{x}$. Moreover, by setting $\Delta = 0$, we recover the classical minimum entropy decoder.

Error probability: Let $\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, b_n)$ denote the set of all source sequences other than $\boldsymbol{x}$, but with the same type and bin as $\boldsymbol{x}$, and a conditional entropy smaller than or equal to that of $\boldsymbol{x}$ given $\boldsymbol{y}$. For an excess distortion error to occur, we must have $|\mathcal{E}(\boldsymbol{x}|\boldsymbol{y}, b_n)| \geq \lfloor \mathrm{e}^{n\Delta} \rfloor = \mathrm{e}^{n(\Delta - \delta_n)}$, where $\delta_n \geq 0$ goes to zero as n grows large. The excess distortion error probability, averaged over B_n, is hence bounded above by

$$\begin{aligned} &\mathbb{P}\left[|\mathcal{E}_n(\boldsymbol{X}|\boldsymbol{Y}, B_n)| \geq \mathrm{e}^{n(\Delta-\delta_n)} \right] \leq \\ &\sum_{\boldsymbol{x},\boldsymbol{y}} P_{XY}^n(\boldsymbol{x}, \boldsymbol{y}) \min \left\{ 1, \mathrm{e}^{-n(\Delta-\delta_n)} \mathbb{E}\big[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)| \big] \right\} \end{aligned} \tag{26}$$

which follows from Markov's inequality combined with the trivial upper bound of 1. We now take a small detour to bound $\mathbb{E}\left[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)| \right]$. To this end, define the set $\mathcal{E}_n'(\boldsymbol{x}|\boldsymbol{y}) \triangleq \big\{ \hat{\boldsymbol{x}} \in \mathcal{T}_n(P_{\boldsymbol{x}}) : \hat{\boldsymbol{x}} \neq \boldsymbol{x}, H\big(\hat{\boldsymbol{x}}|\boldsymbol{y}\big) \leq H\big(\boldsymbol{x}|\boldsymbol{y}\big) \big\}$, and observe that

$$\mathbb{E}\left[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)| \right] = \sum_{\hat{\boldsymbol{x}} \in \mathcal{E}_n'(\boldsymbol{x}|\boldsymbol{y})} \mathbb{P}\left[B_n(\hat{\boldsymbol{x}}) = B_n(\boldsymbol{x}) \right] = \frac{|\mathcal{E}'(\boldsymbol{x}|\boldsymbol{y})|}{M_n}$$

which follows from uniform pairwise independent bin assignment. Next, we note that

$$\begin{aligned} |\mathcal{E}_n'(\boldsymbol{x}|\boldsymbol{y})| &\leq \sum_{Q_{X|Y} \in \mathcal{P}_n(\mathcal{X}|\mathcal{Y}) : H(Q_{X|Y}|P_{\boldsymbol{y}}) \leq H(\boldsymbol{x}|\boldsymbol{y})} |\mathcal{T}_n(Q_{X|Y}|\boldsymbol{y})| \\ &\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \mathrm{e}^{nH(\boldsymbol{x}|\boldsymbol{y})}. \end{aligned} \tag{27}$$

From the above and $M_n = \lfloor (1+n)^{-|\mathcal{X}|} \mathrm{e}^{nR} \rfloor$, we obtain

$$\mathbb{E}\left[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)| \right] \leq \mathrm{e}^{nH(\boldsymbol{x}|\boldsymbol{y}) - nR + n\delta_n'} \tag{28}$$

from some $\delta_n' \geq 0$ which goes to zero as n grows large. Defining $\delta_n'' \triangleq \delta_n + \delta_n'$, and using (28), it follows that

$$\min \left\{ 1, \mathrm{e}^{-n(\Delta-\delta_n)} \mathbb{E}\big[|\mathcal{E}_n(\boldsymbol{x}|\boldsymbol{y}, B_n)| \big] \right\} \leq \mathrm{e}^{-n|R+\Delta-H(\boldsymbol{x}|\boldsymbol{y})|^+ + n\delta_n''}.$$

Plugging this back into (26), and invoking the usual random coding argument of the existence of a code, we obtain

$$\begin{aligned} &\varepsilon_{X|Y}^{\mathrm{WZ}}(n, R, \Delta) \leq \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \sum_{\boldsymbol{x} \in \mathcal{X}^n} P_{XY}^n(\boldsymbol{x}, \boldsymbol{y}) \mathrm{e}^{-n|R+\Delta-H(\boldsymbol{x}|\boldsymbol{y})|^+ + n\delta_n''} \\ &\leq \sum_{Q_{XY} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{Y})} \mathrm{e}^{-nD(Q_{XY} \| P_{XY})} \mathrm{e}^{-n|R+\Delta-H(Q_{X|Y}|Q_Y)|^+ + n\delta_n''} \\ &\leq (n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|} \mathrm{e}^{-n\tilde{F}_{X|Y}(R+\Delta) + n\delta_n''}. \end{aligned} \tag{29}$$

The result in Theorem 3 follows.

V. Concluding Remarks

In the high rate regime where $E_{X|Y}^{\mathrm{r}}(R, \Delta)$ and $E_{X|Y}^{\mathrm{sp}}(R, \Delta)$ diverge, it is possible to derive a tighter achievable exponent, which is a log-loss counterpart of the expurgated exponent in source coding [15]. We skip this due to lack of space. As an extension, it is of interest to derive error exponents and universal schemes for the multi-terminal settings in [1], [2].

References

[1] T. A. Courtade and R. D. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *Proc. IEEE Int. Symp. Inf. Theory*, 2011, pp. 2040–2044.

[2] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[3] Y. Y. Shkel and S. Verdú, "A single-shot approach to lossy source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 129–147, 2018.

[4] Y. Shkel, M. Raginsky, and S. Verdú, "Universal lossy compression under logarithmic loss," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017, pp. 1157–1161.

[5] Y. Shkel, M. Raginsky, and S. Verdu, "Universal compression, list decoding, and logarithmic loss," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 206–210.

[6] I. Csiszár and J. Körner, *Information theory: Coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[7] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 197–199, 1974.

[8] F. Jelinek, *Probabilistic information theory: Discrete and memoryless models*. McGraw-Hill, 1968.

[9] G. Longo and A. Sgarro, "The source coding theorem revisited: A combinatorial approach," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 544–548, 1979.

[10] C. Bunte and A. Lapidoth, "Source coding, lists, and Rényi entropy," in *Proc. IEEE Inf. Theory Workshop*, 2013, pp. 1–5.

[11] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Amsterdam: North-Holland Publ. Co, 1977, pp. 41–52.

[12] R. G. Gallager, "Source coding with side information and universal coding," M.I.T. LIDS, Tech. Rep., 1976.

[13] I. Csiszár and J. Körner, "Towards a general theory of source networks," *IEEE Trans. Inf. Theory*, vol. 26, no. 2, pp. 155–165, 1980.

[14] B. G. Kelly and A. B. Wagner, "Improved source coding exponents via Witsenhausen's rate," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5615–5633, 2011.

[15] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, 1981.

The Method of Types for the AWGN Channel: Correct-Decoding Exponent

Sergey Tridenski
Faculty of Engineering
Bar-Ilan University, Israel
Email: tridens@biu.ac.il

Anelia Somekh-Baruch
Faculty of Engineering
Bar-Ilan University, Israel
Email: somekha@biu.ac.il

***Abstract*—For the discrete-time AWGN channel with a power constraint, we derive a lower bound on the optimal correct-decoding exponent. The derivation uses the method of types with finite alphabets of sizes depending on the block length n and with the number of types sub-exponential in n.**[1]

I. Introduction

We study reliability of communication through the discrete-time additive white Gaussian noise (AWGN) channel with a power constraint imposed on blocks of its inputs. Consider the capacity formula of this channel, which was found by Shannon:

$$C \;=\; \tfrac{1}{2}\log_2(1 + s^2/\sigma^2), \tag{1}$$

where σ^2 is the channel noise variance and s^2 is the power constraint. This capacity corresponds to the maximum of the mutual information $I(p_X, w)$ over p_X, under the power constraint on p_X, where w stands for the channel transition probability density function (PDF) and p_X is the channel input PDF. Let us briefly recall the technicalities [1] of how the expression (1) is obtained from the mutual information:

$$\begin{aligned} & \max_{p_X:\, \mathbb{E}[X^2]\,\le\, s^2} I(p_X, w) \\ = & \max_{p_X:\, \mathbb{E}[X^2]\,\le\, s^2} \Big\{ D\big(w \,\|\, \widehat{p}_Y \mid p_X\big) \;-\; D\big(p_Y \,\|\, \widehat{p}_Y\big)\Big\} \\ = & \max_{p_X:\, \mathbb{E}[X^2]\,\le\, s^2} \bigg\{ \underbrace{\frac{\mathbb{E}[X^2] - s^2}{2\ln(2)(s^2+\sigma^2)} \;-\; D\big(p_Y \,\|\, \widehat{p}_Y\big)}_{\le\, 0} \bigg\} \\ & \qquad\qquad\qquad\qquad +\; \tfrac{1}{2}\log_2(1 + s^2/\sigma^2). \end{aligned} \tag{2}$$

Here $\widehat{p}_Y(y) \triangleq \frac{1}{\sqrt{2\pi(s^2+\sigma^2)}}\exp\Big(-\frac{y^2}{2(s^2+\sigma^2)}\Big)$ and $p_Y(y) \equiv \int_{\mathbb{R}} p_X(x)w(y\,|\,x)dx$, the operator $\mathbb{E}[\,\cdot\,]$ denotes the expectation, and D is the Kullback–Leibler divergence between two probability densities. In this paper we propose an explanation, similar to (2), for Oohama's converse bound on the optimal exponent in the block correct-decoding probability of the AWGN channel [2, Eq. 4]. This bound for the AWGN channel parallels the similar bound for the discrete memoryless channel, given by Arimoto [3], [4].

In the case of discrete memoryless channels, the mutual information enters into the expressions for correct-decoding and error exponents through the method of types [4], [5]. For the moment without any interpretation, let us rewrite the constant-composition correct-decoding exponent [4] with PDF's:

$$\min_{p_{Y|X}} \Big\{ D\big(p_{Y|X} \,\|\, w \mid \widehat{p}_X\big) \,+\, \big|\, R \,-\, I\big(\widehat{p}_X,\, p_{Y|X}\big)\,\big|^{+}\Big\}, \tag{3}$$

where $\widehat{p}_X$ denotes the Gaussian density with zero mean and variance s^2, which maximizes (2), $R > 0$ is the information rate, and $|\,t\,|^{+} \;\triangleq\; \max\{0, t\}$. When $\widehat{p}_X$ is Gaussian, the minimum (3) allows an explicit solution by the method of Lagrange multipliers. The minimizing solution $p_{Y|X}^{*}$ of (3) in this case is also Gaussian. Then it turns out that $p_{Y|X}^{*}$ and the y-marginal PDF of the product $\widehat{p}_X p_{Y|X}^{*}$ play the same roles in the derivation of the converse bound, as w and $\widehat{p}_Y$, respectively, in the maximization (2).

In this paper, in order to derive an expression similar to (3), we extend the method of types [1, Ch. 11.1], [6] to include countable alphabets consisting of real numbers, with the help of power constraints on types. The countable alphabets depend on the block length n, and the number of types satisfying the power constraints is kept sub-exponential in n. The types are empirical distributions of uniformly quantized real numbers in quantized versions of real channel input and output vectors of length n. We emphasize that the quantized versions serve only for classification of channel input and output vectors and not for the communication itself. The uniform quantization step may be different for the quantized versions of channel inputs and outputs, and in both cases it is chosen to be a decreasing function of n.

Similarly as (2), the proposed derivation demonstrates, that, in order to achieve the converse bound on the correct-decoding exponent, it is necessary for the types of the quantized versions of codewords to converge to the Gaussian distribution in characteristic function (CF), or, equivalently, in cumulative distribution function (CDF).

II. Notation

Countable alphabets consisting of real numbers are denoted by $\mathcal{X}_n$, $\mathcal{Y}_n$. The set of types with denominator n over $\mathcal{X}_n$ is denoted by $\mathcal{P}_n(\mathcal{X}_n)$. Capital '$P$' denotes probability mass functions, always corresponding to types. The type class of a type P_X is denoted by $T(P_X)$. Small 'p' denotes probability density functions. Thin letters x, y represent real values, while

[1] This work was supported by the Israel Science Foundation (ISF) grant #1579/23.

thick letters $\mathbf{x}$, $\mathbf{y}$ represent real vectors of length n. Capital letters X, Y represent random variables, a boldface letter $\mathbf{Y}$ represents a random vector of length n. Small w stands for a conditional PDF, and W_n stands for a discrete positive measure, *which does not necessarily add up to* 1. All information-theoretic quantities such as the mutual information $I(P_{XY})$, $I\big(P_X, P_{Y|X}\big)$, $I\big(P_X,\, p_{Y|X}\big)$, the Kullback-Leibler divergence $D\big(P_{Y|X}\,\|\,W_n\,|\,P_X\big)$, $D\big(p_{Y|X}\,\|\,w\,|\,P_X\big)$, and the information rate R are defined with respect to the logarithm to a base $b > 1$, denoted as $\log_b(\cdot)$. The natural logarithm is denoted as $\ln$. Logical "or" and "and" are represented by the symbols $\vee$ and $\wedge$, respectively.

III. Communication system

We consider communication over the time-discrete additive white Gaussian noise channel with real channel inputs $x \in \mathbb{R}$ and channel outputs $y \in \mathbb{R}$ and a transition probability density

$$w(y\,|\,x) \;\triangleq\; \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y\,-\,x)^2}{2\sigma^2}}.$$

Communication is performed by blocks of n channel inputs. Let $R > 0$ denote a nominal information rate. Each block is used for transmission of one out of M messages, where $M = M(n,R) \triangleq \lfloor b^{nR}\rfloor$, for some logarithm base $b > 1$. The encoder is a deterministic function $f : \{1, 2, \ldots, M\} \to \mathbb{R}^n$, which converts a message into a transmitted block, such that

$$\begin{aligned} f(m) \;=\; \mathbf{x}(m) \;=\; \big(x_1(m),\, x_2(m), \ldots,\, x_n(m)\big), \\ m = 1, 2, \ldots, M, \end{aligned}$$

where $x_k(m) \in \mathbb{R}$, for all $k = 1, 2, \ldots, n$. The set of all the codewords $\mathbf{x}(m)$, $m = 1, 2, \ldots, M$, constitutes a codebook $\mathcal{C}$. Each codeword $\mathbf{x}(m)$ in $\mathcal{C}$ satisfies the power constraint:

$$\frac{1}{n}\sum_{k=1}^{n} x_k^2(m) \;\le\; s^2, \qquad m = 1, 2, \ldots, M. \tag{4}$$

The decoder is another deterministic function $g : \mathbb{R}^n \to \{0, 1, 2, \ldots, M\}$, which converts the received block of n channel outputs $\mathbf{y} \in \mathbb{R}^n$ into an estimated message, or, possibly, to a special error symbol '0':

$$g(\mathbf{y}) \;=\; \left\{\begin{array}{ll} 0, & \mathbf{y} \in \bigcap_{m=1}^{M}\mathcal{D}_m^c, \\ m, & \mathbf{y} \in \mathcal{D}_m, \;\; m \in \{1, 2, \ldots, M\}, \end{array}\right. \tag{5}$$

where each set $\mathcal{D}_m \subseteq \mathbb{R}^n$ is either an open region or the empty set, and the regions are disjoint; i.e., $\mathcal{D}_m \cap \mathcal{D}_{m'} = \varnothing$ for $m \neq m'$. Observe that the maximum-likelihood decoder with open decision regions $\mathcal{D}_m^*$, defined for $m = 1, 2, \ldots, M$ as

$$\mathcal{D}_m^* \;=\; \mathbb{R}^n \setminus \bigcup_{\substack{m':\;(m' < m)\,\vee \\ \left(m' > m \;\wedge\; \mathbf{x}(m') \neq \mathbf{x}(m)\right)}} \Big\{\mathbf{y} : \|\mathbf{y}-\mathbf{x}(m')\| \;\le\; \|\mathbf{y}-\mathbf{x}(m)\|\Big\},$$

is a special case of (5). Note that the formal definition of $\mathcal{D}_m^*$ includes the undesirable possibility of $\mathbf{x}(m') = \mathbf{x}(m)$ for $m' \neq m$.

IV. Definitions

For each n, we define two discrete countable alphabets $\mathcal{X}_n$ and $\mathcal{Y}_n$ as one-dimensional lattices:

$$\begin{aligned} &\alpha, \beta \in (0, 1), \quad \alpha + \beta < 1, \\ &\Delta_{\alpha,\,n} \;\triangleq\; 1/n^{\alpha}, \quad \Delta_{\beta,\,n} \;\triangleq\; 1/n^{\beta}, \end{aligned} \tag{6}$$

$$\mathcal{X}_n \;\triangleq\; \bigcup_{i\,\in\,\mathbb{Z}}\{i\Delta_{\alpha,\,n}\}, \qquad \mathcal{Y}_n \;\triangleq\; \bigcup_{i\,\in\,\mathbb{Z}}\{i\Delta_{\beta,\,n}\}. \tag{7}$$

For each n, we define also a discrete positive measure (not necessarily a distribution), which will approximate the channel w:

$$W_n(y\,|\,x) \;\triangleq\; w(y\,|\,x)\cdot\Delta_{\beta,\,n}, \quad \forall x \in \mathcal{X}_n,\; \forall y \in \mathcal{Y}_n. \tag{8}$$

Denoting by $C^0(A)$ a class of functions $f : \mathbb{R} \to \mathbb{R}_{\ge 0}$ continuous on an open subset $A \subseteq \mathbb{R}$, for each n we define

$$\begin{aligned} \mathcal{F}_n \;\triangleq\; \bigg\{ f : \mathbb{R} \to \mathbb{R}_{\ge 0} \;\;\bigg|\;\; & f \in C^0\big(\mathbb{R} \setminus \{\mathcal{Y}_n + \Delta_{\beta,\,n}/2\}\big), \\ & \sup_{y\,\in\,\mathbb{R}} f(y) < +\infty, \;\; \int_{\mathbb{R}} f(y)dy = 1 \bigg\}. \end{aligned} \tag{9}$$

With a parameter $\rho \in (-1, +\infty)$, we define the following Gaussian probability density functions [7], [8]:

$$p_{Y|X}^{(\rho)}(y\,|\,x) \;\triangleq\; \tfrac{1}{\sigma_{Y|X}(\rho)\sqrt{2\pi}}\exp\Big\{-\,\tfrac{(y\,-\,k_\rho\,\cdot\,x)^2}{2\sigma_{Y|X}^2(\rho)}\Big\}, \tag{10}$$

$$k_\rho \;\triangleq\; \tfrac{\mathrm{SNR}\,-\,\rho\,-\,1\,+\,\sqrt{(\mathrm{SNR}\,-\,\rho\,-\,1)^2\,+\,4\,\cdot\,\mathrm{SNR}}}{2\,\cdot\,\mathrm{SNR}}, \;\; \mathrm{SNR} \triangleq \tfrac{s^2}{\sigma^2}, \tag{11}$$

$$\sigma_{Y|X}^2(\rho) \;\triangleq\; (1+\rho)k_\rho\,\sigma^2, \tag{12}$$

$$\widehat{p}_Y^{(\rho)}(y) \;\triangleq\; \tfrac{1}{\sigma_Y(\rho)\sqrt{2\pi}}\exp\Big\{-\,\tfrac{y^2}{2\sigma_Y^2(\rho)}\Big\}, \tag{13}$$

$$\sigma_Y^2(\rho) \;\triangleq\; \sigma^2 \,+\, k_\rho\, s^2, \tag{14}$$

$$\widehat{p}_X(x) \;\triangleq\; \tfrac{1}{s\sqrt{2\pi}}\exp\Big\{-\,\tfrac{x^2}{2s^2}\Big\}. \tag{15}$$

The first property of the following lemma shows that $\widehat{p}_Y^{(\rho)}$ is the y-marginal PDF of the product $\widehat{p}_X p_{Y|X}^{(\rho)}$.

Lemma 1 (Properties of (10)-(15)):
The following properties hold:

$$\sigma_Y^2(\rho) \;=\; \sigma_{Y|X}^2(\rho) \,+\, k_\rho^2\, s^2, \tag{16}$$

$$\frac{1+\rho}{\sigma_{Y|X}^2(\rho)} \;=\; \frac{\rho}{\sigma_Y^2(\rho)} \,+\, \frac{1}{\sigma^2}, \tag{17}$$

$$\sigma_{Y|X}^2(\rho) \;=\; \sigma^2 \,+\, k_\rho(1-k_\rho)s^2, \tag{18}$$

$$1 \;\ge\; k_\rho \;>\; 0, \qquad \rho \ge 0, \tag{19}$$

$$\tfrac{1}{2}\big[1 \,+\, \sqrt{1 \,+\, 4\sigma^2 s^{-2}}\,\big] \;\ge\; k_\rho \;\ge\; 1, \qquad -1 \le \rho \le 0, \tag{20}$$

and for any two jointly distributed random variables (X, Y), *such that* $\mathbb{E}\big[X^2\big] = \sigma_X^2 \;\le\; s^2 + \epsilon,\;\; \epsilon > 0$, *and* $Y\,|\,X = x \;\sim\; \mathcal{N}\big(k_\rho\, x,\; \sigma_{Y|X}^2(\rho)\big)$, *it holds that*

$$\begin{aligned} \mathbb{E}\big[(Y-X)^2\big] \;&=\; \sigma^2 \,+\, (1-k_\rho)s^2 \,+\, (1-k_\rho)^2(\sigma_X^2 - s^2) \\ &\le\; \sigma^2 \,+\, \epsilon\sigma^2 s^{-2}, \qquad -1 < \rho \le 0. \end{aligned} \tag{21}$$

Here (13) corresponds to [7, Eq. 63], definition (14) combined with (11) corresponds to [7, Eq. 64], relationships (12), (17), and (10) can be found in [7, Eq. 65], while (10), (11), (18) correspond respectively to [8, Eq. 327, 302, 328].

Proof of Lemma 1: The first property (16) can be verified using (14), (12), (11). Then (17) can be obtained from (16), (14), (12). Property (18) follows by (16) and (14). It can be verified from (11) that k_ρ is a positive monotonically decreasing function of ρ, such that $k_0 = 1$. Then we get (19) and (20). The equality of (21) can be obtained using (18). Then, the inequality of (21) can be verified using (20). □

The following expression will describe our result for the correct-decoding exponent:

$$E_c(R) \triangleq \sup_{-1 < \rho \leq 0} \Big\{ D\big(p_{Y|X}^{(\rho)} \,\|\, w \mid \widehat{p}_X\big) + \rho\big[I\big(\widehat{p}_X, p_{Y|X}^{(\rho)}\big) - R\big]\Big\}. \tag{22}$$

The following identity can be obtained using (10), (12), (13), (15), (17):

$$D\big(p_{Y|X}^{(\rho)} \,\|\, w \mid \widehat{p}_X\big) + \rho I\big(\widehat{p}_X, p_{Y|X}^{(\rho)}\big) \equiv c_0(\rho) + c_1(\rho)s^2,$$
$$c_0(\rho) \triangleq \frac{1}{\ln b}\ln\left(\frac{\sigma \cdot \sigma_Y^{\rho}(\rho)}{\sigma_{Y|X}^{1+\rho}(\rho)}\right), \quad c_1(\rho) \triangleq \frac{1 - k_\rho}{2\sigma^2 \ln b}. \tag{23}$$

It can be verified that with $\rho \geq 0$ the expression inside the supremum of (22) is equivalent to the expression for the Gaussian random-coding error exponent of Gallager prior to the maximization over ρ [9, Eq. 7.4.24 with Eq. 7.4.28]. Therefore, if the supremum is taken over $\rho \geq 0$, the expression (22) coincides with Shannon's sphere-packing converse bound on the error exponent [10, Eq. 3, 4, 11] in the limit of a large block length.

V. Converse bounds in terms of types

For a vector of n channel inputs $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ let $Q_\alpha(\mathbf{x}) = \mathbf{x}^q = (x_1^q, x_2^q, \dots, x_n^q) \in \mathcal{X}_n^n \equiv (\mathcal{X}_n)^n$ be its quantized version, with components

$$x_k^q = Q_\alpha(x_k) \triangleq \Delta_{\alpha,n} \cdot \lfloor x_k/\Delta_{\alpha,n} + 1/2 \rfloor, k = 1, \dots, n. \tag{24}$$

For a vector of n channel outputs $\mathbf{y} = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$, similarly, let $Q_\beta(\mathbf{y}) \triangleq \mathbf{y}^q = (y_1^q, y_2^q, \dots, y_n^q) \in \mathcal{Y}_n^n$ be its quantized version, with $y_k^q = Q_\beta(y_k)$ for all $k = 1, \dots, n$. We derive a converse bound on the correct-decoding exponent in terms of types in two steps. At first, we derive the bound for a codebook with all the *quantized versions* of its codewords belonging to the same type. We start with the conditional probability of correct decoding in such a codebook, given the joint type of the transmitted codeword and the received block.

Lemma 2 (Conditional probability of correct decoding):

Let $P_{XY} \in \mathcal{P}_n(\mathcal{X}_n \times \mathcal{Y}_n)$ be a joint type, such that $\mathbb{E}_{P_X}\big[X^2\big] \leq c_X$, $\mathbb{E}_{P_Y}\big[Y^2\big] \leq c_Y$, and $\mathbb{E}_{P_{XY}}\big[(Y - X)^2\big] \leq c_{XY}$, and let $\mathcal{C}$ be a codebook, such that the quantized versions (24) of its codewords $\mathbf{x}(m)$, $m = 1, 2, \dots, M(n, R)$, are all of the type P_X, that is:

$$\mathbf{x}^q(m) = Q_\alpha(\mathbf{x}(m)) \in T(P_X), \quad \forall m.$$

Let $J \sim$ Unif$\big(\{1, 2, \dots, M\}\big)$ be a random variable, independent of the channel, and let $\mathbf{x}(J) \rightarrow \mathbf{Y}$ be the random channel-input and channel-output vectors, respectively. Let $\mathbf{Y}^q = Q_\beta(\mathbf{Y}) \in \mathcal{Y}_n^n$. Then

$$\Pr\Big\{g(\mathbf{Y}) = J \;\Big|\; \big(\mathbf{x}^q(J), \mathbf{Y}^q\big) \in T(P_{XY})\Big\} \leq b^{-n\left(\widetilde{R} - I(P_{XY}) + o(1)\right)},$$

where $\widetilde{R} = \frac{1}{n}\log_b M(n, R)$, and $o(1) \rightarrow 0$, as $n \rightarrow \infty$, depending only on α, β, c_X, c_Y, c_{XY}, and σ^2.

The proof is given in [11, Lemma 10].

Then, the unconditional converse bound, for a codebook with all the quantized versions of its codewords belonging to the same type, is given by the following lemma:

Lemma 3 (Correct-decoding exponent of mono-composition codebooks):

Let $P_X \in \mathcal{P}_n(\mathcal{X}_n)$ be a type, such that $\mathbb{E}_{P_X}\big[X^2\big] \leq c_X$, and let $\mathcal{C}$ be a codebook, such that the quantized versions (24) of its codewords $\mathbf{x}(m)$, $m = 1, 2, \dots, M(n, R)$, are all of the type P_X, that is:

$$\mathbf{x}^q(m) = Q_\alpha(\mathbf{x}(m)) \in T(P_X), \quad \forall m.$$

Let $J \sim$ Unif$\big(\{1, 2, \dots, M\}\big)$ be a random variable, independent of the channel, and let $\mathbf{x}(J) \rightarrow \mathbf{Y}$ be the random channel-input and channel-output vectors, respectively. Then for any $\epsilon > 0$ and $\widetilde{\sigma}^2 \geq \sigma^2$ there exists $n_0 = n_0(\alpha, \beta, \widetilde{\sigma}^2, \epsilon) \in \mathbb{N}$, such that for any $n > n_0$

$$-\frac{1}{n}\log_b \Pr\big\{g(\mathbf{Y}) = J\big\} \geq \min\big\{E_n(P_X, R, \widetilde{\sigma}, \epsilon), E(\widetilde{\sigma})\big\} + o(1), \tag{25}$$

where

$$E(\widetilde{\sigma}) \triangleq \frac{1}{2\ln b}\big[\widetilde{\sigma}^2/\sigma^2 - 1 - \ln\big(\widetilde{\sigma}^2/\sigma^2\big)\big], \tag{26}$$
$$E_n(P_X, R, \widetilde{\sigma}, \epsilon) \triangleq \min_{\substack{P_{Y|X}: \\ P_{XY} \in \mathcal{P}_n(\mathcal{X}_n \times \mathcal{Y}_n), \\ \mathbb{E}[(Y-X)^2] \leq \widetilde{\sigma}^2 + \epsilon}} \Big\{D\big(P_{Y|X} \,\|\, W_n \,|\, P_X\big) + \big|R - I\big(P_X, P_{Y|X}\big)\big|^+\Big\}, \tag{27}$$

and where $|\,t\,|^+ \triangleq \max\{0, t\}$, and $o(1) \rightarrow 0$, as $n \rightarrow \infty$, depending only on α, β, c_X, $\widetilde{\sigma}^2 + \epsilon$, and σ^2.

The proof is given in [11, Lemma 16].

Next, similarly as in [4, Lemma 5], minimization over types P_X extends the bound of Lemma 3 to arbitrary codebooks:

Lemma 4 (Correct-decoding exponent):

Let $J \sim$ Unif$\big(\{1, 2, \dots, M\}\big)$ be a random variable, independent of the channel, and let $\mathbf{x}(J) \rightarrow \mathbf{Y}$ be the random channel-input and channel-output vectors, respectively. Then for any $\widetilde{\epsilon}, \epsilon > 0$ and $\widetilde{\sigma}^2 \geq \sigma^2$ there exists $n_0 = n_0(\alpha, \beta, s^2, \widetilde{\sigma}^2, \widetilde{\epsilon}, \epsilon) \in \mathbb{N}$, such that for any $n > n_0$

$$-\frac{1}{n}\log_b \Pr\big\{g(\mathbf{Y}) = J\big\} \geq \min\big\{E_n(R, \widetilde{\sigma}, \widetilde{\epsilon}, \epsilon), E(\widetilde{\sigma})\big\} + o(1), \tag{28}$$

where

$$E_n(R,\, \widetilde{\sigma},\, \widetilde{\epsilon},\, \epsilon) \;\triangleq\; \min_{\substack{P_X:\\ P_X \in \mathcal{P}_n(\mathcal{X}_n),\\ \mathbb{E}[X^2] \le s^2 + \epsilon}} E_n(P_X,\, R,\, \widetilde{\sigma},\, \widetilde{\epsilon}), \tag{29}$$

and where $E(\widetilde{\sigma})$ and $E_n(P_X,\, R,\, \widetilde{\sigma},\, \widetilde{\epsilon})$ are as defined in (26) and (27), respectively, and $o(1) \to 0$, as $n \to \infty$, depending only on the parameters α, β, $s^2 + \epsilon$, $\widetilde{\sigma}^2 + \widetilde{\epsilon}$, and σ^2.
The proof is given in [11, Lemma 17].

The final lemma relates between the minimum over types (27) and an infimum over PDF's of the class defined by (9):

Lemma 5 (Type to PDF):
For any c_{XY} and $\epsilon > 0$ there exists $n_0 = n_0(\beta,\, c_{XY},\, \epsilon) \in \mathbb{N}$, such that for any $n > n_0$ and for any type $P_X \in \mathcal{P}_n(\mathcal{X}_n)$:

$$\min_{\substack{P_{Y|X}:\\ P_{XY} \in \mathcal{P}_n(\mathcal{X}_n \times \mathcal{Y}_n),\\ \mathbb{E}_{P_{XY}}[(Y-X)^2] \le c_{XY}}} \Big\{D\big(P_{Y|X} \,\|\, W_n \,|\, P_X\big) \,+\, \big|\, R \,-\, I\big(P_X, P_{Y|X}\big)\,\big|^+\Big\}$$
$$\ge \inf_{\substack{p_{Y|X}:\\ p_{Y|X}(\,\cdot\,|\,x) \in \mathcal{F}_n,\, \forall x,\\ \mathbb{E}_{P_X p_{Y|X}}[(Y-X)^2] \le c_{XY} + \epsilon}} \Big\{D\big(p_{Y|X} \,\|\, w \,|\, P_X\big) \,+\, \big|\, R \,-\, I\big(P_X,\, p_{Y|X}\big)\,\big|^+\Big\} \,+\, o(1),$$

where $o(1) \to 0$, as $n \to \infty$, and depends only on the parameters β and c_{XY}.
The proof is given in [11, Lemma 20].

VI. Main result

In this section we prove the main theorem and conclude with a proposition, which gives an alternative representation for (22). The proof of Theorem 1 relies on Lemmas 4 and 5.

Theorem 1 (Correct-decoding exponent):
Let $J \sim$ Unif$\big(\{1, 2, \ldots, M\}\big)$ be a random variable, independent of the channel, and let $\mathbf{x}(J) \to \mathbf{Y}$ be the random channel-input and channel-output vectors, respectively. Then

$$\liminf_{n\,\to\,\infty}\; \inf_{\mathcal{C}}\; \inf_{g}\; \left\{-\,\frac{1}{n}\log_b \Pr\big\{g(\mathbf{Y}) = J\big\}\right\} \;\ge\; E_c(R),$$

where $E_c(R)$ is defined by (22), decoder functions g are defined by (5), and codebooks $\mathcal{C}$ satisfy (4).

Proof: Starting from Lemma 4, for each $R > 0$ we can choose a different parameter $\widetilde{\sigma} = \widetilde{\sigma}(R) \ge \sigma$, such that there is equality $E(\widetilde{\sigma}(R)) = E_c(R)$ between (26) and (22). Then by (28) we obtain

$$\liminf_{n\,\to\,\infty}\; \inf_{\mathcal{C}}\; \inf_{g}\; \left\{-\,\frac{1}{n}\log_b \Pr\big\{g(\mathbf{Y}) = J\big\}\right\} \;\ge$$
$$\min\Big\{\liminf_{n\,\to\,\infty} E_n(R,\, \widetilde{\sigma}(R),\, \widetilde{\epsilon},\, \epsilon),\;\; E_c(R)\Big\}.$$

With the choice $2\widetilde{\epsilon} = \epsilon\sigma^2 s^{-2}$, the first term in the minimum can be lower-bounded as follows:

$$\liminf_{n\,\to\,\infty} \min_{\substack{P_X:\\ P_X \in \mathcal{P}_n(\mathcal{X}_n),\\ \mathbb{E}[X^2] \le s^2 + \epsilon}} \; \min_{\substack{P_{Y|X}:\\ P_{XY} \in \mathcal{P}_n(\mathcal{X}_n \times \mathcal{Y}_n),\\ \mathbb{E}[(Y-X)^2] \le \widetilde{\sigma}^2(R) + \widetilde{\epsilon}}}$$
$$\Big\{D\big(P_{Y|X} \,\|\, W_n \,|\, P_X\big) \,+\, \big|\, R \,-\, I\big(P_X, P_{Y|X}\big)\,\big|^+\Big\}$$
$$\overset{a}{\ge} \liminf_{n\,\to\,\infty} \min_{\substack{P_X:\\ P_X \in \mathcal{P}_n(\mathcal{X}_n),\\ \mathbb{E}[X^2] \le s^2 + \epsilon}} \; \inf_{\substack{p_{Y|X}:\\ p_{Y|X}(\,\cdot\,|\,x) \in \mathcal{F}_n,\, \forall x,\\ \mathbb{E}[(Y-X)^2] \le \widetilde{\sigma}^2(R) + 2\widetilde{\epsilon}}}$$
$$\Big\{D\big(p_{Y|X} \,\|\, w \,|\, P_X\big) \,+\, \big|\, R \,-\, I\big(P_X,\, p_{Y|X}\big)\,\big|^+\Big\}$$
$$\overset{b}{\ge} \liminf_{n\,\to\,\infty} \min_{\substack{P_X:\\ P_X \in \mathcal{P}_n(\mathcal{X}_n),\\ \mathbb{E}[X^2] \le s^2 + \epsilon}} \; \inf_{\substack{p_{Y|X}:\\ p_{Y|X}(\,\cdot\,|\,x) \in \mathcal{F}_n,\, \forall x,\\ \mathbb{E}[(Y-X)^2] \le \widetilde{\sigma}^2(R) + 2\widetilde{\epsilon}}}$$
$$\Big\{D\big(p_{Y|X} \,\|\, w \,|\, P_X\big) \,-\, \rho\Big[R \,-\, D\big(p_{Y|X} \,\|\, \widehat{p}_Y^{(\rho)} \,|\, P_X\big)\Big]\Big\}$$
$$\overset{c}{\equiv} \liminf_{n\,\to\,\infty} \min_{\substack{P_X:\\ P_X \in \mathcal{P}_n(\mathcal{X}_n),\\ \mathbb{E}[X^2] \le s^2 + \epsilon}} \; \inf_{\substack{p_{Y|X}:\\ p_{Y|X}(\,\cdot\,|\,x) \in \mathcal{F}_n,\, \forall x,\\ \mathbb{E}[(Y-X)^2] \le \widetilde{\sigma}^2(R) + 2\widetilde{\epsilon}}} \Big\{c_0(\rho)$$
$$+\, c_1(\rho)\,\mathbb{E}\big[X^2\big] \,-\, \rho R \,+\, (1 + \rho)D\big(p_{Y|X} \,\|\, p_{Y|X}^{(\rho)} \,|\, P_X\big)\Big\}$$
$$\overset{d}{=} \liminf_{n\,\to\,\infty} \min_{\substack{P_X:\\ P_X \in \mathcal{P}_n(\mathcal{X}_n),\\ \mathbb{E}[X^2] \le s^2 + \epsilon}} \Big\{c_0(\rho) \,+\, c_1(\rho)\,\mathbb{E}\big[X^2\big] \,-\, \rho R\Big\} \tag{30}$$
$$\overset{e}{\ge} \qquad c_0(\rho) \,+\, c_1(\rho)(s^2 + \epsilon) \,-\, \rho R, \tag{31}$$

where:

(a) follows by Lemma 5 with $c_{XY} = \widetilde{\sigma}^2(R) + \widetilde{\epsilon}$.

(b) holds for $\rho \in (-1, 0\,]$, because $\big|\, R \,-\, I\big(P_X,\, p_{Y|X}\big)\,\big|^+ \ge -\rho\big[\, R \,-\, I\big(P_X,\, p_{Y|X}\big)\big]$ for any such ρ, and because $I\big(P_X,\, p_{Y|X}\big) \;\le\; D\big(p_{Y|X} \,\|\, \widehat{p}_Y^{(\rho)} \,|\, P_X\big)$, where $\widehat{p}_Y^{(\rho)}$ is the Gaussian PDF defined in (13).

(c) holds as an identity inside the infimum by the definitions (10), (13), (23), and properties (12), (17).

(d) holds if $2\widetilde{\epsilon} \ge \epsilon\sigma^2 s^{-2}$ and $\rho \in (-1, 0\,]$, because then by (21) and (9) the function $p_{Y|X}^{(\rho)}$ satisfies the conditions under the infimum and achieves the infimum.

(e) follows by the condition under the minimum of (30) since $c_1(\rho) \le 0$ for $\rho \in (-1, 0\,]$.

In conclusion, since (31) is the lower bound for any $\rho \in (-1, 0\,]$ and $2\widetilde{\epsilon} \ge \epsilon\sigma^2 s^{-2}$, we obtain

$$\liminf_{n\,\to\,\infty} E_n\big(R,\, \widetilde{\sigma}(R),\, \epsilon\sigma^2 s^{-2}/2,\, \epsilon\big) \;\ge$$
$$\sup_{-1\,<\,\rho\,\le\,0} \big\{c_0(\rho) \,+\, c_1(\rho)(s^2 + \epsilon) \,-\, \rho R\big\} \;\overset{\epsilon\,\to\,0}{\longrightarrow}\; E_c(R). \;\; \square$$

Remark: Observe that the inequality (b) in the proof of Theorem 1 cannot be met with equality unless $D\big(p_Y^{(\rho)} \,\|\, \widehat{p}_Y^{(\rho)}\big) \to 0$, where $p_Y^{(\rho)}$ is the y-marginal PDF of $P_X p_{Y|X}^{(\rho)}$. Accordingly, since $\widehat{p}_Y^{(\rho)}$ is Gaussian, while $p_Y^{(\rho)}$ is a convolution of P_X with the Gaussian PDF $p_{Y|X}^{(\rho)}$, the type P_X must converge to the

Gaussian distribution in CF[2] and CDF in order to achieve the exponent of Theorem 1. In the proof, the type P_X represents the histograms of codewords, i.e., the empirical distributions of their quantized versions.

Proposition 1 (Parametric representation of E_c): For every $R \geq I(\widehat{p}_X, w)$ there exists a unique $\rho \in (-1, 0\,]$, such that

$$R = I\big(\widehat{p}_X, p^{(\rho)}_{Y|X}\big), \quad E_c(R) = D\big(p^{(\rho)}_{Y|X} \,\|\, w \mid \widehat{p}_X\big). \tag{32}$$

The parametric representation of (32) is analogous to [9, Eq. 7.4.30, Eq. 7.4.31], it is equivalent to [2, Eq. 22] and appears in [12, Eq. 25, 26]. Here we present an alternative proof of Proposition 1 in the vein of the proof of Theorem 1.

Proof: Let us denote $R_\beta \triangleq I\big(\widehat{p}_X, p^{(\beta)}_{Y|X}\big)$. Then for $\beta \in (-1, 0\,]$ we can write a sandwich proof:

$$\inf_{\substack{p_{Y|X}:\\ \widehat{p}_X p_{Y|X} \in \mathcal{N}}} \Big\{ D\big(p_{Y|X} \,\|\, w \mid \widehat{p}_X\big) + \big|\, R_\beta - I\big(\widehat{p}_X, p_{Y|X}\big)\,\big|^+ \Big\} \tag{33}$$

$$\overset{a}{\geq} \sup_{-1<\rho\leq 0} \inf_{\substack{p_{Y|X}:\\ \widehat{p}_X p_{Y|X} \in \mathcal{N}}} \Big\{ D\big(p_{Y|X} \,\|\, w \mid \widehat{p}_X\big) - \rho\big[R_\beta - D\big(p_{Y|X} \,\|\, \widehat{p}^{(\rho)}_Y \mid \widehat{p}_X\big)\big] \Big\}$$

$$\overset{b}{\equiv} \sup_{-1<\rho\leq 0} \inf_{\substack{p_{Y|X}:\\ \widehat{p}_X p_{Y|X} \in \mathcal{N}}} \Big\{ D\big(p^{(\rho)}_{Y|X} \,\|\, w \mid \widehat{p}_X\big) - \rho\big[R_\beta - R_\rho\big] + (1+\rho) D\big(p_{Y|X} \,\|\, p^{(\rho)}_{Y|X} \mid \widehat{p}_X\big) \Big\}$$

$$\overset{c}{=} \sup_{-1<\rho\leq 0} \Big\{ D\big(p^{(\rho)}_{Y|X} \,\|\, w \mid \widehat{p}_X\big) - \rho\big[R_\beta - R_\rho\big] \Big\}$$

$$\equiv E_c(R_\beta) \overset{d}{\geq} D\big(p^{(\beta)}_{Y|X} \,\|\, w \mid \widehat{p}_X\big), \tag{34}$$

where $\mathcal{N}$ denotes the set of all bivariate non-degenerate Gaussian PDF's. Here (a) follows similarly to the inequality (b) in Theorem 1; (b) is an identity; (c) follows because $\widehat{p}_X p^{(\rho)}_{Y|X}$ is Gaussian and $p^{(\rho)}_{Y|X}$ achieves the infimum; (d) is a lower bound on the supremum at $\rho = \beta$. Finally, since the RHS of (34) is further lower-bounded by the infimum (33), we conclude that $E_c(R_\beta) = D\big(p^{(\beta)}_{Y|X} \,\|\, w \mid \widehat{p}_X\big)$.

From $I\big(\widehat{p}_X, p^{(\rho)}_{Y|X}\big) = \frac{1}{2}\log_b\big(\sigma^2_Y(\rho)/\sigma^2_{Y|X}(\rho)\big)$ using (14) and (18) we obtain $\frac{dR_\rho}{d\rho} = \frac{dR_\rho}{dk_\rho} \cdot \frac{dk_\rho}{d\rho} < 0$. Hence for every $R \geq I(\widehat{p}_X, w)$ the parameter $\rho(R) \in (-1, 0\,]$ is unique. □

References

[1] T. M. Cover and J. A. Thomas, *"Elements of Information Theory,"* John Wiley & Sons, 2006.

[2] Y. Oohama, "The Optimal Exponent Function for the Additive White Gaussian Noise Channel at Rates above the Capacity," in *IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun 2017.

[3] S. Arimoto, "On the Converse to the Coding Theorem for Discrete Memoryless Channels," *IEEE Trans. on Information Theory*, vol. 19, no. 3, pp. 357–359, May 1973.

[4] G. Dueck and J. Körner, "Reliability Function of a Discrete Memoryless Channel at Rates above Capacity," *IEEE Trans. on Information Theory*, vol. 25, no. 1, pp. 82–85, Jan 1979.

[5] I. Csiszár and J. Körner, *"Information Theory: Coding Theorems for Discrete Memoryless Systems,"* Academic Press, 1981.

[6] I. Csiszár, "The Method of Types," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2505–2523, Oct 1998.

[7] B. Nakiboğlu, "The Sphere Packing Bound for Memoryless Channels," *Problems of Information Transmission*, vol. 56, no. 3, pp. 201–244, Jul 2020.

[8] S. Verdú, "Error Exponents and α-Mutual Information," *Entropy*, 23, 199, 2021.

[9] R. G. Gallager, *"Information Theory and Reliable Communication,"* John Wiley & Sons, 1968.

[10] C. E. Shannon, "Probability of Error for Optimal Codes in a Gaussian Channel," *The Bell System Technical Journal*, vol. 38, no. 3, pp. 611–656, May 1959.

[11] S. Tridenski, A. Somekh-Baruch, "The Method of Types for the AWGN Channel," arXiv:2307.13322, Jul 2023.

[12] H.-C. Cheng, B. Nakiboğlu, "Refined Strong Converse for the Constant Composition Codes," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 2149–2154, Jun 2020.

[2] It follows because the expression for the characteristic function of the zero-mean Gaussian distribution also has a Gaussian form.

The shifted composition rule

Jason M. Altschuler
University of Pennsylvania
email: alts@upenn.edu

Sinho Chewi
Institute for Advanced Study
email: schewi@ias.edu

This is an extended abstract for an invited talk based on [AC23; AC24].

I. Shifted composition rule

We formulate a new technique for bounding information-theoretic divergences. For KL divergence, this Shifted Chain Rule (SCR) states

$$\mathsf{KL}(\boldsymbol{\mu}^Y \,\|\, \boldsymbol{\nu}^Y) \le \mathsf{KL}(\boldsymbol{\mu}^{X'} \,\|\, \boldsymbol{\nu}^X) + \mathbb{E}\,\mathsf{KL}(\boldsymbol{\mu}^{Y|X=x} \,\|\, \boldsymbol{\nu}^{Y|X=x'})$$

where $\boldsymbol{\mu}$ is a joint distribution on X, X', Y; $\boldsymbol{\nu}$ is a joint distribution on X, Y; and the expectation is over any coupling (x, x') of $\boldsymbol{\mu}^X$ and $\boldsymbol{\mu}^{X'}$. By taking $X = X'$, the SCR generalizes the standard KL chain rule which (combined with data-processing) gives the bound

$$\begin{aligned}\mathsf{KL}(\boldsymbol{\mu}^Y \,\|\, \boldsymbol{\nu}^Y) &\le \mathsf{KL}(\boldsymbol{\mu}^{X,Y} \,\|\, \boldsymbol{\nu}^{X,Y}) \\ &= \mathsf{KL}(\boldsymbol{\mu}^X \,\|\, \boldsymbol{\nu}^X) + \mathbb{E}\,\mathsf{KL}(\boldsymbol{\mu}^{Y|X=x} \,\|\, \boldsymbol{\nu}^{Y|X=x})\,.\end{aligned}$$

The key advantage of the SCR is the additional flexibility in X', which intuitively enables modifying the "history" of the process $X \mapsto Y$ to $X' \mapsto Y$ (first term) at a price given by how different X and X' are (second term). This enables addressing applications where the standard chain rule would not suffice, such as situations where $\mathsf{KL}(\boldsymbol{\mu}^X \,\|\, \boldsymbol{\nu}^X)$ is large or even infinite (e.g., $\boldsymbol{\mu}^X$, $\boldsymbol{\nu}^X$ are different Dirac measures).

More generally, our papers consider Rényi divergences of any positive order. The SCR then becomes the Shifted Composition Rule, analogously extending the standard Rényi composition rule via this additional flexibility in X'. In this abstract, we focus on KL for simplicity of exposition.

II. Reverse transport inequalities

In these two papers, our main application is the derivation of reverse transport inequalities for the Langevin diffusion

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad (1)$$

where $(B_t)_{t\ge0}$ is Brownian motion. Let $(P_t)_{t\ge0}$ denote the Markov semigroup for (1) and fix probability measures μ, ν, ν' and $t > 0$. As a representative example of our results, we use the SCR to show that if $\nabla^2 V \succeq \alpha I$, then

$$\mathsf{KL}(\mu P_t \,\|\, \nu P_t) \le \frac{\alpha\, W_2^2(\mu, \nu)}{2\,(\exp(2\alpha T) - 1)}\,, \qquad (2)$$

and if $-\beta I \preceq \nabla^2 V \preceq \beta I$, then

$$\mathsf{KL}(\mu P_t * \nu \,\|\, \mu P_t * \nu') \le \frac{\beta\, W_2^2(\nu, \nu')}{2\,(1 - \exp(-2\beta t))}\,. \qquad (3)$$

These inequalities capture complementary aspects of the diffusion: (2) measures sensitivity w.r.t. the initial condition (indeed, for $\alpha \ge 0$ it yields a mixing time bound), whereas (3) captures the regularity of the marginal law of the process. In other words, they encode regularity for Kolmogorov's backward and forward equations, respectively.

As an illustration of the use of the SCR, suppose that we want to establish (2) when $\mu = \delta_x$, $\nu = \delta_y$ are Diracs. To formulate an argument in discrete time, we first replace the continuous-time semigroup $(P_t)_{t\ge0}$ by a discretized one and apply a limiting argument. Then, the question is to bound $\mathsf{KL}(\delta_x P^N \,\|\, \delta_y P^N)$ for a Markov kernel P. A naïve application of the KL chain rule is vacuous, since $\mathsf{KL}(\delta_x \,\|\, \delta_y) = \infty$. Instead, we construct an auxiliary process $\{X_n'\}_{n=0}^N$ such that $X_0' = y$ and $X_N' \sim \delta_x P^N$, and we instead bound $\mathsf{KL}(\mathrm{law}(X_N') \,\|\, \delta_y P^N)$ via the SCR (details in [AC23]). In this context, this argument can be seen as a generalization of the shifted divergence technique from the differential privacy and sampling literature [Fel+18; AT22; AT23] or as a discrete-time analogue of the coupling in [ATW06].

III. Functional analysis, geometry, and probability

Inequalities (2) and (3) are part of a larger story—called *Bakry–Émery* theory [BGL14]—which relates analytic properties of the semigroup, through functional inequalities, to the curvature of the underlying space and of the measure (i.e., the Hessian $\nabla^2 V$ of the negative log-density), and to probabilistic aspects such as concentration of measure and mixing. Indeed, it is well-known that via duality, (2) is equivalent to the celebrated dimension-free *Harnack* inequality of [Wan97], and implies back the curvature lower bound $\nabla^2 V \succeq \alpha I$.

On the other hand, the inequality (3), which is equivalent to a *shift Harnack* inequality [Wan14], appears in its sharp form for the first time in our paper [AC24]. This allows us to prove that (3) implies back the curvature *upper bound* $\nabla^2 V \preceq \beta I$. In our paper, we leave open the intriguing question of whether this observation can form the basis of a Bakry–Émery theory for curvature upper bounds.

References

[AC24] J. M. Altschuler and S. Chewi. “Shifted composition II: shift Harnack inequalities and curvature upper bounds”. In: *arXiv preprint arXiv:2401.00071* (2024).

[AC23] J. M. Altschuler and S. Chewi. “Shifted composition I: Harnack and reverse transport inequalities”. In: *arXiv preprint 2311.14520* (2023).

[AT22] J. M. Altschuler and K. Talwar. “Privacy of noisy stochastic gradient descent: more iterations without more privacy loss”. In: *Advances in Neural Information Processing Systems*. 2022.

[AT23] J. M. Altschuler and K. Talwar. “Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling”. In: *Conference on Learning Theory*. 2023.

[ATW06] M. Arnaudon, A. Thalmaier, and F.-Y. Wang. “Harnack inequality and heat kernel estimates on manifolds with curvature unbounded below”. In: *Bull. Sci. Math.* 130.3 (2006), pp. 223–233.

[BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.

[Fel+18] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta. “Privacy amplification by iteration”. In: *Symposium on Foundations of Computer Science*. 2018.

[Wan97] F.-Y. Wang. “Logarithmic Sobolev inequalities on noncompact Riemannian manifolds”. In: *Probability Theory and Related Fields* 109.3 (1997), pp. 417–424.

[Wan14] F.-Y. Wang. “Integration by parts formula and shift Harnack inequality for stochastic equations”. In: *The Annals of Probability* 42.3 (2014), pp. 994–1019.

Bounds on transport maps via diffusion processes

Max Fathi
Université Paris Cité and Sorbonne Université, CNRS,
Laboratoire Jacques-Louis Lions
and Laboratoire de Probabilités, Statistique et Modélisation,
F-75013 Paris, France
and DMA, École normale supérieure,
Université PSL, CNRS,
75005 Paris, France
and Institut Universitaire de France
email: mfathi@lpsm.paris

***Abstract*—Globally lipschitz transport maps have found many applications in the study of probabilistic functional inequalities such as logarithmic Sobolev and Poincaré inequalities, by transporting an inequality from a nice reference measure to another one. For example, a theorem of Caffarelli states that optimal transport maps from the standard Gaussian measure onto uniformly log-concave measures are 1-lipschitz. This then recovers the sharp bounds of Bakry and Emery on the logarithmic Sobolev constant of such measures.**

In this talk, I will discuss a construction of non-optimal transport maps using the heat flow, due to Kim and Milman, and explain how it allows to get dimension-free lipschitz maps in new settings, including certain Riemannian manifolds. Joint work with D. Mikulincer and Y. Shenfeld.

Caffarelli's contraction theorem [2] states that probability measures on $\mathbb{R}^d$ with density of the form e^{-V} with $\operatorname{Hess} V \geq \mathrm{Id}$ can be realized as the image of a standard Gaussian measure on $\mathbb{R}^d$ by a 1-lipchitz map. The map is actually the Brenier map from optimal transport theory [1], viewed as a solution of a Monge-Ampère partial differential equation. This result found various applications to the study of sharp constants in fucntional inequalities and concentration inequalities. We refer to [5], [6] for surveys and further developments.

Later on, Kim and Milman [4] gave an alternative construction of Lipschitz transport maps under the same assumptions as Caffarelli's theorem, using heat flow to define a map, and deducing regularity properties of the map from regularity properties of the heat flow.

In this talk, I will present results of [3] that show how Kim and Milman's argument works in other situations, where the quantitative regularity of Brenier maps is not well-understood. In particular, we show that there exists globally Lipschitz maps with dimension-free constants between the standard Gaussian measure and log-lipshcitz perturbations of it. A precise statement is

Theorem 1: Let γ be the standard Gaussian measure on $\mathbb{R}^d$ and $\nu = e^f d\gamma$ be a probability measure such that f is L-Lipschitz. Then there exists a transport map T sending γ onto ν such that $||T||_{lip} \leq \exp(c(L + L^2))$, where c is a numerical constant that does not depend on f nor on the dimension.

The Lispchitz constant is explicit, and of the sharp order of magnitude when applied to deduce functional inequalities such as Poincaré inequalities. Similar results are also obtained for non-Gaussian measures, under assumptions of uniform convexity of the potential and bounds on the third-order derivatives.

We also investigated Riemannian manifolds satisfying certain geometric assumptions on the curvature, showing for example existence of dimension-free Lispchitz transport maps between certain probability measures on spheres with appropriately scaled radius.

Acknowledgements

M.F. was supported by the Projects MESA (ANR-18-CE40-006) and EFI (ANR-17-CE40-0030) of the French National Research Agency (ANR).

References

[1] Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* vol. 44, no. 2, pp. 375–417, 1991.

[2] L. A. Caffarelli, Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.*, vol. 214, (2000), pp. 547–563

[3] M. Fathi, D. Mikulincer and Y. Shenfeld, Transportation onto log-Lipschitz perturbations. *Arxiv preprint* 2023.

[4] Y.-H. Kim and E. Milman, A generalization of Caffarelli's contraction theorem via (reverse) heat flow. *Math. Ann.* vol. 354 (2012) pp. 827–862.

[5] A. Kolesnikov, Mass transportation and contractions. *Arxiv preprint* 2011.

[6] E. Milman, Spectral estimates, contractions and hypercontractivity. *J. Spectral Theory*, vol. 8, (2018), pp. 669–714.

Gromov–Wasserstein Alignment: Statistical and Computational Advancements via Duality

Ziv Goldfeld
Cornell University
email: goldfeld@cornell.edu

***Abstract*—The Gromov-Wasserstein (GW) distance quantifies dissimilarity between metric measure (mm) spaces and provides a natural correspondence between them. As such, it serves as a figure of merit for applications involving alignment of heterogeneous datasets, including object matching, single-cell genomics, and language models translation. While various heuristic methods for approximately evaluating the GW distance from data have been developed, formal guarantees for such approaches—both statistical and computational—remained elusive. This work closes these gaps for the quadratic GW distance between Euclidean mm spaces of different dimensions. At the core of our proofs is a novel dual representation of the GW problem as an infimum of a certain class of optimal transportation problems. The dual form enables deriving, for the first time, sharp empirical convergence rates for the GW distance by providing matching upper and lower bounds. For computational tractability, we consider the entropically regularized GW distance. We derive bounds on the entropic approximation gap, establish sufficient conditions for smoothness and convexity of the objective in the dual problem, and devise efficient algorithms with local and, under convexity, even global convergence guarantees. These advancements facilitate principled estimation and inference methods for GW alignment problems, that are efficiently computable via the said algorithms.**

I. Extended Abstract

The Gromov-Wasserstein (GW) distance quantifies discrepancy between probability distributions supported on different metric spaces by aligning them with one another. Given two metric measure (mm) spaces $(\mathcal{X}, \mathsf{d}_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, \mathsf{d}_{\mathcal{Y}}, \nu)$, the (p,q)-GW distance between them is [1], [2]

$$\mathsf{D}_{p,q}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left(\iint_{\mathcal{X}\times\mathcal{Y}} \Delta_q^p \, d\pi\otimes\pi \right)^{\frac{1}{p}}, \qquad (1)$$

where $\Delta_q(x,y,x',y') := \left|\mathsf{d}_{\mathcal{X}}(x,x')^q - \mathsf{d}_{\mathcal{Y}}(y,y')^q\right|$ is the distance distortion cost, $\Pi(\mu,\nu)$ is the set of all couplings between μ and ν. The GW distances thus equals the least amount of distance distortion one can achieve between the mm spaces when optimizing over all possible alignments thereof (as modeled by couplings). This approach, which is rooted in optimal transport (OT) theory, is an L^p relaxation of the Gromov-Hausdorff distance between metric spaces and enjoys various favorable properties. Among others, the GW distance (i) identifies pairs of mm spaces between which there exists an measure preserving isometry; (ii) defines a metric on the space of all mm spaces modulo the aforementioned isomorphic relation; and (iii) captures empirical convergence of mm space, i.e., when μ,ν are replaced with their empirical measures $\hat{\mu}_n, \hat{\nu}_n$ based on n samples. Although alignment schemes inspired by the GW framework have seen many applications in computer vision, machine learning, single-cell genomics, and more, existing estimation and computation methods are heuristic and lack formal sample or time complexity guarantees.

To close these gaps, we develop a duality theory for the GW distance, which linearizes this quadratic functional and ties it to the well-understood OT problem. This is done by introducing an auxiliary, matrix-valued optimization variable $\mathbf{A} \in \mathbb{R}^{d_x\times d_y}$ that enables linearizing the dependence on the coupling. We then interchange the optimization over $\mathbf{A}$ and π and identify the inner problem as classical OT problem with respect to a cost function $c_{\mathbf{A}}$ that depends on $\mathbf{A}$. This representation allows us to lift tool from statistical OT to derive, for the first time, the sample complexity of the empirical plug-in estimator of the GW distance. The derived two-sample rate is $n^{-2/\max\{\min\{d_x,d_y\},4\}}$ (up to a log factor when $\min\{d_x,d_y\}=4$), which matches the corresponding rates for empirical OT. We then provide matching lower bounds, thereby establishing sharpness of the derived rates.

From a computational standpoint, evaluation of the GW distance requires solving a quadratic assignment problem, which is known to be NP-complete. A popular, computationally tractable proxy is the entropic GW (EGW) problem, which regularizes the distance distortion cost from (1) by the Kullback-Leibler divergence penalty $\epsilon\mathsf{D}_{\mathsf{KL}}(\pi\|\mu\otimes\nu)$; $\epsilon > 0$ is the regularization parameter. We show that the entropic approximation gap is at most $O(\epsilon\log(1/\epsilon))$, whereby EGW can approximate GW to an arbitrary precision. We then note that our GW duality naturally extends to the EGW distance, up to replacing the OT cost in the variational problem with its entropic OT (EOT) counterpart. Leveraging the connection to EOT, we derive smoothness and convexity properties of the objective in this variational problem, which enable computing it via accelerated gradient descent. Gradients are evaluated by employing Sinkhorn's algorithm to solve the EOT problem, which we model as an inexact oracle and account for it in our analysis. This results in the first efficient algorithms for solving the EGW problem that are subject to formal guarantees in both the convex and non-convex regimes. There results enable principled estimation and computation of GW alignment.

References

[1] F. Mémoli, “Gromov-Wasserstein distances and the metric approach to object matching,” *Found. Comput. Math.*, vol. 11, no. 4, 2011.

[2] K.-T. Sturm, “The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces,” *arXiv:1208.0434*, 2012.

Training Generative Models from Privatized Data via Entropic Optimal Transport

Daria Reshetova Wei-Ning Chen Ayfer Özgür

Stanford University

{resh, wnchen, aozgur}@stanford.edu

Local differential privacy has been a powerful method for privacy-preserving data collection. It is typically done by adding proper noise so that only population-level statistics can be inferred, hence protecting sensitive individual information. In this work, we develop a framework for training Generative Adversarial Networks (GANs) on locally differentially privatized data. We show that entropic regularization of optimal transport – a popular regularization method in the literature that has often been leveraged for its computational benefits – enables the generator to learn the raw (unprivatized) data distribution even though it only has access to privatized samples. We prove that, at the same time, this leads to fast statistical convergence at the parametric rate. This shows that entropic regularization of optimal transport uniquely enables the mitigation of both the effects of privatization noise and the curse of dimensionality in statistical convergence. We provide experimental evidence to support the efficacy of our framework in practice.

Linearized Brascamp–Lieb Inequalities

Thomas A. Courtade
University of California, Berkeley
Department of Electrical Engineering and Computer Sciences
email: courtade@berkeley.edu

***Abstract*—Combining Valdimarsson's characterization of extremizers for the Brascamp–Lieb inequalities together with their dual entropic form, a linearization argument reveals that several well-known inequalities in probability can be viewed as consequences of the Brascamp–Lieb inequalities. The resulting "linearized Brascamp–Lieb inequalities" admit interpretation as a sharp spectral gap inequality for a simple physical process.**

I. Introduction

Fix a Euclidean space E, linear subspaces $E_i \subset E$, $i = 1, \dots, k$, a collection of linear maps $\mathbf{B} = (B_i : E \to E_i)_{i=1}^k$, and non-negative real numbers $\mathbf{c} = (c_i)_{i=1}^k \subset (0, \infty)$. The pair $(\mathbf{B}, \mathbf{c})$ is called a *(Brascamp–Lieb) datum*. The *Brascamp–Lieb inequalities* [3], [4] take the form

$$\int_E \prod_{i=1}^k (f_i \circ B_i)^{c_i} \leq e^{\mathrm{BL}(\mathbf{B},\mathbf{c})} \prod_{i=1}^k \left(\int_{E_i} f_i \right)^{c_i}, \tag{1}$$

where the *Brascamp–Lieb constant* $\mathrm{BL}(\mathbf{B}, \mathbf{c})$ is defined to be the smallest constant such that (1) holds for all non-negative $f_i \in L^1(E_i)$, $i = 1, \dots, k$. Here, the integrals are with respect to Lebesgue measure, and a theorem of Lieb [15] is that $\mathrm{BL}(\mathbf{B}, \mathbf{c})$ can be computed by considering centered Gaussian functions $(f_i)_{i=1}^k$.

For a linear subspace $V \subset E$, we let $P_V : E \to E$ denote the orthogonal projection of E onto V. A datum $(\mathbf{B}, \mathbf{c})$ is said to be *geometric* if $B_i^* B_i = P_{E_i}$ for each $i = 1, \dots, k$, and the following *frame condition* holds:

$$\sum_{i=1}^k c_i P_{E_i} = \mathrm{id}_E\,. \tag{2}$$

When $(\mathbf{B}, \mathbf{c})$ is geometric, we have $\mathrm{BL}(\mathbf{B}, \mathbf{c}) = 0$ [1].

For a given datum $(\mathbf{B}, \mathbf{c})$, inequality (1) is said to be *extremizable* if there exist admissible $(f_i)_{i=1}^k$ such that (1) is met with equality. Modulo an equivalence relation that amounts to a linear change of variables, it is known that all extremizable data are equivalent to geometric data [1], and the extremizers in this case have been completely characterized by Valdimarsson [18].

For a Euclidean space E, let $\mathcal{M}(E)$ denote the set of Borel probability measures on E, absolutely continuous with respect to Lebesgue measure. For $\mu \in \mathcal{M}(E)$ with density $d\mu = f dx$, We define the *(Shannon) entropy*

$$h(\mu) = -\int_E f \log f dx,$$

provided the integral exists in the Lebesgue sense. Carlen and Cordero-Erausquin [8] observed the following dual formulation of the Brascamp–Lieb inequalities: For every $\mu \in \mathcal{M}(E)$ with finite entropy,

$$h(\mu) \leq \sum_{i=1}^k c_i h(B_i \# \mu) + \mathrm{BL}(\mathbf{B}, \mathbf{c}), \tag{3}$$

where # denotes the usual pushforward operation. We say that (3) is *extremizable* if there exists $\mu \in \mathcal{M}(E)$ such that (3) is an equality, and all entropies therein are finite; such a μ is called an *extremizer*. As one expects, (3) is extremizable if and only if (1) is extremizable. Hence, we can simply refer to the datum $(\mathbf{B}, \mathbf{c})$ as being extremizable without confusion.

Recall that for two probability measures $\nu, \mu \in \mathcal{M}(E)$, the *relative entropy* is defined as

$$D(\nu \| \mu) := \begin{cases} \int_E \log(\frac{d\nu}{d\mu}) d\nu & \text{if } \nu \ll \mu \\ +\infty & \text{otherwise.} \end{cases}$$

Having recalled all of the above, we can now state a variation of the Brascamp–Lieb inequalities involving relative entropies, for reference measure equal to an extremizer of (3).

Theorem 1. *Let $(\mathbf{B}, \mathbf{c})$ be extremizable, and $\mu \in \mathcal{M}(E)$ an extremizer in (3). For any $\nu \in \mathcal{M}(E)$, we have*

$$\sum_{i=1}^k c_i D(B_i \# \nu \| B_i \# \mu) \leq D(\nu \| \mu). \tag{4}$$

A linearization argument leads to the following family of variance inequalities, which is the subject of this note.

Theorem 2. *Let $(\mathbf{B}, \mathbf{c})$ be extremizable, and $\mu \in \mathcal{M}(E)$ an extremizer in (3). For $X \sim \mu$ and integrable $\varphi : E \to \mathbb{R}$,*

$$\sum_{i=1}^k c_i \operatorname{Var}(\mathbb{E}[\varphi(X) | B_i X]) \leq \operatorname{Var}(\varphi(X)). \tag{5}$$

In order to apply (5) in practice, we need two things: (i) a characterization of extremizable data; and (ii) a characterization of extremal $\mu \in \mathcal{M}(E)$ in (3). The first has been already addressed, and in particular, it suffices to consider geometric data, which are concisely characterized by the frame condition (2). The second can also be addressed easily enough. In particular, Valdimarsson's characterization of extremal $(f_i)_{i=1}^k$ in (1) can be translated to a neat characterization of extremal μ in (3). To state it, let $\mu \in \mathcal{M}(E)$, and let μ_{E_i} (resp. $\mu_{E_i^\perp}$) denote the marginal of μ on E_i (resp. $E_i^\perp$). We say that μ *splits*

The author acknowledges support from NSF-CCF 1750430, the hospitality of the Laboratoire de Probabilités, Statistique et Modélisation, and the Invited Professor Program of the Fondation Sciences Mathématiques de Paris.

along $(E_i, E_i^\perp)$ if we have the decomposition $\mu = \mu_{E_i} \otimes \mu_{E_i^\perp}$. In other words, μ splits along $(E_i, E_i^\perp)$, if it is product with respect to the orthogonal decomposition $E = E_i \oplus E_i^\perp$.

The following can be distilled from Valdimarsson's characterization of extremal $(f_i)_{i=1}^k$ in (1), and provides a satisfactory answer to the second issue noted above.

Proposition 1. *Let* $(\mathbf{B}, \mathbf{c})$ *be geometric, and let* $\mu \in \mathcal{M}(E)$ *have finite entropy. The following are equivalent:*
1) μ *is an extremizer in* (3)*;*
2) μ *splits along* $(E_i, E_i^\perp)$ *for each* $i = 1, \dots, k$.

We remark that Valdimarsson [18] actually leads to a more explicit characterization of extremal μ than above (roughly speaking, an extremal μ has a rigid factorization into independent components, with some factors chosen freely, and others isotropic Gaussians). However, for our purposes, the characterization in Proposition 1 suffices, and is easily stated.

We thus arrive at the following simple and explicit statement, which we call *linearized Brascamp–Lieb inequalities*.

Corollary 1 (Linearized Brascamp–Lieb inequalities)**.** *Let* $\mathbf{c}$ *and* $(E_i)_{i=1}^k$ *satisfy the frame condition* (2)*. If* X *has law that splits along* $(E_i, E_i^\perp)$ *for each* $i = 1, \dots, k$*, then for all integrable* $\varphi : E \to \mathbb{R}$*,*

$$\sum_{i=1}^k c_i \operatorname{Var}(\mathbb{E}[\varphi(X) | P_{E_i} X]) \leq \operatorname{Var}(\varphi(X)). \tag{6}$$

The remainder of this note is organized as follows. Section II illustrates a few applications of (6) to inequalities in probability. Section III explains how (6) may be interpreted as a sharp spectral gap inequality. Section IV contains the proofs, and Section V gives some brief concluding remarks.

II. Applications

It's well-known that the Brascamp–Lieb inequalities (1) contain many classical analytic and geometric inequalities (e.g., the Hölder, Young, and Loomis–Whitney inequalities), and their dual formulation (3) can be seen as generalizing the information-theoretic inequality known as subadditivity of entropy. All of these applications require only the evaluation of $\mathrm{BL}(\mathbf{B}, \mathbf{c})$, which can be accomplished in practice due to the Gaussian saturation property. By incorporating the characterization of extremizers into the picture, we obtain (4) and (5). As a consequence, we find that a variety of probabilistic inequalities may also be obtained from the Brascamp–Lieb inequalities. Toward that end, let us now demonstrate some special cases of the linearized Brascamp–Lieb inequalities.

Example 1 (Efron–Stein inequality [13], [17])**.** *Let* $X = (X_i)_{i=1}^k$ *be a random vector with independent components* $(X_i)_{i=1}^k$*, and define*

$$X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots X_k).$$

For any measurable φ *with* $\operatorname{Var}(\varphi(X)) < \infty$*,*

$$\operatorname{Var}(\varphi(X)) \leq \sum_{i=1}^k \mathbb{E}[\operatorname{Var}(\varphi(X) | X^{(i)})]. \tag{7}$$

Proof. We can assume X takes values in E, and choose E_i such that X_i is the component of X in $E_i^\perp$. This implies the orthogonal decomposition $E = \oplus_{i=1}^k E_i^\perp$, which yields the frame condition

$$\frac{1}{k-1} \sum_{i=1}^k P_{E_i} = \mathrm{id}_E \,.$$

By the independence hypothesis, the law of X splits along $(E_i, E_i^\perp)$ for each $i = 1, \dots, k$, and therefore (7) follows from (6) by invoking the classical variance decomposition $\operatorname{Var}(\varphi(X)) = \mathbb{E}[\operatorname{Var}(\varphi(X)|Y)] + \operatorname{Var}(\mathbb{E}[\varphi(X)|Y])$ with $Y = X^{(i)}$. □

More generally, the classical variance decomposition can be applied directly to (6) to obtain a generalized version of the Efron–Stein inequality. We'll return to this in our interpretation of (5) as a spectral gap inequality in Section III.

Example 2 (Dembo–Kagan–Shepp inequality [12])**.** *Let* $(X_i)_{i\geq 1}$ *be a sequence of i.i.d. random vectors, and define* $S_n = \sum_{j=1}^n X_j$*. If function* g *satisfies* $\mathbb{E}[g(S_n)] < \infty$*, then*

$$\operatorname{Var}(\mathbb{E}[g(S_n)|S_m]) \leq \frac{m}{n} \operatorname{Var}(g(S_n)), \qquad n \geq m \geq 1. \tag{8}$$

Proof. For simplicity of notation, we'll assume each X_i is one-dimensional. Consider the random vector $X = (X_1, \dots, X_n)$ taking values in $E := \mathbb{R}^n$, with X_j the projection of X along natural basis vector e_j, $j = 1, \dots, n$. Take $(E_i)_{i=1}^k$ be an enumeration of all $k = \binom{n}{m}$ subspaces of E, equal to the linear span of exactly m natural basis vectors. By construction, X splits along $(E_i, E_i^\perp)$, and the frame condition (2) holds with $c_i = \frac{n}{m} / \binom{n}{m}$ for each i. By symmetry, $\mathbb{E}[g(S_n)|P_{E_i}X]$ are equal in law for each $i = 1, \dots, k$. So, an application of (6) with $\varphi(X) = g(S_n)$ gives

$$\operatorname{Var}(\mathbb{E}[g(S_n)|X_1, \dots, X_m]) \leq \frac{m}{n} \operatorname{Var}(g(S_n)), \qquad n \geq m \geq 1.$$

The claim follows since S_m is a sufficient statistic of $(X_1, \dots, X_m)$ for S_n. □

By L^2 duality, (5) is equivalent to the following "variance drop" inequality.

Corollary 2 (Variance Drop[1])**.** *Let the notation and assumptions of Theorem 2 prevail. For any real-valued* $\psi_i : B_i X \mapsto \psi_i(B_i X)$ *with finite variance,*

$$\operatorname{Var}\left(\sum_{i=1}^k c_i \psi_i(B_i X) \right) \leq \sum_{i=1}^k c_i \operatorname{Var}\left(\psi_i(B_i X) \right). \tag{9}$$

Moreover, this is equivalent to (5)*.*

It is tempting to regard (9) as a consequence of Jensen's inequality applied to convexity of variance. To see that it is not, assume without loss of generality that $(\mathbf{B}, \mathbf{c})$ is geometric. Taking traces of the frame condition implies $\sum_{i=1}^k c_i \leq 1$, with

[1] Inequality (9) can be obtained by applying (1) to functions $f_i = e^{\epsilon \psi_i} \tilde{f}_i$, for extremal $(\tilde{f}_i)_{i=1}^k$ and vanishing ϵ. However, the interpretation of $\tilde{f}_i$ as the marginal density of $B_i \# \mu$ for some meaningful μ only becomes apparent upon inspection of passage between (1) and (3) via duality.

equality only in the case where $E_i = E$ for every i. In this latter case, every X is an admissible extremizer. Hence, (9) is a strict improvement of Jensen's inequality except in degenerate cases.

Proof. To see that (5) implies (9), put $\varphi := \sum_{i=1}^k c_i \psi_i \circ B_i$. Applying Cauchy-Schwarz twice followed by (5), we have

$$\begin{aligned}
&\operatorname{Var}(\varphi(X)) \\
&= \sum_{i=1}^k c_i \operatorname{Cov}(\mathbb{E}[\varphi(X)|B_i X], \psi_i(B_i X)) \\
&\leq \sum_{i=1}^k c_i \operatorname{Var}\left(\mathbb{E}[\varphi(X)|B_i X]\right)^{1/2} \operatorname{Var}\left(\psi_i(B_i X)\right)^{1/2} \\
&\leq \left(\sum_{i=1}^k c_i \operatorname{Var}\left(\mathbb{E}[\varphi(X)|B_i X]\right)\right)^{1/2} \left(\sum_{i=1}^k c_i \operatorname{Var}\left(\psi_i(B_i X)\right)\right)^{1/2} \\
&\leq \operatorname{Var}\left(\varphi(X)\right)^{1/2} \left(\sum_{i=1}^k c_i \operatorname{Var}\left(\psi_i(B_i X)\right)\right)^{1/2} .
\end{aligned}$$

To see the reverse implication (9) ⇒(5), observe that

$$\begin{aligned}
&\sum_{i=1}^k c_i \operatorname{Var}(\mathbb{E}[\varphi(X)|B_i X]) \\
&\quad= \sum_{i=1}^k c_i \operatorname{Cov}(\varphi(X), \mathbb{E}[\varphi(X)|B_i X]) \\
&\quad\leq \operatorname{Var}(\varphi(X))^{1/2} \operatorname{Var}\left(\sum_{i=1}^k c_i \mathbb{E}[\varphi(X)|B_i X]\right)^{1/2} \\
&\quad\leq \operatorname{Var}\left(\varphi(X)\right)^{1/2} \left(\sum_{i=1}^k c_i \operatorname{Var}\left(\mathbb{E}[\varphi(X)|B_i X]\right)\right)^{1/2} ,
\end{aligned}$$

where the first inequality is Cauchy–Schwarz, and the second follows from (9) with $\psi_i(B_i X) = \mathbb{E}[\varphi(X)|B_i X]$. □

As a special case, we recover an inequality due to Madiman and Barron [16], which is itself a generalization of a classical result on U-statistics due to Hoeffding [14]. To state it, recall that $\mathcal{T} \subset 2^{[n]}$ is said to be an r-cover of $[n] := \{1, \dots, n\}$ if each element of $[n]$ is contained in exactly r members of $\mathcal{T}$.

Example 3 (Madiman–Barron inequality [16])**.** *Let $X = (X_m)_{m=1}^n$ be a collection of n independent random random vectors, and let $(S_i)_{i=1}^k \subset 2^{[n]}$ be an r-cover of $[n]$. For any real-valued $\psi_i : B_i X \mapsto \psi_i(B_i X)$ with finite variance,*

$$\operatorname{Var}\left(\sum_{i=1}^k \psi_i(X_{S_i})\right) \leq r \sum_{i=1}^k \operatorname{Var}\left(\psi_i(X_{S_i})\right), \qquad (10)$$

where $X_{S_i} := (X_m)_{m \in S_i}$.

Proof. Let E be the space in which the random vector $X = (X_1, \dots, X_n)$ takes values. Consider the geometric datum with $c_i = 1/r$ and E_i equal to the subspace of E in which the coordinate X_{S_i} lives, and apply (9). □

We've focused this section on implications of (5), but we emphasize that the relative entropy inequalities (4) also contain useful results. To give a quick example, we note that Shearer's inequality corresponds to the case where $(\mathbf{B}, \mathbf{c})$ is geometric, and μ has suitable product structure.

Example 4 (Shearer's inequality [10])**.** *Let E admit an orthogonal decomposition $E = \oplus_{m=1}^n V_m$, and let $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ enjoy product structure with respect to this decomposition (μ_m is a probability measure on V_m, $m = 1, \dots, n$). Fix a collection of subsets $(S_i)_{i=1}^k \subset 2^{[n]}$. If $\mathbf{c} = (c_i) \subset (0, \infty)_{i=1}^k$ satisfies $\sum_{i: S_i \ni m} c_i = 1$ for each $m = 1, \dots, n$, then for all probability measures ν*

$$\sum_{i=1}^k c_i D(\nu_{S_i} \| \mu_{S_i}) \leq D(\nu \| \mu), \qquad (11)$$

where μ_{S_i} (resp. ν_{S_i}) denotes the marginal of μ (resp. ν) on $\oplus_{m \in S_i} V_i$.

Proof. Put $E_i = \oplus_{m \in S_i} V_i$, and note that $\sum_{i: S_i \ni m} c_i = 1$ coincides with the frame condition (2). Thus, the claim follows from (4). □

Remark 1. *The most common statement of Shearer's inequality assumes $(S_i)_{i=1}^k$ is an r-cover of $[n]$, and has all $(c_i)_{i=1}^k$ equal to $1/r$. However, inequality* (11) *can be regarded as a simple self-strengthening obtained by iteration. A weighted version of* (10) *also appears in [16].*

Remark 2. *In the terminology of Valdimarsson [18], Shearer's inequality* (11) *corresponds to* (4) *in the special case of a geometric datum* $(\mathbf{B}, \mathbf{c})$ *with no "dependent subspace".*

The author has previously observed that the Dembo–Kagan–Shepp inequality and the Madiman–Barron inequality can be derived directly by linearizing Shearer's inequality [11], as can be the Efron–Stein inequality. Of course, each of these classical inequalities has its own direct proof by ad hoc arguments. Nevertheless, these examples are worth repeating to emphasize their interpretation as special cases of linearized Brascamp–Lieb inequalities. The following is a simple explicit example of a linearized Brascamp–Lieb inequality that is not a linearization of Shearer's inequality.

Example 5. *Let $X \sim N(0, \mathrm{id}_{\mathbb{R}^2})$, and let $(u_i)_{i=1}^3 \subset \mathbb{R}^2$ be equiangular unit vectors (i.e., $u_i^T u_i = 1$ and $u_i^T u_{i'} = \cos(2\pi/3) = -1/2$ for $i \neq i'$). For any integrable φ,*

$$\sum_{i=1}^3 \operatorname{Var}(\mathbb{E}[\varphi(X)|u_i^T X]) \leq \frac{3}{2} \operatorname{Var}(\varphi(X)).$$

III. Spectral gap interpretation

We've seen how several inequalities in probability follow as special cases of the linearized Brascamp–Lieb inequalities. Now, we turn attention to the most general statement of the linearized Brascamp–Lieb inequalities and give a simple physical interpretation, inspired by the folklore interpretation of the Efron–Stein inequality as a Poincaré (or, spectral gap) inequality. Toward this end, for a linear subspace $V \subset E$

and $x \in E$, write $x = (x_V, x_{V^\perp})$, where $x_V := P_V x$, and $x_{V^\perp} := P_{V^\perp} x = (\mathrm{id}_E - P_V)x$.

Consider an experiment where two particles of equal mass and respective velocities $x, x' \in E$ undergo an elastic collision. By conservation of energy and momentum, the particles necessarily exchange velocity components on some subspace V. That is, the post-collision velocities of the first and second particles are, respectively:

$$x_+ = (x'_V, x_{V^\perp}), \quad \text{and} \quad x'_+ = (x_V, x'_{V^\perp}).$$

Suppose we now adopt a probabilistic collision model in which the subspace V is randomly chosen from some set $\{V_1, \dots, V_k\}$, with respective probabilities $p_1, \dots, p_m$. Then, given pre-collision velocities x, x', the expected change of velocity imparted to the first particle through collision is

$$\Delta v(x, x') = \sum_{i=1}^{k} p_j P_{V_i}(x' - x).$$

If the incoming velocities x, x' undergo a common orthogonal transformation, then a natural physical constraint imposed on the model is that the expected change in velocity $\Delta v(x, x')$ should undergo the same orthogonal transformation. That is, we require Δv to satisfy $\Delta v(Ux, Ux') = U \Delta v(x, x')$ for all $x, x' \in E$ and orthogonal $U : E \to E$. Using definitions, this invariance implies that there must exist some $\lambda \in \mathbb{R}$ such that

$$\sum_{i=1}^{k} p_i P_{V_i} = \lambda \, \mathrm{id}_E \, .$$

Moreover, it is easy to check that $0 \le \lambda \le 1$, with equality only in the trivial cases where $V_i = \{0\}$ for every i (non-interacting particles), or where $V_i = E$ for every i (particles completely exchange velocities).

Now, let μ be a probability measure on E, and consider a stochastic process $(X(t); t \ge 0)$ where a particle with initial velocity $X(0)$ is placed in contact with a bath containing particles with velocities distributed i.i.d. according to μ, and collisions between our particle and particles in the bath occur at rate 1, according to a Poisson point process. Note that if a collision happens at time t, the post-collision velocity of our particle will be

$$X(t+) = (X'_{V_i}, X_{V_i^\perp}(t-)) \quad \text{with probability } p_i, \ 1 \le i \le k,$$

where $X' \sim \mu$ is independent of the pre-collision velocity $X(t-)$ of the particle of interest. Assuming the bath is in equilibrium, the background measure μ must be invariant under these dynamics, which is true if and only if it splits along each $(V_i, V_i^\perp)$, $i = 1, \dots, k$.

The linearized Brascamp–Lieb inequalities can be interpreted as a spectral gap inequality for this stochastic process. Indeed, define $E_i := V_i^\perp$ and $c_i := \frac{p_i}{1-\lambda}$, which can be checked to satisfy the frame condition (2). For $X \sim \mu$, the linearized Brascamp–Lieb inequalities can be rewritten as

$$\mathrm{Var}(\varphi(X)) \le \frac{1}{\lambda} \sum_{i=1}^{k} p_i \mathbb{E}[\mathrm{Var}(\varphi(X) | X_{V_i^\perp})],$$

by the classical variance decomposition. Thus, in general, the linearized Brascamp–Lieb inequalities coincide with the sharp Poincaré inequality for the described dynamics.

The inequality (4) can similarly be interpreted as governing convergence to equilibrium, but in the stronger sense of relative entropy. In our setting, (4) can be written as

$$\sum_{i=1}^{k} p_i D(\mu_{V_i} \otimes \nu_{V_i^\perp} \| \mu) \le (1 - \lambda) D(\nu \| \mu), \tag{12}$$

where μ_{V_i} and $\nu_{V_i^\perp}$ denote the marginals of μ and ν on V_i and $V_i^\perp$, respectively. If our particle has pre-collision velocity with law ν, then the post-collision velocity of the particle will have $\mu_{V_i} \otimes \nu_{V_i^\perp}$ with probability p_i, and therefore the law of the post-collision velocity averaged over the collision model is the mixture $\nu_+ := \sum_{i=1}^{k} \mu_{V_i} \otimes \nu_{V_i^\perp}$. By convexity of relative entropy, the above inequality implies $D(\nu_+ \| \mu) \le (1 - \lambda) D(\nu \| \mu)$, demonstrating a strict trend to equilibrium in relative entropy with each collision. Since we assume collisions occur at rate 1, if our particle has initial velocity with law ν_0 and $(\nu_t)_{t \ge 0}$ denotes the evolution of ν_0 along these dynamics, an application of Grönwall's lemma yields the exponential decay of entropy

$$D(\nu_t \| \mu) \le e^{-\lambda t} D(\nu_0 \| \mu), \quad t \ge 0.$$

Remark 3. *There seems to be no fundamental reason to limit ourselves to a discrete set of collision possibilities. For example, if $E = \mathbb{R}^n$, we could take the frame to be $\{P_{\mathrm{span}\{\sigma\}}; \sigma \in \mathbb{S}^{n-1}\}$, equipped with the uniform measure on $\mathbb{S}^{n-1}$. This would give spectral gap $\lambda = 1/n$, and the unique invariant measures are the isotropic Gaussians.*

IV. Proofs

The hard work has already been done by Bennett, Carbery, Christ and Tao [1], Valdimarsson [18], and Carlen and Cordero-Erausquin [8]. We only need to point out how the ingredients fit neatly together. We only sketch the proofs due to space constraints.

Proof of Proposition 1. Let $(\mathbf{B}, \mathbf{c})$ be geometric. As observed in [8, Theorem 2.2], inspection of the duality argument that allows passage between (1) and (3) reveals that μ is an extremizer in (3) if and only if it admits a density f satisfying

$$f = \prod_{i=1}^{k} (f_i \circ B_i)^{c_i}, \tag{13}$$

where f_i denotes the density of $B_i \# \mu$. Moreover, if (13) holds, the marginal densities $(f_i)_{i=1}^k$ will be extremizers in (1). Now, the asserted splitting property can be obtained from the splitting property in Valdimarsson's characterization of extremizers for (1) in geometric settings [18]. □

With the identity (13) already noted, Theorem 1 follows easily.

Proof of Theorem 1. All extremizable data are equivalent to geometric data by a linear change of variables. Hence, by the data processing property of relative entropy, we may assume $(\mathbf{B},\mathbf{c})$ is geometric without any loss of generality.

To prove (4), it clearly suffices to assume $D(\nu\|\mu)<\infty$, since otherwise the claim is trivial; note that this implies $\nu\ll\mu$, and also $D(B_i\#\nu\|B_i\#\mu)<\infty$ for each i by the data processing inequality. Now, let $d\mu=fdx$, and write

$$\begin{aligned}D(\nu\|\mu)&=-h(\nu)+\int\log f d\nu\\&\geq-\sum_{i=1}^{k}c_i\left(h(B_i\#\nu)+\int\log(f_i\circ B_i)d\nu\right)\\&=-\sum_{i=1}^{k}c_i\left(h(B_i\#\nu)+\int\log(f_i)d(B_i\#\nu)\right)\\&=\sum_{i=1}^{k}c_iD(B_i\#\nu\|B_i\#\mu),\end{aligned}$$

where the first and last lines are definitions, the inequality follows from (3) and (13), and the penultimate line follows from the definition of pushforward. □

The standard program for deriving a spectral gap inequality from an entropy inequality is to linearize it to reveal the local behavior (see, e.g., [9]). Toward that end, recall that the relative entropy of $P\ll Q$ can be written as $D(P\|Q)=\int\frac{dP}{dQ}\log\left(\frac{dP}{dQ}\right)dQ$. Therefore, if P is a perturbation of Q in the sense that $dP=(1+\epsilon\varphi)dQ$ for a bounded function φ and ϵ sufficiently small, then Taylor expansion of $x\in\mathbb{R}^+\mapsto x\log x$ about $x=1$ gives the local behavior of relative entropy

$$D(P\|Q)=\frac{\epsilon^2}{2}\operatorname{Var}_Q(\varphi)+o(\epsilon^2),$$

where the first-order term is absent since φ necessarily satisfies $\int\varphi dQ=0$ for P to be a probability measure.

Proof of Theorem 2. It suffices to assume φ is bounded, since the general statement follows by localization. Thus, let $X\sim\mu$ be an extremizer in (3), assume $\int\varphi d\mu=0$ and define $d\mu_\epsilon:=(1+\epsilon\varphi)d\mu$, which is a valid probability measure for all ϵ sufficiently small. Definitions imply

$$d(B_i\#\mu_\epsilon)=(1+\epsilon\mathbb{E}[\varphi(X)|B_iX])d(B_i\#\mu),$$

where $\mathbb{E}[\varphi(X)|B_iX]$ is the conditional expectation of $\varphi(X)$ with respect to the σ-algebra generated by B_iX. Thus, by linearization and Theorem 1, we have

$$\begin{aligned}&\frac{\epsilon^2}{2}\sum_{i=1}^{k}c_i\operatorname{Var}(\mathbb{E}[\varphi(X)|B_iX])+o(\epsilon^2)\\&=\sum_{i=1}^{k}c_iD\left(B_i\#\mu_\epsilon\middle\|B_i\#\mu\right)\\&\leq D(\mu_\epsilon\|\mu)=\frac{\epsilon^2}{2}\operatorname{Var}(\varphi(X))+o(\epsilon^2).\end{aligned}$$

Dividing through by ϵ^2 and letting $\epsilon\downarrow 0$ completes the proof. □

In view of Proposition 1 and Theorem 2, Corollary 1 holds whenever μ is absolutely continuous with respect to Lebesgue measure and has finite entropy. It is straightforward to extend the statement to the case when either (or both) of these qualifications do not hold.

V. Closing Remarks

The duality between functional Brascamp–Lieb inequalities and their entropic form continues to hold in abstract settings. The transference principle of extremizers introduced here to obtain inequalities of the type (4) continues to apply. This suggests many interesting questions. For example, do the Shearer-type inequalities for non-product measures in [2], [5]–[7] fit into the context of Brascamp–Lieb-type inequalities on suitable spaces, as happens with $\mathbb{S}^n$ [9]? Does an approximate form of (4) hold when μ is a near-extremizer in a quantitative sense? Answers could lead to a systematic development of spectral gap inequalities for interesting classes of processes.

Acknowledgement: The author thanks Efe Aras, Pietro Caputo, and Max Fathi for stimulating conversations.

References

[1] J. Bennett, A. Carbery, M. Christ, and T. Tao. The Brascamp-Lieb inequalities: finiteness, structure and extremals. *Geometric and Functional Analysis*, 17(5):1343–1415, 2008.

[2] A. Blanca, P. Caputo, Z. Chen, D. Parisi, D. Štefankovič, E Vigoda. On Mixing of Markov Chains: Coupling, Spectral Independence, and Entropy Factorization. *Proc. of 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 3670-3692.

[3] H. J. Brascamp, E. H. Lieb, and J. M. Luttinger. A general rearrangement inequality for multiple integrals. *Journal of functional analysis*, 17(2):227–237, 1974.

[4] H. J. Brascamp and E. H. Lieb. Best constants in Young's inequality, its converse, and its generalization to more than three functions. *Advances in Mathematics*, 20(2):151–173, 1976.

[5] A. Bristiel and P. Caputo. Entropy inequalities for random walks and permutations. *arXiv preprint* arXiv:2109.06009 (2021).

[6] P. Caputo and D. Parisi. Block factorization of the relative entropy via spatial mixing. *Comm. in Mathematical Physics*, 388.2 (2021): 793-818.

[7] P. Caputo and A. Sinclair. Entropy production in nonlinear recombination models. *Bernoulli* 24(4B): 3246-3282, Nov. 2018.

[8] E. A. Carlen and D. Cordero-Erausquin. Subadditivity of the entropy and its relation to Brascamp–Lieb type inequalities. *Geometric and Functional Analysis*, 19(2):373–405, 2009.

[9] E. A. Carlen, E. H. Lieb, and M. Loss. A sharp analog of Young's inequality on S^N and related entropy inequalities. *The Journal of Geometric Analysis*, 14(3):487–520, 2004.

[10] F. R. Chung, R. L. Graham, P. Frankl, and J. B. Shearer. Some intersection theorems for ordered sets and graphs. *Journal of Combinatorial Theory, Series A*, 43(1):23–37, 1986

[11] T. A. Courtade. Bounds on the Poincaré constant for convolution measures. *Ann. Inst. H. Poincaré Probab. Statist.* 56 (1):566-579, 2020.

[12] A. Dembo, A. Kagan, and L. A. Shepp. Remarks on the maximum correlation coefficient. *Bernoulli*, pages 343–350, 2001.

[13] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics* 9:586-596, 1981.

[14] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pages 293–325, 1948

[15] E. H. Lieb. Gaussian kernels have only Gaussian maximizers. *Inventiones mathematicae*, 102(1):179–208, 1990.

[16] M. Madiman and A. Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Transactions on Information Theory*, 53(7):2317–2329, 2007.

[17] M. J. Steele, J. Michael. An Efron–Stein inequality for nonsymmetric statistics. *The Annals of Statistics* 14, no. 2: 753-758, 1986.

[18] S. I. Valdimarsson. Optimisers for the Brascamp-Lieb inequality. *Israel J. Math.*, 168:253–274, 2008.

On the Error Exponent Benefit of Sequentiality in Universal Binary Classification

Ching-Fang Li
GIEE, National Taiwan University, Taipei, Taiwan
Email: r10921046@ntu.edu.tw

I-Hsiang Wang
GICE, National Taiwan University, Taipei, Taiwan
Email: ihwang@ntu.edu.tw

***Abstract*—In the binary classification problem, the decision maker observes a testing sequence sampled from one of the two unknown distributions P_0 and P_1, along with two training sequences following each of the distributions. Since the two distributions are unknown, it is natural to impose universal guarantees on certain performances. The focus of this paper is a "semi-sequential" setup where the training sequences have fixed length and testing samples arrive sequentially, with universality constraints set on the expected stopping time and the type-I error exponent. The goal is to study the error exponent benefit under a competitive Neyman-Pearson criterion à la Levitan and Merhav [1], that is, the type-I error exponent is required to achieve a pre-set distribution-dependent constraint $\lambda(P_0, P_1)$ for all P_0, P_1. A novel upper bound on the optimal type-II error exponent is proved, showing that unlike the fully-sequential case considered by Hsu, Li, and Wang [2] where testing and training samples all arrive sequentially, in general there remains a trade-off between the type-I and type-II error exponents. We also propose a two-phase test and prove that it achieves the upper bound when $\lambda(P_0, P_1)$ is continuous and not greater than a Rènyi divergence of P_1 from P_0. The benefit of sequentiality is demonstrated by comparison with the cases where testing and training sequences are both of fixed length or both sequentially observed.**

I. Introduction

In a binary hypothesis testing problem, the decision maker observes a sequence of samples drawn i.i.d. from one of the two known distributions P_0 or P_1. The task is to decide from which distribution the sequence is generated and the performance is measured by the type-I and type-II error probabilities. Both error probabilities decay exponentially as the number of samples n goes to infinity, and the exponential rates are denoted as the error exponents. Blahut [3] characterized the optimal trade-off between the two exponents. When samples are taken sequentially and the decision maker is free to decide when to stop as long as the expected stopping time is less than a given constraint n, Wald's Sequential Probability Ratio Test (SPRT) [4] is shown to be optimal [5] and the error exponents can simultaneously achieve the two extremes, namely, the two KL divergences $\mathrm{D}(P_1 \| P_0)$ and $\mathrm{D}(P_0 \| P_1)$. Sequentiality in taking samples eradicates the trade-off in error exponents.

While the benefit of sequentiality of taking samples in hypothesis testing is well understood, a caveat is that the underlying distributions P_0 and P_1 are fully known to the decision maker. This is not the case in many real-world problems: instead of knowing the two distributions, the decision maker may only observe two training sequences generated from P_0 and P_1 respectively. Since the underlying distributions are unknown, it is natural to ask for a *universal* guarantee on certain performances. Ziv [6] considered this problem, assuming the number of training samples scales linearly with that of the testing samples and the ratio is $\alpha > 0$. He gave a formulation under a generalized Neyman-Pearson criterion, where a universality constraint is set on the type-I error exponent to be no less than a given constant $\lambda_0 > 0$ regardless of the underlying distributions. Subsequently Gutman [7] characterized the optimal type-II error exponent, which depends on the underlying distributions and the constant λ_0.

However, satisfying the universality constraint in [6], [7] on the type-I error exponent comes at a price: it can be shown that no matter how small λ_0 is, there always exists a pair of distinct distributions P_0, P_1 so close to each other that the type-II error probability tends to 1. Levitan and Merhav [1] proposed the competitive Neyman-Pearson criterion, where a distribution-dependent constraint $\lambda(P_0, P_1)$ replaces the constant λ_0. Under this criterion, they characterized the optimal type-II error exponent and proposed an asymptotically optimal test, and a trade-off between $\lambda(P_0, P_1)$ and the type-II error exponent is observed. They also established necessary and sufficient conditions on $\lambda(P_0, P_1)$, for which there exist tests with exponentially vanishing error probabilities. For such tests, the two error exponents are bounded by the two Rényi divergences $\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0)$ and $\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_0 \| P_1)$ respectively, analogous to $\mathrm{D}(P_1 \| P_0)$ and $\mathrm{D}(P_0 \| P_1)$ in Blahut's result [3].

In this work, we aim to investigate the benefit of sequentiality of taking samples in the binary classification problem under the competitive Neyman-Pearson criterion. Universality constraints are set on the expected stopping time and the type-I error exponent. Our focus is on the *semi-sequential* setup where the training sequences are fixed-length and the testing samples are taken sequentially. The optimal type-II error exponent is characterized when the type-I error exponent universality constraint $\lambda(P_0, P_1)$ is continuous in (P_0, P_1) and upper bounded by $\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0)$. For achievability, we propose a two-phase test. To ensure the expected stopping time is bounded by n, the test stops at $n-1$ if the type of the testing sequence is close enough to one of the type of the training sequences, similar to the Sequential Type Matching Test (STMT) in [2], and this is shown to happen with high probability. Otherwise, n testing samples are not enough, so it continues until n^2 testing samples are collected and uses the fixed-length test in [1].

To prove the optimality of our test, we provide two upper bounds on the type-II error exponent, and the minimum of the two upper bounds is shown to be achievable by our proposed test under the aforementioned condition on $\lambda(P_0, P_1)$. For the first bound we follow a standard argument and utilize the data processing inequality of KL divergence. The second bound is obtained by a reduction from a new composite hypothesis testing problem where the distribution of the testing samples is assumed to be known. By the results in [1], we have an upper bound on the performance of tests for the new problem. Since any tests that solve the binary classification problem can be used to solve the new problem and a correspondence in the error probabilities is clearly established, the upper bound can be transformed into a bound for the binary classification problem.

Finally, we show the benefit of sequentiality by comparing the optimal error exponents with those of fixed-length tests in [1] and fully-sequential (both testing and training samples arrive sequentially) tests in [2]. For fair comparison, we follow [2] and focus on tests with vanishing error probabilities under all possible distributions. Aside from numerical comparison, we also prove that there is a strict gap between the optimal error exponents of semi-sequential and fully-sequential tests if and only if $\alpha < 1$. In other words, in general there exists a trade-off between error exponents, which makes sense since the training sequences are still fixed-length. Nevertheless, the trade-off is completely eradicated when $\alpha \geq 1$.

Related works: Haghifam *et al.* [8] considered the semi-sequential classification problem as well. They proposed a test and showed that it achieves larger Bayesian error exponent over the fixed-length case. Under the same setting, Bai *et al.* [9] proposed an *almost fixed-length* two-phase test with performance lying between Gutman's fixed-length test and the semi-sequential test in [8]. However, in both [8] and [9], they did not have a universality constraint on the expected stopping time, nor other universal guarantees over all possible distributions. For example, in [8], the expected stopping time of their test depends implicitly on the unknown distributions P_0, P_1. Moreover, in order to achieve certain performance guarantees, parameters have to be chosen to satisfy some conditions that depend on the underlying distribution.

Notations: A finite-length sequence $(x_1, x_2, ..., x_n)$ is denoted as x^n. Logarithms are of base 2 if not specified. $\mathcal{P}(\mathcal{X})$ is the set of all probability distributions over alphabet $\mathcal{X}$. Given positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \doteq b_n$ if $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. The relation $\dot{\leq}$ is defined similarly. Also, $\mathbb{1}\{\cdot\}$ denotes the indicator function.

II. Problem Formulation

Let $\mathcal{X}$ be a finite alphabet with $|\mathcal{X}| = d \geq 2$ and consider the set of distributions bounded away from the boundary of the probability simplex. Specifically, fix some $\varepsilon > 0$ and let

$$\mathcal{P}_\varepsilon = \{P \in \mathcal{P}(\mathcal{X}) \,|\, \forall\, x \in \mathcal{X},\ P(x) \geq \varepsilon\}.$$

This assumption is made to ensure that the KL divergences between these distributions are bounded and uniformly continuous. Note that $\mathcal{P}_\varepsilon$ is compact. The underlying distributions are described by a pair of distinct distributions $(P_0, P_1) \in \mathcal{D}_\varepsilon = \{(P,Q) \,|\, P, Q \in \mathcal{P}_\varepsilon,\ P \neq Q\}$, and (P_0, P_1) is *unknown* to the decision maker. The decision maker observes a testing sequence $\{X_k\}$ consisting of i.i.d. samples following P_θ, where $\theta \in \{0,1\}$ is the unknown ground truth. Given $n \in \mathbb{N}$ and $\alpha > 0$, to learn about the unknown underlying distributions, the decision maker also has access to two training sequences T_0^N and T_1^N, where $N = \lceil \alpha n \rceil$, $T_{0,k} \overset{\text{i.i.d.}}{\sim} P_0$ and $T_{1,k} \overset{\text{i.i.d.}}{\sim} P_1$. The testing and training sequences are mutually independent.

The objective of the decision maker is to output $\hat{\theta} \in \{0,1\}$ as an estimation of the ground truth θ, based on the observed samples. A test is a pair $\Phi_n = (\tau_n, \delta_n)$ where $\tau_n \in \mathbb{N}$ is a Markov stopping time with respect to the filtration $\mathcal{F}_k = \sigma(X^k, T_0^N, T_1^N)$. We may write τ_n as τ when it is clear from the context. The decision rule $\delta_n : \mathcal{X}^\tau \times \mathcal{X}^N \times \mathcal{X}^N \to \{0,1\}$ is a $\mathcal{F}_\tau$-measurable function. For simplicity, denote the output of δ_n as $\hat{\theta}$. Note that here n can be viewed as an index of the problem, indicating the length of the training sequences. Moreover, Φ_n refers specifically to a test for the problem with $N = \lceil \alpha n \rceil$ samples in each training sequence. In the following, when X^k is observed, denote the empirical distribution (type) as $\hat{P}^k$, where $\hat{P}^k(x) = \frac{1}{k}\sum_{i=1}^k \mathbb{1}\{X_i = x\}$ for $x \in \mathcal{X}$. The empirical distributions of T_0^N and T_1^N are denoted as $\hat{P}_0^N$ and $\hat{P}_1^N$ (we omit N if it is clear from the context).

The above problem formulation is described as a *semi-sequential* version. The *fixed-length* setting can be viewed as restricting $\tau_n = n$. In the *fully-sequential* setting, the testing and training samples are all sequentially observed. At time k, there are k testing samples and $N_k = \lceil \alpha k \rceil$ training samples from each distribution. A test is similarly defined with the constant N replaced by a time-dependent variable N_k.

To evaluate the performance of tests, we consider the error probability and the number of samples used. Given $(P_0, P_1) \in \mathcal{D}_\varepsilon$ and $\theta \in \{0,1\}$, the error probability is defined as $\pi_\theta(\Phi_n|P_0, P_1) = \mathbb{P}_\theta\{\hat{\theta} \neq \theta\}$, where $\mathbb{P}_\theta$ is the shorthand notation for the joint probability law of the testing sequence and training sequences. The average number of samples used can be described by the expected stopping time $\mathbb{E}_\theta[\tau_n]$, where $\mathbb{E}_\theta$ denotes the expectation under $\mathbb{P}_\theta$.

Since the underlying distributions are unknown, it is natural to ask for some universal guarantees on the performance. The universality constraints are twofold. First, to compare with fixed-length tests, we set a *constraint on the expected stopping time* to be at most n under all possible distributions $(P_0, P_1) \in \mathcal{D}_\varepsilon$. The error exponents can be defined accordingly.

Definition 1 (Error Exponents)**.** Let $\{\Phi_n\}$ be a sequence of tests where Φ_n satisfies $\mathbb{E}_\theta[\tau_n] \leq n$ for all underlying distributions and ground truth θ. The type-I and type-II error exponents of $\{\Phi_n\}$ with $(P_0, P_1) \in \mathcal{D}_\varepsilon$ are defined as

$$e_\theta(P_0, P_1) = \liminf_{n\to\infty} -\frac{1}{n} \log \pi_\theta(\Phi_n|P_0, P_1), \quad \theta = 0, 1.$$

Second, a universality constrained is set on the type-I error exponent, where we adopt the competitive Neyman-Pearson criterion proposed in [1]. We focus on tests satisfying

$$e_0(P_0,P_1) \geq \lambda(P_0,P_1), \quad \forall\,(P_0,P_1)\in\mathcal{D}_\varepsilon, \tag{1}$$

where $\lambda:\mathcal{P}_\varepsilon\times\mathcal{P}_\varepsilon\to[0,\infty)$ is a function with positive value on $\mathcal{D}_\varepsilon$. Among all the tests satisfying (1), the goal is to find an optimal test that maximizes $e_1(P_0,P_1)$ uniformly over the underlying distributions.

III. Main Results

To present the results, we first introduce the Rényi Divergence. The Rényi Divergence of order $\frac{\alpha}{1+\alpha}$ of P from Q can be expressed as

$$\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P\|Q) = \min\nolimits_{V\in\mathcal{P}(\mathcal{X})}\{\mathrm{D}(V\|Q)+\alpha\mathrm{D}(V\|P)\}.$$

The results are summarized in the following theorem.

Theorem 1. *Let $\lambda:\mathcal{P}_\varepsilon\times\mathcal{P}_\varepsilon\to[0,\infty)$ be a function with positive value on $\mathcal{D}_\varepsilon=\{(P,Q)\,|\,P,Q\in\mathcal{P}_\varepsilon,\ P\neq Q\}$, and $\{\Phi_n\}$ be a sequence of semi-sequential tests such that*

- *for each $\Phi_n=(\tau_n,\delta_n)$, $\mathbb{E}_\theta[\tau_n]\leq n$ for all $(P_0,P_1)\in\mathcal{D}_\varepsilon$ and ground truth $\theta\in\{0,1\}$,*
- *the type-I error exponent satisfies* (1).

Then for any $(P_0,P_1)\in\mathcal{D}_\varepsilon$,

$$e_1(P_0,P_1)\leq\min\left\{\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_0\|P_1),\ \mu(P_0,P_1)\right\}, \tag{2}$$

where

$$\mu(P_0,P_1)=\inf_{\substack{Q_0,Q_1\in\mathcal{P}(\mathcal{X})\\ g(Q_0,Q_1)<0}}\Big(\alpha\mathrm{D}(Q_0\|P_0)+\alpha\mathrm{D}(Q_1\|P_1)\Big),$$

and $g(Q_0,Q_1)=$

$$\inf_{P_1'\in\mathcal{P}_\varepsilon\setminus\{P_1\}}\alpha\mathrm{D}(Q_0\|P_1)+\alpha\mathrm{D}(Q_1\|P_1')-\lambda(P_1,P_1').$$

If $\lambda(P_0,P_1)\leq\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$ for all $(P_0,P_1)\in\mathcal{D}_\varepsilon$, then there exist tests that satisfy the above universality constraints. Moreover, if λ is continuous, the upper bound (2) *can be achieved simultaneously for all $(P_0,P_1)\in\mathcal{D}_\varepsilon$.*

The proof of achievability and converse are delegated to Section IV and V respectively. Next we briefly state the results for the fixed-length setting [1] and fully-sequential setting [2], serving as comparisons.

Theorem 2 ([1, Theorem 1 & 2])**.** *Let $\lambda:\mathcal{P}_\varepsilon\times\mathcal{P}_\varepsilon\to[0,\infty)$ be a function with positive value on $\mathcal{D}_\varepsilon$, and $\{\Phi_n\}$ be a sequence of fixed-length tests satisfying* (1)*. Then for any $(P_0,P_1)\in\mathcal{D}_\varepsilon$, the type-II error exponent $e_1(P_0,P_1)$ is upper bounded by*

$$\inf_{\substack{Q,Q_0,Q_1\in\mathcal{P}(\mathcal{X})\\ g_1(Q,Q_0,Q_1)<0}}\Big(\mathrm{D}(Q\|P_1)+\alpha\mathrm{D}(Q_0\|P_0)+\alpha\mathrm{D}(Q_1\|P_1)\Big),$$

where $g_n(Q,Q_0,Q_1)=$

$$\inf_{(P_0',P_1')\in\mathcal{D}_\varepsilon}n\mathrm{D}(Q\|P_0')+\alpha\mathrm{D}(Q_0\|P_0')+\alpha\mathrm{D}(Q_1\|P_1')-\lambda(P_0',P_1').$$

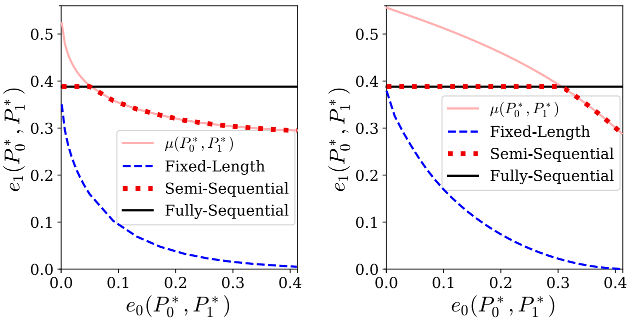
Moreover, using λ as a threshold function, there exist tests that satisfy (1) *and achieve the upper bound simultaneously for all $(P_0,P_1)\in\mathcal{D}_\varepsilon$.*

Theorem 3 ([2, Theorem 1])**.** *Let $\{\Phi_n\}$ be a sequence of fully-sequential tests such that*

- *for each $\Phi_n=(\tau_n,\delta_n)$, $\mathbb{E}_\theta[\tau_n]\leq n$ for all $(P_0,P_1)\in\mathcal{D}_\varepsilon$ and ground truth $\theta\in\{0,1\}$,*
- *the error probabilities vanish regardless of the underlying distributions as n goes to infinity,*

then for any $(P_0,P_1)\in\mathcal{D}_\varepsilon$, $e_0(P_0,P_1)\leq\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$ and $e_1(P_0,P_1)\leq\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_0\|P_1)$. Moreover, there exist tests that satisfy the above universality constraints and achieve these upper bounds simultaneously for all $(P_0,P_1)\in\mathcal{D}_\varepsilon$.

A sequence of tests is said to be *efficient* [1] or *universally exponentially consistent* [10] if the error probabilities decay to zero exponentially for all the underlying distributions as the number of samples grows to infinity. Theorem 3 implies that there exist fully-sequential tests that are universally exponentially consistent, and there is no need to set constraints on the type-I error exponent. In [1], a necessary and sufficient condition on $\lambda(\cdot,\cdot)$ for the existence of efficient fixed-length tests is provided. Also, it is shown that by taking a natural choice $\lambda(P_0,P_1)=\xi\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$ for some $0<\xi<1$, there exist efficient fixed-length tests. Since a fixed-length test can be viewed as a special case of the semi-sequential tests, we know that there indeed exist semi-sequential tests with exponentially decay error probabilities under all possible distributions, given some appropriately chosen $\lambda(\cdot,\cdot)$.



(a) $\min\{\lambda_0,\rho\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)\}$

(b) $\xi\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1\|P_0)$

Fig. 1: The optimal type-II error exponent under different choices of $\lambda(P_0,P_1)$. Here we fix $\mathcal{X}=\{0,1\}$, $\varepsilon=0.01$, $\alpha=0.7$, and choose $P_0^*=\mathrm{Ber}(0.4)$, $P_1^*=\mathrm{Ber}(0.9)$. In Figure 1a, ρ is chosen as 0.9999, and λ_0 increases from 0.001 to 0.42. In Figure 1b, ξ increases from 0.001 to 0.999.

To see the benefit of sequentiality, we focus on universally exponentially consistent tests. Fix some $(P_0^*,P_1^*)\in\mathcal{D}_\varepsilon$ and plot $e_1(P_0^*,P_1^*)$ versus $e_0(P_0^*,P_1^*)$, it is shown that the achievable region lies inside the rectangle bounded by two Rényi divergences. As stated in Theorem 3, fully-sequential tests can achieve the corner point. For fixed-length tests and semi-sequential tests, there is a trade-off between the two error exponents, and the trade-off curve depends on the function $\lambda(\cdot,\cdot)$. Specifically, given two different constraint functions $\lambda(\cdot,\cdot)$ and $\lambda'(\cdot,\cdot)$, even if $\lambda(P_0^*,P_1^*)=\lambda'(P_0^*,P_1^*)$, the corresponding optimal $e_1(P_0^*, P_1^*)$ may be different. We demonstrate by plotting the error exponent regions with respect to two different choices of $\lambda(\cdot,\cdot)$. The first one is modified from Gutman's setting [7]. We define $\lambda(P_0, P_1) = \min\{\lambda_0, \rho \mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0)\}$ with some $\lambda_0 > 0$ and $0 < \rho < 1$. The result for different constant λ_0 is shown in Figure 1a. Alternatively, in Figure 1b, we choose $\lambda(P_0, P_1) = \xi \mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0)$ for some $0 < \xi < 1$. It is clear that the trade-off curves are different in two figures.

In Figure 1a and Figure 1b, the performance of semi-sequential tests lies between those of fixed-length tests and fully-sequential tests. However, it turns out semi-sequential tests can achieve the same error exponents as fully-sequential tests under certain circumstances. A necessary and sufficient condition is provided in the following.

Proposition 1. *Semi-sequential tests can achieve the same error exponents as fully-sequential tests for all $(P_0, P_1) \in \mathcal{D}_\varepsilon$ if and only if $\alpha \geq 1$.*

Proposition 1 shows that if $\alpha < 1$, then for any semi-sequential tests, there exists some distributions $(P_0, P_1) \in \mathcal{D}_\varepsilon$ such that the point $\left(\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0), \mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_0 \| P_1)\right)$ is not achievable, and the trade-off between type-I and type-II error exponent remains. Meanwhile, when $\alpha \geq 1$, the trade-off is completed eradicated as in the fully-sequential case. The proof is given in Appendix A of the extended version [11].

IV. Achievability

We propose a test $\Phi_n = (\tau_n, \delta_n)$ for the semi-sequential problem. First consider the empirical distributions at time $n-1$, namely, $(\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in \mathcal{P}(\mathcal{X})^3$. In the fixed-length setting, in order to satisfy (1), the test should output 0 if the empirical distributions are close to some possible distributions (P_0', P_0', P_1') with $(P_0', P_1') \in \mathcal{D}_\varepsilon$. However, this will make the type-II error exponent small, so one should get more samples and decide later. Nevertheless, the expected stopping time should not exceed n, meaning the probability of taking more samples should be kept small. Observe that with high probability, the empirical distributions are close to the true underlying distributions. This happens if and only if the type of the testing sequence is close to the type of one of the training sequences. We measure the closeness with

$$\mathrm{GJS}(P, Q, \alpha) = \min_{V \in \mathcal{P}(\mathcal{X})} \{\mathrm{D}(Q \| V) + \alpha \mathrm{D}(P \| V)\},$$

α-weighted Generalized Jensen-Shannon Divergence (GJS). Let us now define two sets, for $i \in \{0, 1\}$,

$$\Lambda_i^n = \left\{(Q, Q_0, Q_1) \in \mathcal{P}(\mathcal{X})^3 \,\middle|\, \mathrm{GJS}(Q, Q_i, \alpha) < \eta_n \right\},$$

where $\eta_n = \left[(d+2)\log n + d \log(N+1)\right]/(n-1)$. By the method of types, with high probability, the empirical distributions will lie in these sets. Hence, the stopping time $\tau_n = n-1$ if $(\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in \Lambda_0^n \cup \Lambda_1^n$. Otherwise, $\tau_n = n^2$.

Next we specify the decision rule. When $\tau_n = n-1$, the decision rule is similar to the STMT in [2]:

$$\delta_n(X^{n-1}, T_0^N, T_1^N) = \begin{cases} 0 & \text{if } (\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in \Lambda_0^n, \\ 1 & \text{if } (\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in \Lambda_1^n. \end{cases}$$

When $\tau_n = n^2$, we use the fixed-length test in [1] for n^2 testing samples and N training samples with threshold function $\overline{\lambda}_n(P_0, P_1) = \lambda(P_0, P_1)/n$. As a result, we can ensure that the type-I error probability is of the same order as when $\tau_n = n-1$. Specifically, the decision rules is

$$\delta_n(X^{n^2}, T_0^N, T_1^N) = \mathbb{1}\left\{g_n(\hat{P}^{n^2}, \hat{P}_0, \hat{P}_1) \geq 0\right\}.$$

Now the test is clearly defined, we are ready to analyze its performance. First we show that the proposed test satisfies the universality constraint on the expected stopping time. Given any $(P_0, P_1) \in \mathcal{D}_\varepsilon$ and $\theta \in \{0, 1\}$, we have

$$\begin{aligned} \mathbb{E}_\theta[\tau_n] &\leq n - 1 + \mathbb{P}_\theta\left\{(\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in (\Lambda_\theta^n)^{\mathrm{c}}\right\} \times n^2 \\ &\leq n - 1 + n^d (N+1)^d \times 2^{-(n-1)\eta_n} \times n^2 = n, \end{aligned}$$

where the last inequality follows from the method of types.

Next we calculate the type-I error exponent. Based on the stopping time, the error events can be divided into two parts. When $\tau_n = n-1$, there is an error only if $(\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in \Lambda_1^n$. Since η_n vanishes as n goes to infinity, following the proof of error exponents in [2], it can be shown that

$$\mathbb{P}_0\left\{(\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in \Lambda_1^n\right\} \dot{\leq} 2^{-n\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0)}.$$

When $\tau_n = n^2$, by the method of types,

$$\begin{aligned} &\mathbb{P}_0\left\{\tau_n = n^2,\ \delta_n(X^{n^2}, T_0^N, T_1^N) = 1\right\} \\ &\leq (n^2+1)^d (N+1)^{2d} \times 2^{-n^2 \overline{\lambda}_n(P_0, P_1)} \doteq 2^{-n\lambda(P_0, P_1)}. \end{aligned}$$

Notice that under the assumption $\lambda(P_0, P_1) \leq \mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_1 \| P_0)$, the error probability is dominated by the second part, where the exponential rate is $\lambda(P_0, P_1)$. We conclude that $\{\Phi_n\}$ satisfies the universality constraint on the type-I error exponent (1).

For the type-II error exponent, we use the same method. When $\tau_n = n-1$, we have

$$\mathbb{P}_1\left\{(\hat{P}^{n-1}, \hat{P}_0, \hat{P}_1) \in \Lambda_0^n\right\} \dot{\leq} 2^{-n\mathrm{D}_{\frac{\alpha}{1+\alpha}}(P_0 \| P_1)}. \tag{3}$$

When $\tau_n = n^2$, by the method of types,

$$\mathbb{P}_1\left\{\tau_n = n^2,\ \hat{\theta} = 0\right\} \leq (n^2+1)(N+1)^{2d} \times 2^{-n\mu_n(P_0, P_1)},$$

where $\mu_n(P_0, P_1) =$

$$\inf_{\substack{Q, Q_0, Q_1 \in \mathcal{P}(\mathcal{X}) \\ g_n(Q, Q_0, Q_1) < 0}} n\mathrm{D}(Q \| P_1) + \alpha \mathrm{D}(Q_0 \| P_0) + \alpha \mathrm{D}(Q_1 \| P_1).$$

Lemma 1. *For $(P_0, P_1) \in \mathcal{D}_\varepsilon$, the sequence $\{\mu_n(P_0, P_1)\}$ is non-decreasing in n, and $\lim_{n \to \infty} \mu_n(P_0, P_1) = \mu(P_0, P_1)$.*

Proof. It is not hard to see that $\{\mu_n(P_0, P_1)\}$ is non-decreasing and bounded above, implying the convergence. Now, as n grows, to minimize $n\mathrm{D}(Q \| P_1) + \alpha \mathrm{D}(Q_0 \| P_0) + \alpha \mathrm{D}(Q_1 \| P_1)$, Q should be close to P_1, otherwise $n\mathrm{D}(Q \| P_1)$ gets too large. Also, to have $g_n(Q, Q_0, Q_1)$ less than 0, P_0' should be close to Q. Setting $P_0' = Q = P_1$ gives $\mu(P_0, P_1)$. The details are given in Appendix B of the extended version [11]. Notice that in some steps, we utilize the compactness of $\mathcal{P}_\varepsilon$ and the continuity of λ. □

By Lemma 1, we know the exponential error rate when $\tau_n = n^2$ is $\mu(P_0, P_1)$. Along with (3), the type-II error exponent is shown to achieve the upper bound (2).

V. Converse

In this section we show the upper bound on the type-II error exponent. For the first half of (2), the Rényi Divergence, we follow the proof of converse in [2] and replace the expected number of training samples $\mathbb{E}_\theta[N_\tau]$ by the deterministic N. For $\mu(P_0, P_1)$, we use a reduction from a fixed-length composite hypothesis testing problem. Intuitively, if we are allowed to drop the constraint on the expected stopping time and take infinitely many testing samples, P_θ can be fully known. In this situation, consider the following equivalent problem. Suppose a distribution $P \in \mathcal{P}_\varepsilon$ is *fixed* and *known*. The decision maker observes two independent fixed-length sequences T_0^N, T_1^N, where $N = \lceil \alpha n \rceil$ and the objective is to decide between the following two hypotheses:

$$\mathcal{H}_0 : T_{0,k} \overset{\text{i.i.d.}}{\sim} P,\ T_{1,k} \overset{\text{i.i.d.}}{\sim} \bar{P} \text{ for some } \bar{P} \text{ s.t. } (P, \bar{P}) \in \mathcal{D}_\varepsilon$$
$$\mathcal{H}_1 : T_{1,k} \overset{\text{i.i.d.}}{\sim} P,\ T_{0,k} \overset{\text{i.i.d.}}{\sim} \bar{P} \text{ for some } \bar{P} \text{ s.t. } (P, \bar{P}) \in \mathcal{D}_\varepsilon$$

Equivalently, using the language of composite hypothesis testing, we can view the two sequences as a sequence of pairs $\{(T_{0,k}, T_{1,k})\}_{k=1}^N$, and define the following two sets of product distributions:

$$\mathscr{P}_0 = \left\{ P\bar{P} \,\middle|\, (P, \bar{P}) \in \mathcal{D}_\varepsilon \right\} \text{ and } \mathscr{P}_1 = \left\{ \bar{P}P \,\middle|\, (P, \bar{P}) \in \mathcal{D}_\varepsilon \right\}.$$

Here we further emphasize the connection between this problem and the original problem. Given a sequence of semi-sequential tests $\{\Phi_n\}$ satisfying the universality constraints on the type-I error exponent, we can use it for this new composite hypothesis testing problem. Specifically, with the knowledge of P, we can generate the testing sequence. Along with the observed fixed-length sequences T_0^N, T_1^N, the test Φ_n will output a decision. One can now observe a correspondence between the error probabilities:

$$\begin{aligned} &\text{(original)} && \text{(new problem)} \\ \pi_0(\Phi_n | P_0, P_1) &= \pi_0(\Phi_n | P = P_0, \bar{P} = P_1) &&= \pi_0(\Phi_n | P_0 P_1), \\ \pi_1(\Phi_n | P_0, P_1) &= \pi_1(\Phi_n | P = P_1, \bar{P} = P_0) &&= \pi_1(\Phi_n | P_0 P_1). \end{aligned}$$

The above method gives us a randomized test, yet theoretically it can be easily derandomized and the error probabilities are at most twice the original, so it would not affect the exponential rate. Given $P, \bar{P}$, the error exponents in the new problem are defined as

$$e_0(P\bar{P}) := \liminf_{n\to\infty} \frac{-\log \pi_0(\Phi_n | P\bar{P})}{N} = \frac{e_0(P,\bar{P})}{\alpha}, \tag{4}$$
$$e_1(\bar{P}P) := \liminf_{n\to\infty} \frac{-\log \pi_1(\Phi_n | \bar{P}P)}{N} = \frac{e_1(\bar{P},P)}{\alpha}. \tag{5}$$

Since $\{\Phi_n\}$ satisfies the universality constraints on the type-I error exponent and by (4), for any fixed $P \in \mathcal{P}_\varepsilon$, we have $e_0(P\bar{P}) \geq \lambda(P, \bar{P})/\alpha$, for all $P\bar{P} \in \mathscr{P}_0$. Using the results in [1], we know that $e_1(\bar{P}P) \leq \inf_{Q \in \mathcal{Q}} \mathrm{D}\left(Q \middle\| \bar{P}P\right)$, where $\mathcal{Q}$ is the following set of distributions

$$\left\{ Q \in \mathcal{P}(\mathcal{X}^2) \,\middle|\, \inf_{P\bar{P}' \in \mathscr{P}_0} \left(\mathrm{D}\left(Q \middle\| P\bar{P}'\right) - \tfrac{\lambda(P,\bar{P}')}{\alpha} \right) < 0 \right\}.$$

For a distribution $Q \in \mathcal{P}(\mathcal{X}^2)$, let $Q_0, Q_1 \in \mathcal{P}(\mathcal{X})$ denote its marginal distributions. Since $\bar{P}P$ is a product distribution, by the chain rule of KL divergence,

$$\mathrm{D}\left(Q \middle\| \bar{P}P\right) = \mathrm{D}\left(Q \| Q_0 Q_1\right) + \mathrm{D}\left(Q_0 \middle\| \bar{P}\right) + \mathrm{D}\left(Q_1 \| P\right).$$

As $\mathrm{D}(Q \| Q_0 Q_1) \geq 0$, we can restrict Q to be a product distribution in the above results. Combining with (5) leads us to the upper bound $\mu(P_0, P_1)$.

Note that when deriving this bound, we assume infinite testing samples, which is closer to the reality when $\alpha < 1$. This provides an intuition of Proposition 1, explaining why $\mu(P_0, P_1)$ is a more accurate bound when $\alpha < 1$.

VI. Concluding Remarks

In this work, we mainly focus on universally exponentially consistent tests. Also in Theorem 1, it is required that λ is upper bounded by the Rényi Divergence, which makes it impossible to choose λ as a constant. However, in the fixed-length setting [1], there is no such limitation. An interesting question is what the optimal type-II error exponents will be if we drop this requirement. Similarly, in the fully-sequential setting, what will happen if we replace the constraint on vanishing error probabilities with the universality constraint on the type-I error exponent? It turns out the optimal error exponents can be characterized in both the above scenarios. As a result, the benefit of sequentiality can also be observed in the case without exponential consistency. In particular, we can take $\lambda(\cdot, \cdot) = \lambda_0$ and compare the type-II error exponent with Gutman's result. The proof involves some modifications in both the achievability and converse, which will be presented in a future paper.

References

[1] E. Levitan and N. Merhav, "A competitive neyman-pearson approach to universal hypothesis testing with applications," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2215–2229, 2002.

[2] C.-Y. Hsu, C.-F. Li, and I.-H. Wang, "On universal sequential classification from sequentially observed empirical statistics," in *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 642–647.

[3] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 405–417, July 1974.

[4] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, June 1945.

[5] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, September 1948.

[6] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Transactions on Information Theory*, vol. 34, no. 2, pp. 278–286, March 1988.

[7] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, March 1989.

[8] M. Haghifam, V. Y. F. Tan, and A. Khisti, "Sequential classification with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 3095–3113, May 2021.

[9] L. Bai, J. Diao, and L. Zhou, "Achievable error exponents for almost fixed-length binary classification," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 1336–1341.

[10] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal sequential outlier hypothesis testing," *Sequential Analysis*, vol. 36, no. 3, pp. 309–344, 2017.

[11] C.-F. Li and I.-H. Wang, "On the error exponent benefit of sequentiality in universal binary classification," *Preprint*, 2023, http://homepage.ntu.edu.tw/~ihwang/Eprint/izs24usc.pdf.

Batches Stabilize the Minimum Norm Risk in High Dimensional Overparameterized Linear Regression

Shahar Stein Ioushua, Inbar Hasidim, Ofer Shayevitz and Meir Feder
Tel Aviv University
email: {steinioushua, inbarhasidim}@mail.tau.ac.il, ofersha@eng.tau.ac.il, meir@tauex.tau.ac.il

Abstract—**Learning algorithms that divide the data into batches are prevalent in many machine-learning applications, typically offering useful trade-offs between computational efficiency and performance. In this paper, we examine the benefits of batch-partitioning through the lens of a minimum-norm overparameterized linear regression with isotropic Gaussian features. We suggest a natural small-batch version of the minimum-norm estimator, and derive an upper bound on its quadratic risk, showing it is inversely proportional to the noise level and to the overparameterization ratio, for the optimal choice of batch size. In contrast to minimum-norm, our estimator admits a stable risk behavior that is monotonically increasing in the overparameterization ratio, eliminating both the blowup at the interpolation point and the double-descent phenomenon. Interestingly, we observe that this implicit regularization offered by the batch partition is partially explained by feature overlap between the batches. Our bound is derived via a novel combination of techniques, in particular normal approximation in the Wasserstein metric of noisy projections over random subspaces.**

I. Introduction

Batch-based algorithms are used in various machine-learning problems. Particularly, partition into batches is natural in distributed settings, where data is either collected in batches by remote sensors who can send a small number of bits to a central server, or collected locally but offloaded to multiple remote workers for computational savings, see e.g. [1], [2]. Learning in batches is also employed in centralized settings; this is often done to reduce computational load, but is also known (usually empirically) to sometimes achieve better convergence, generalization, and stability, see e.g., [3], [4]. One of the most basic and prevalent learning tasks is linear regression, which has been extensively studied in both centralized and distributed settings. Linear regression is of particular contemporary interest in the overparameterized regime, where the number of parameters exceeds the number of samples. In this regime there are infinitely many interpolators, and a common regularization method is to pick the minimum norm (min-norm) solution, i.e., the interpolator whose ℓ_2 norm is minimal. However, this method requires inverting a matrix whose dimensions are the number of samples, a task that can be computationally costly and also result in a non-stable risk, growing unbounded close to the interpolation point [5], [6]. Performing linear regression separately in batches and combining the solutions (usually by averaging) can help with the computational aspects, and has been studied before mainly for large (linear in the number of samples) batches [2]. However, such solutions break down and cannot control the risk for sublinear batch size; they also shed no light on the performance benefits heuristically known to be offered by small batches. Can the min-norm solution benefit more from small batch partitioning? We answer this question in the affirmative, by suggesting a simple and natural min-norm-based small-batch regression algorithm, and showing it stabilizes the min-norm risk. We discuss the ramifications of our result in several settings.

Our contribution. We consider a linear model with isotropic features, in the overparameterized regime with n data samples and $p > n$ parameters, where the $n \times p$ feature matrix is i.i.d. Gaussian. The risk attained by min-norm in this setting and related ones was previously analyzed in [6]. Here, we suggest the following small batch variation of min-norm. First, the data is partitioned into small disjoint batches of equal size b, and a simple min-norm estimator is computed separately for each batch. Then, the resulting $\frac{n}{b}$ weak estimators are pooled together to form a new $\frac{n}{b} \times p$ feature matrix for a modified "linear" model, with suitably weighted modified samples. The modified model is not truly linear, since both the new features and noise depend on the parameter. Finally, a min-norm estimator is computed in this new setting, yielding our suggested batch minimum norm (batch-min-norm) estimator. While the modified model is far more overparameterized than the original model (by a factor of b), its features are now favorably correlated with the underlying parameters. We shall see that this trade-off can be beneficial.

To that end, we derive an upper bound on the risk obtained by our estimator, in the limit of $n, p \to \infty$ with a fixed overparameterization ratio $\gamma = p/n$, as a function of the SNR. Our bound is compared to simulations and is demonstrated to be quite tight. We then analytically find the batch size minimizing the bound, and show that it is inversely proportional to both γ and SNR; in particular, there is a low-SNR threshold point below which increasing the batch size (after taking $n, p \to \infty$) is always beneficial (albeit at very low SNR we can do worse than the null solution), see Figure 1. Unlike min-norm, and similarly to optimally-tuned ridge regression [7], the risk attained by batch-min-norm is generally stable; it is monotonically increasing in γ, does not explode near the interpolation point $\gamma = 1$, and does not exhibit a *double descent* phenomenon [6] (all this assuming $\mathsf{SNR} \geq 1$, see Figure 2). It is (trivially) always at least as good (and often much better) than min-norm. Another interesting observation is that the batch algorithm exactly coincides with the regular min-norm algorithm for any batch size, whenever the feature matrix has orthogonal rows. Thus, somewhat intriguingly, the reason that batches are useful can be partially attributed to the fact that feature vectors are slightly linearly dependent between batches, i.e., there is small overlap between the subspaces spanned by the batches. From a technical perspective, as we shall later see, this overlap implicitly regularizes the noise amplification suffered by the standard min-norm.

The derivation of the upper bound is the main technical contribution of the paper. The main difficulty lies in the second step of the algorithm, namely analyzing the min-norm with the feature matrix comprised of per-batch min-norm estimators. This step is no longer under a standard linear model, since the new feature matrix and the corresponding noise vector depend on the parameters, in a generally non-linear way. This poses a significant technical barrier requiring the use of several nontrivial mathematical tools. To compute the bias of the algorithm, we first write it as a recursive perturbed-

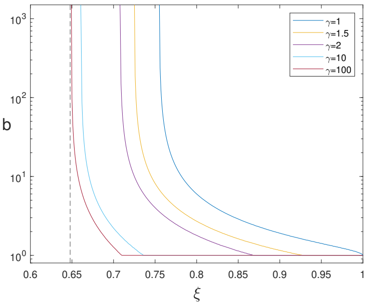


Fig. 1: *Optimal batch size vs. normalized signal-to-noise ratio* $\xi = \frac{\mathsf{SNR}}{1+\mathsf{SNR}}$. *When* $\xi < 0.6478$, *the optimal batch size* $b \to \infty$, *for any* $\gamma > 1$.

projection onto a random per-batch subspace, drawn from the Haar measure on the Stiefel manifold. We show that the statistics of these projections are asymptotically close in the Wasserstein metric to i.i.d. Gaussian vectors, a fact that allows us to obtain a recursive expression for the bias as batches are being added, with a suitable control over the error term. We then translate this recursion into a differential equation, whose solution yields the asymptotic expression for the bias. The full version of this paper, which includes further discussions, full proofs and numerical experiments, can be found in [8].

A. *Ramifications*

Distributed linear regression. In this setting, the goal is usually to offload the regression task from the main server by distributing it between multiple workers; the main server then merges the estimates given by the workers. This merging is typically done by simple averaging, e.g., [1], with the number of workers large but fixed (i.e., the batch size is linear in the sample size). The regime of sublinear batch sizes is less explored in the literature, perhaps due to practical reasons. When b is sublinear, the server-averaging approach breaks down since its risk is trivially dominated by the per-batch bias, and hence it attains the null risk asymptotically. In contrast, our algorithm projects the modified observations onto the subspace spanned by the entire collection of weak estimators, therefore resulting in much less bias than each weak estimator separately. Hence, our algorithm is far superior to server-averaging for fixed batch size. Numerical results indicate that this is also true in the general sublinear regime.

Mini-batch learning, High-dimensional overparameterized linear regression is known to sometimes serve as a reasonable proxy (via linearization) to more complex settings such as deep neural networks [9]. Furthermore, min-norm is equivalent to full Gradient Descent (GD) in linear regression, and even exhibits similar behavior observed when using GD in complex models, e.g. the double-descent [6]. Learning using small batches is a common approach that originated from computational considerations [10], but was also observed to improve generalization [11]. There is hence clear impetus to study the impact of small batches on min-norm-flavor algorithms in the linear regression setting. Indeed, mini-batch SGD for linear regression has been recently studied in [12], who gave closed-form solutions for the risk in terms of Volterra integral equations. In contrast to a practically observed phenomenon in deep networks, [12] showed that the linear regression risk of mini-batch SGD does not depend on the batch size b as long as $b \ll n^{1/5}$. Our batch-min-norm algorithm, while clearly not equivalent to mini-batch SGD in the linear regression setting, does exhibit the small-batch gain phenomenon, and hence could perhaps shed some light on similar effects empirically observed in larger models. In particular, the batch regularization effect we observe can be traced back to a data "overlap" between the batches, and it is interesting to explore whether this effect manifests itself in other settings. Moreover, from a high-level perspective, our algorithm "summarizes" each batch to create a "representative sample", and then trains again via GD only on these representative samples. It is interesting to explore whether this approach can be rigorously generalized to more complex models.

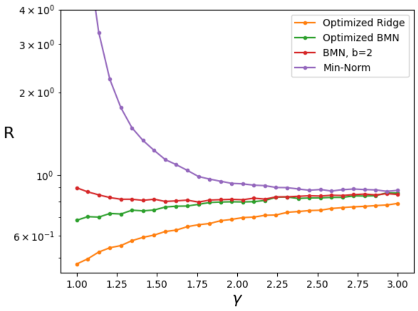


Fig. 2: *Risk of batch-min-norm with optimal batch size vs. overparametrization ratio* γ *with* $\xi = 0.7$. *Optimized ridge is ridge regression with the optimal regularization parameter.*

II. Preliminaries

A. *Notation*

We denote by x, $\boldsymbol{x}$, scalars and vectors, respectively. Vectors can be either row or column, which will be clear from context. We use X, $\boldsymbol{X}$ to denote random variables, and random vectors or matrices, respectively. The ℓ_2 norm of $\boldsymbol{X}$ is denoted $\|\boldsymbol{X}\|_2$ (and sometimes $\|\boldsymbol{X}\|$). For b orthogonal random vectors $\boldsymbol{U}_1, \cdots, \boldsymbol{U}_b \in \mathbb{R}^n$, $b \leq n$, with unit norm we say that $\{\boldsymbol{U}_i\}$ were uniformly drawn if the matrix $\boldsymbol{U} = [\boldsymbol{U}_1, \cdots, \boldsymbol{U}_b]$ was drawn from the Haar measure on the Stiefel manifold $\mathbb{V}_b(\mathbb{R}^n) \triangleq \{\boldsymbol{A} \in \mathbb{R}^{b\times n} : \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_b\}$, where $\boldsymbol{I}_b$ is the $b \times b$ identity matrix. We sometimes drop the subscript and write $\boldsymbol{I}$, when the dimension is clear from the context.

B. *Wasserstein Distance*

The p-Wasserstein distance between two probability measures μ and ν on $\mathbb{R}^n$ is

$$\mathcal{W}_p(\mu, \nu) \triangleq \left(\inf \mathbb{E}\|\boldsymbol{X} - \boldsymbol{Y}\|_2^p\right)^{\frac{1}{p}}, \tag{1}$$

where the infimum is taken over all random vector pairs $(\boldsymbol{X}, \boldsymbol{Y})$ with marginals $\boldsymbol{X} \sim \mu$ and $\boldsymbol{Y} \sim \nu$. Throughout this paper, we say Wasserstein distance to mean the 1-Wasserstein distance $\mathcal{W}_1(\mu, \nu)$, unless explicitly mentioned otherwise. With a slight abuse of notations, we write $\mathcal{W}_1(\boldsymbol{X}, \boldsymbol{Y})$ to indicate the Wasserstein distance between the corresponding probability measures of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The Wasserstein distance plays a key role in our proofs, mainly due to the following facts. First, our batch-min-norm algorithm performs projections onto small random subspaces. Wasserstein distance can be used to quantify how far these are from projections onto i.i.d. Gaussian vectors (Theorem 2 in

[8]). Second, closeness in Wasserstein distance implies change-of-measure inequalities for expectations of Lipschitz functions via the famous Kantorovich-Rubinstein duality theorem, which allows us to compute expectations in the Gaussian domain with proper error control. For further details on the properties of Wasserstein distance, see [8].

III. Problem setup

Let $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ be a data samples vector obtained from the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W} \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown parameters, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a given feature matrix with i.i.d. standard Gaussian entries, and $\boldsymbol{W} \in \mathbb{R}^n$ is a noise vector independent of $\boldsymbol{X}$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Here, the norm $\|\boldsymbol{\beta}\|_2 = r$ is assumed unknown unless otherwise stated. Define the overparametrization ratio $\gamma \triangleq \frac{p}{n}$. When $\gamma > 1$ we call the problem *overparameterized* and when $\gamma < 1$ we say it is *underparametrized*. An *estimator* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{Y}, \boldsymbol{X})$ for $\boldsymbol{\beta}$ from the samples and features is a mapping $\hat{\boldsymbol{\beta}} : \mathbb{R}^n \times \mathbb{R}^{n\times p} \to \mathbb{R}^p$. We measure the performance of an estimator via the quadratic (normalized) *risk* $R(\hat{\boldsymbol{\beta}}) \triangleq \frac{1}{r^2}\,\mathbb{E}\|\hat{\boldsymbol{\beta}}(\boldsymbol{Y}, \boldsymbol{X}) - \boldsymbol{\beta}\|^2$ it attains. Note that while $R(\hat{\boldsymbol{\beta}})$ is the estimation risk, it is also equal in this case (up to a constant) to the associated prediction risk, namely the mean-squared prediction error $\mathbb{E}\|\mathbf{x}^T\hat{\boldsymbol{\beta}} - \mathbf{x}^T\boldsymbol{\beta}\|^2$ when using $\hat{\boldsymbol{\beta}}$ to estimate the response to a new i.i.d. feature vector $\mathbf{x}$.

A. Minimum-Norm Estimation

In the overparametrized case $\gamma > 1$, there is an infinite number of solutions to the linear model $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$, and in order to choose one we need to impose some regularization. A common choice is the ℓ_2-norm regularization, which yields the *min-norm estimator*, defined as

$$\hat{\boldsymbol{\beta}}_{\text{MN}} \triangleq \operatorname{argmin} \|\boldsymbol{\beta}\|_2^2, \ \text{s.t. } \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}, \tag{3}$$

and explicitly given by

$$\hat{\boldsymbol{\beta}}_{\text{MN}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}. \tag{4}$$

The risk of the min-norm estimator is then

$$R(\hat{\boldsymbol{\beta}}_{\text{MN}}) = (1 - \tfrac{1}{r^2}\mathbb{E}\,\|\boldsymbol{L}\boldsymbol{\beta}\|^2) + \tfrac{1}{r^2}\mathbb{E}[\boldsymbol{W}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{W}], \tag{5}$$

where $\boldsymbol{L} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}$ is the orthogonal projection onto the row space of $\boldsymbol{X}$, and we used the fact that the matrix $\boldsymbol{X}$ is orthogonal to its null space $\boldsymbol{I} - \boldsymbol{L}$. The first term in the above is the (normalized) *bias* of the estimator, which represents the part of $\boldsymbol{\beta}$ that is not captured in the subspace spanned by $\boldsymbol{X}$. The second term is the (normalized) *variance* of the estimator. As previously shown in [6], the asymptotic risk of the min-norm estimator under the above model is

$$\lim_{p\to\infty} R(\hat{\boldsymbol{\beta}}_{\text{MN}}) = 1 - \gamma^{-1} + \frac{1-\xi}{\xi}\cdot\frac{1}{\gamma - 1}. \tag{6}$$

where $\xi \triangleq \frac{r^2}{r^2+\sigma^2} = \frac{\mathsf{SNR}}{1+\mathsf{SNR}}$ (also known as the Wiener coefficient), and $\mathsf{SNR} \triangleq \frac{r^2}{\sigma^2}$ is the *normalized* SNR.

IV. Batch Minimum-Norm Estimation

We proceed to suggest and study a natural batch version of the min-norm estimator. Let us divide the samples $\boldsymbol{Y}$ to n/b batches of some fixed size b, and denote by $\boldsymbol{Y}_j \in \mathbb{R}^b$ the jth batch. From each batch $\boldsymbol{Y}_j$ we can obtain a min-norm estimate of $\boldsymbol{\beta}$, given by

$$\hat{\boldsymbol{\beta}}_j \triangleq \boldsymbol{X}_j^T(\boldsymbol{X}_j\boldsymbol{X}_j^T)^{-1}\boldsymbol{Y}_j, \quad j = 1, \cdots, n/b, \tag{7}$$

where $\boldsymbol{X}_j$ are the feature vectors that correspond to $\boldsymbol{Y}_j$. We now have n/b weak estimators for $\boldsymbol{\beta}$, each predicting only a tiny portion of $\boldsymbol{\beta}$'s energy. However, these estimators are clearly better correlated with $\boldsymbol{\beta}$ compared to random features. Hence, it makes sense to think of each $\hat{\boldsymbol{\beta}}_j^T$ as a *modified feature vector* $\boldsymbol{x}'_j$ that summarizes what was learned from batch j. Since each modified feature $\boldsymbol{x}'_j$ is a linear combination the jth batch's feature vectors, with coefficients $\boldsymbol{Y}_j^T \cdot (\boldsymbol{X}_j\boldsymbol{X}_j^T)^{-1}$, we can similarly construct the corresponding jth *modified sample*

$$Y'_j \triangleq \boldsymbol{Y}_j^T \cdot (\boldsymbol{X}_j\boldsymbol{X}_j^T)^{-1}\boldsymbol{Y}_j = \boldsymbol{x}'_j\boldsymbol{\beta} + W'_j, \tag{8}$$

where $W'_j = \boldsymbol{Y}_j^T(\boldsymbol{X}_j\boldsymbol{X}_j^T)^{-1}\boldsymbol{W}_j$ is the *modified noise*.

We can now pool all these modified quantities to form a new model:

$$\boldsymbol{Y}' = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{W}', \tag{9}$$

where the *modified feature matrix* $\boldsymbol{X}'$ and *modified noise vector* $\boldsymbol{W}'$ are given by $\boldsymbol{X}' = [\boldsymbol{x}'^T_1, \cdots, \boldsymbol{x}'^T_{n/b}]^T$, and $\boldsymbol{W}' = [W'_1, \cdots, W'_{n/b}]^T$. Of course, the above is not truly a linear model, since both the matrix $\boldsymbol{X}'$ and the noise $\boldsymbol{W}'$ depend on the parameter $\boldsymbol{\beta}$. But we can nevertheless naturally combine all the batch estimators by simply applying min-norm estimation to (9). This yields our suggested batch-min-norm estimator:

$$\hat{\boldsymbol{\beta}}_{\text{BMN}} \triangleq \boldsymbol{X}'^T(\boldsymbol{X}'\boldsymbol{X}'^T)^{-1}\boldsymbol{Y}'. \tag{10}$$

The risk of $\hat{\boldsymbol{\beta}}_{\text{BMN}}$ is then given by

$$R(\hat{\boldsymbol{\beta}}_{\text{BMN}}) = (1 - \tfrac{1}{r^2}\mathbb{E}\,\|\boldsymbol{L}'\boldsymbol{\beta}\|^2) + \tfrac{1}{r^2}\mathbb{E}[\boldsymbol{W}'^T(\boldsymbol{X}'\boldsymbol{X}'^T)^{-1}\boldsymbol{W}'], \tag{11}$$

where $\boldsymbol{L}'$ is now the projection operator onto the subspace spanned by the rows of $\boldsymbol{X}'$, again using the fact that $\boldsymbol{X}'$ is orthogonal to its null space $\boldsymbol{I} - \boldsymbol{L}'$. We can see that the first and second terms in (11) are the (normalized) bias and variance of $\hat{\boldsymbol{\beta}}_{\text{BMN}}$, denoted $\mathsf{Bias}(\hat{\boldsymbol{\beta}}_{\text{BMN}})$ and $\mathsf{Var}(\hat{\boldsymbol{\beta}}_{\text{BMN}})$, respectively. Unlike in the min-norm estimator case, the rows of $\boldsymbol{X}'$ depend on the parameter $\boldsymbol{\beta}$, and the noise $\boldsymbol{W}'$ depends on $\boldsymbol{X}'$, which makes the analysis of the risk significantly more challenging.

It is interesting to point out that if $\boldsymbol{X}$ happens to have orthogonal rows, then $\boldsymbol{L} = \boldsymbol{L}'$, which means the bias of the batch estimator coincides with that of min-norm. Moreover, in this case, the variance of both batch- and regular min-norm is simply the variances of the noises $\boldsymbol{W}'$ and $\boldsymbol{W}$, respectively, which are identical. Therefore, the risk of both estimators coincides in the orthogonal case, for any batch size. However, as we show in the next section, in the general case the risk can benefit from batch partition. This suggests that the gain of batch-min-norm can be partially attributed to the linear dependence between the feature vectors in different batches.

V. Main Result

Our main result is an upper bound on the risk of batch-min-norm. Throughout the paper limits are taken as $n, p \to \infty$ and $\gamma = p/n$ held fixed.

Theorem 1. *For any $\gamma > 1/b$, the asymptotic risk of batch-min-norm is upper bounded by*

$$\lim_{p\to\infty} R(\hat{\boldsymbol{\beta}}_{\text{BMN}}) \le \frac{\gamma b - 1}{\gamma b + (b-1)\xi} + \frac{(1-\xi)(b-(b-1)\xi)}{\xi(\gamma b - 1)}, \tag{12}$$

where the first addend upper bounds the asymptotic bias and the second addend upper bounds the asymptotic variance.

Note that while we are interested in the $\gamma > 1$ regime, our bound applies verbatim to $\gamma > 1/b$. Loosely speaking, the reason is that working in batches "shifts" the interpolation point from 1 to $1/b$, similarly to what would happen if we naively discarded all but n/b samples and applied min-norm to the remaining ones (batch-min-norm is superior to this naive approach, as demonstrated in [8, Section 7]).

Our upper bound turns out to be quite tight for a wide range of parameters (see [8, Section 7]). It can therefore be used to obtain a very good estimate for the optimal batch size (which we therefore loosely refer to as "optimal" in the sequel), by minimizing (12) with respect to b. To that end, we need to assume that the SNR (namely ξ) is known; this is often a reasonable assumption, but otherwise the SNR can be estimated well from the data for almost all $\boldsymbol{\beta}$ (see, implicitly, in [6, Section 7]). The minimization yields an explicit formula for the optimal batch size as a function of the overparameterization ratio γ and the SNR, see [8, Subsection 5.1]. In particular, it can be analytically verified that the optimal batch size is inversely proportional to both γ and SNR; more specifically, there is a low-SNR threshold point below which increasing the batch size (after taking $n, p \to \infty$) is always beneficial. This can be seen in Figure 1, which plots the optimal batch size for different values of SNR and γ. For further discussion see [8, Subsection 5.1].

Let us briefly outline the proof of Theorem 1. We start with the bias of the algorithm, and write it as a recursive relation, where a single new batch is added each time, and its expected contribution to the bias reduction is quantified. In a nutshell, we keep track of the projection of $\boldsymbol{\beta}$ onto the complementary row space of the modified feature vectors from all preceding batches. We then write the batch's contribution as a function of the inner products between the (random) batch's basis vectors and the basis of that space. We show that this collection of inner products is close in Wasserstein distance to a Gaussian vector with independent entries, and derive an explicit recursive rule for the bias as a function of the number of batches processed, under this approximating Gaussian distribution. This function is then shown to be Lipschitz in a region where most of the distribution is concentrated, which facilitates the use of Wasserstein duality to show that the recursion rule is asymptotically correct under the true distribution. Finally, we convert the recursive rule into a certain differential equation, whose solution yields the bias bound. This is done in Section VI.

To bound the variance, we note that the jth modified sample Y'_j, features vector $\boldsymbol{x}'_j$, and noise W'_j, are all linear combinations of the corresponding batch elements, $\boldsymbol{Y}_j$, $\boldsymbol{X}_j$ and $\boldsymbol{W}_j$, with the same (random) coefficients. These coefficients converge almost surely to the original samples $\boldsymbol{Y}_j$. We use this to show that the variance converges to that of a Gaussian mixture noise with χ^2-distributed weights that is projected onto the rows of a Wishart matrix. This part of the proof can be found in [8].

VI. Proof of Main Result - Bias Part

In order to estimate the asymptotic bias, we rewrite the $\mathsf{Bias}(\hat{\boldsymbol{\beta}}_{\text{BMN}})$ term in (11) as a recursive equation where at the jth step we add the jth batch $\boldsymbol{Y}_j$, that corresponds to the matrix rows $\boldsymbol{X}_j$, and update the contribution of this batch to the overall projection. Recall that $\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{W}_j = \boldsymbol{X}_j(\boldsymbol{\beta} + \boldsymbol{Z}_j)$, with $\boldsymbol{Z}_j \triangleq \boldsymbol{X}_j^T(\boldsymbol{X}_j\boldsymbol{X}_j^T)^{-1}\boldsymbol{W}_j$. with $\boldsymbol{Z}_j \triangleq \boldsymbol{X}_j^T(\boldsymbol{X}_j\boldsymbol{X}_j^T)^{-1}\boldsymbol{W}_j$. Then, the jth row in the modified feature matrix $\boldsymbol{X}'$ is $\boldsymbol{x}'_j = (\boldsymbol{\beta}^T + \boldsymbol{Z}_j^T)\boldsymbol{D}_j$, with $\boldsymbol{D}_j$ the projection matrix onto the row space of $\boldsymbol{X}_j$. Denote by $\boldsymbol{X}'_j = [\boldsymbol{x}'^T_1, \cdots, \boldsymbol{x}'^T_j]^T$ the modified feature matrix after the first j steps and let $\boldsymbol{L}'_j \triangleq \boldsymbol{X}'^T_j(\boldsymbol{X}'_j\boldsymbol{X}'^T_j)^{-1}\boldsymbol{X}'_j$ be the projection onto the row space of $\boldsymbol{X}'_j$. Then, applying the rank-one update rule of the inverse of matrix product (see e.g. [13]) on $(\boldsymbol{X}'_j\boldsymbol{X}'^T_j)^{-1}$ we get

$$\|\boldsymbol{L}'_j\boldsymbol{\beta}\|^2 = \|\boldsymbol{L}'_{j-1}\boldsymbol{\beta}\|^2 + \text{update}(j), \tag{13}$$

with $\text{update}(j) = \mathbb{E}\left[\frac{((\boldsymbol{\beta}^T+\boldsymbol{Z}_j^T)\boldsymbol{D}_j^T(\boldsymbol{I}-\boldsymbol{L}'_{j-1})\boldsymbol{\beta})^2}{(\boldsymbol{\beta}^T+\boldsymbol{Z}_j^T)\boldsymbol{D}_j(\boldsymbol{I}-\boldsymbol{L}'_{j-1})\boldsymbol{D}_j(\boldsymbol{\beta}+\boldsymbol{Z}_j)}\right]$.

It can be seen that at each step the numerator of update(j) is the projection of the part of $\boldsymbol{\beta}$ that lies in the null space of $\boldsymbol{L}'_{j-1}$, namely the part of $\boldsymbol{\beta}$ that was not captured by the first $j-1$ rows of $\boldsymbol{X}'$, onto the row space of the new batch. However, the projection is affected by the current batch's noise, $\boldsymbol{Z}_j$. We can view this as a noisy version of the projection $\boldsymbol{D}_j$, a perspective that will be made clear in the next lemma, which is the key tool for analyzing the recursive rule (13).

Lemma 1. *Let $\boldsymbol{P}$ be a projection onto a subspace of dimension δp for $\delta \in [0,1]$. Write $\|\boldsymbol{\beta}\| = r$ and $\|\boldsymbol{P}\boldsymbol{\beta}\|^2 = \alpha r^2$ for $\alpha \in [0,1]$. Let $\{\boldsymbol{U}_i\}_{i=1}^b$ be uniformly drawn orthonormal vectors, and $\tilde{\boldsymbol{D}}$ a noisy projection onto the span of $\{\boldsymbol{U}_i\}$ given by $\tilde{\boldsymbol{D}}\boldsymbol{v} = \boldsymbol{D}\boldsymbol{v} + \sum_{i=1}^b \boldsymbol{U}_i Z_i$, where $\boldsymbol{D} = \sum_{i=1}^b \boldsymbol{U}_i\boldsymbol{U}_i^T$, $Z_i \sim \mathcal{N}(0, T_i \cdot \sigma^2/p)$, and T_i are r.vs. mutually independent of $\{\boldsymbol{U}_i\}$ and concentrated in the interval $[1-o(1), 1+o(1)]$ with probability at least $1-o(1/p)$. Then, the expected squared noisy projection of $\boldsymbol{\beta}$ in the direction of $\boldsymbol{P}\tilde{\boldsymbol{D}}\boldsymbol{\beta}$ is*

$$\mathbb{E}\Big[\frac{\langle \boldsymbol{P}\tilde{\boldsymbol{D}}\boldsymbol{\beta}, \boldsymbol{\beta}\rangle^2}{\|\boldsymbol{P}\tilde{\boldsymbol{D}}\boldsymbol{\beta}\|^2}\Big] = \frac{1}{p}\left(\frac{\alpha r^2}{\delta}\Big(1 + \frac{\alpha r^2(b-1)}{\sigma^2 + r^2}\Big) + o(1)\right). \tag{14}$$

The remainder of this section is dedicated to the proof of Lemma 1, via a Gaussian approximation technique. But first, we use this lemma to prove the upper bound on the bias in Theorem 1.

Proof of bias part in Theorem 1. We give the high level sketch of the proof. Full proof can be found in [8]. Write $B_j = 1 - \frac{1}{r^2}\|\boldsymbol{L}'_j\boldsymbol{\beta}\|^2$ for the bias after j steps and assume without loss of generality that $r = 1$. Then $B_0 = 1$, and the desired bias is given by $\lim \mathsf{Bias}(\hat{\boldsymbol{\beta}}_{\text{BMN}}) = \lim \mathbb{E}B_{\frac{n}{b}}$. Let $\{\boldsymbol{U}_i\}_1^b$, be the orthonormal basis for the row space of $\boldsymbol{X}_j$, then $\boldsymbol{D}_j = \sum_{i=1}^b \boldsymbol{U}_i\boldsymbol{U}_i^T$, and $\boldsymbol{D}_j(\boldsymbol{\beta} + \boldsymbol{Z}_j)$ can be written as $\tilde{\boldsymbol{D}}_j\boldsymbol{\beta}$. Then we get from Lemma 1 with $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{L}'_{j-1}$ that the update term in (13) is given by

$$\text{update}(j) = \mathbb{E}\left[\tfrac{1}{p-j}B_j(1 + (b-1)\xi B_j)\right] + o(1/p), \tag{15}$$

where we set $\alpha = B_j$ and used the fact that with probability 1 the dimension of $\bar{\boldsymbol{L}}_{j-1}$ is $\delta = 1 - j/p$. Then, using (13) and taking expectation we get

$$\mathbb{E}B_{j+1} = \mathbb{E}B_j - \tfrac{1}{p-j}(\mathbb{E}B_j + (b-1)\xi\mathbb{E}B_j^2) + o(1/p). \quad (16)$$

Next, we will show that (16) can be translated to a differential equation, which we then solve to produce the desired upper bound. Define $t_j \triangleq \mathbb{E}B_j$. Using the (trivial) bound $(\mathbb{E}B_j)^2 \leq \mathbb{E}B_j^2 \leq \mathbb{E}B_j$ that holds for any random variable with support in the unit interval, we obtain

$$\frac{1+(b-1)\xi t_j}{1-j/p} \leq \frac{t_{j+1}-t_j}{1/p} + o(\tfrac{1}{p}) \leq -\frac{1+(b-1)\xi t_j^2}{1-j/p}. \quad (17)$$

We are interested in $t_{n/b}$ under the initial condition $t_0 = 1$, as $p \to \infty$ and $p/n = \gamma$ is held fixed. The upper bound in (17) is monotonically increasing in t_j, and therefore we can use it iteratively. Noticing that over n/b iterations of the recursive bound (17) the error term $o(1/p)$ can grow to at most $o(1)$, we drop it hereafter and add it back later. Then, loosely speaking, by taking $g(x)$ to be some convex function such that

$$g'(x) \leq \frac{g(x+1/p)-g(x)}{1/p} \leq -\frac{g(x)\cdot(1+(b-1)\xi g(x))}{1-x},$$

we will get $t_j \leq g(j/p)$. The above is a differential inequality with the initial condition $g(0) = 1$. Solving it we get $g(x) \leq \frac{1-x}{1+(b-1)\xi\cdot x}$, hence, we have $t_j \leq g(j/p) \leq \frac{1-j/p}{1+(b-1)\xi\cdot j/p}$. Then, adding back the error term we get

$$\lim_{p\to\infty} \tfrac{1}{r^2}\,\mathsf{Bias}\left(\hat{\boldsymbol{\beta}}_{\mathrm{BMN}}\right) \leq \lim_{p\to\infty} g(\tfrac{1}{\gamma b}) + o(1) \leq \frac{\gamma b - 1}{\gamma b + (b-1)\xi},$$

□

Next, we turn to prove Lemma 1. Let $\boldsymbol{S} = [S_1, \cdots, S_{2b}]$ be a random vector given by

$$S_i \triangleq \begin{cases} \langle \boldsymbol{P}\boldsymbol{U}_i, \boldsymbol{\beta}\rangle, & i = 1, \cdots b, \\ \langle (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{U}_i, \boldsymbol{\beta}\rangle, & i = b+1, \cdots, 2b \end{cases}, \quad (18)$$

and define the function

$$f(\boldsymbol{s}, \boldsymbol{z}) \triangleq \frac{\left(\sum_{i=1}^b s_i\left(s_i + s_{b+i} + z_i\right)\right)^2}{\sum_{i=1}^b \left(s_i + s_{b+i} + z_i\right)^2}. \quad (19)$$

The outline of the proof for Lemma 1 is as follows. First, in Lemma 2 we show that the expected squared projection (14) is approximately equal to $\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z})$, for any noise $\boldsymbol{Z}$ that is sufficiently close in distribution to i.i.d. $\mathcal{N}(0, 1/p)$ noise. Then, we show that $\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z})$ can be calculated with good accuracy by replacing the vector $\boldsymbol{S}$ with a Gaussian vector $\boldsymbol{G}$ with independent entries that have the same variance as the elements of $\boldsymbol{S}$. To do so, in Lemma 3 we bound the Wasserstein distance between $\boldsymbol{S}$ and $\boldsymbol{G}$ using Corollary 1, and show that f is Lipschitz where $\boldsymbol{S}$ and $\boldsymbol{G}$ are concentrated, hence $|\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z}) - \mathbb{E}f(\boldsymbol{G}, \boldsymbol{Z})| \lessapprox \mathcal{W}_1(\boldsymbol{S}, \boldsymbol{G})$. Then in Lemma 4 we explicitly calculate $\mathbb{E}f(\boldsymbol{G}, \boldsymbol{Z})$ as a (random) weighted sum of MMSE estimators, yielding (14) and concluding the proof.

First, we show that indeed $\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z})$ approximates the mean-squared projection (14). The proofs of all the lemmas can be found in [8].

Lemma 2. *Let $\boldsymbol{S}$ be given by* (18), *$\boldsymbol{Z} = [Z_1, \cdots, Z_b]$ as in Lemma 1, and $f(\boldsymbol{S}, \boldsymbol{Z})$ as in* (19). *Then,*

$$\left|\mathbb{E}[\tfrac{1}{\delta}f(\boldsymbol{S}, \boldsymbol{Z})] - \mathbb{E}\left[\tfrac{\langle \boldsymbol{P}\tilde{\boldsymbol{D}}\boldsymbol{\beta}, \boldsymbol{\beta}\rangle^2}{\|\boldsymbol{P}\tilde{\boldsymbol{D}}\boldsymbol{\beta}\|^2}\right]\right| \leq \tfrac{1}{\sqrt[4]{p}}\mathbb{E}[\tfrac{1}{\delta}f(\boldsymbol{S}, \boldsymbol{Z})] + o(\tfrac{1}{p}). \quad (20)$$

Next, we show that for large p the vector $\boldsymbol{S}$ is close to Gaussian in the Wasserstein distance, a fact we can utilize to approximate $\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z})$.

Corollary 1. *Let $\boldsymbol{S} \in \mathbb{R}^{2b}$ be the random vector given in* (18), *and let $\boldsymbol{G} \in \mathbb{R}^{2b}$ have independent entries such that*

$$G_i \sim \begin{cases} \mathcal{N}(0, \frac{\alpha}{p}), & i = 1, \cdots b, \\ \mathcal{N}(0, \frac{1-\alpha}{p}), & i = b+1, \cdots, 2b \end{cases}. \quad (21)$$

Then $\mathcal{W}_1(\boldsymbol{S}, \boldsymbol{G}) \leq \sqrt{\frac{b}{p}} \cdot \frac{2\sqrt{2}b}{p-1}$.

We see that $\boldsymbol{S}$ in (18) is $O(p^{-\frac{3}{2}})$-close in Wasserstein distance to $\boldsymbol{G}$ in (21). If f was k-Lipschitz in $\boldsymbol{s}$, this would yield a $O(kp^{-\frac{3}{2}})$ approximation for $\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z})$, by calculating the latter using the Gaussian statistics. This is however not the case, since f's gradient diverges along certain curves. Nonetheless, we will show that f is Lipschitz in the region where $\boldsymbol{S}$ and $\boldsymbol{G}$ are concentrated, which along with the fact that $\boldsymbol{Z}$ is independent of $\boldsymbol{S}$ and $\boldsymbol{G}$, will yield an upper bound on $|\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z}) - \mathbb{E}f(\boldsymbol{G}, \boldsymbol{Z})|$.

Lemma 3. *Let $f : \mathbb{R}^{2b} \times \mathbb{R}^b \to \mathbb{R}$ be defined in* (19), *$\boldsymbol{S}$ and $\boldsymbol{Z}$ defined in* (18) *and Lemma 1, respectively, and $\boldsymbol{G}$ be distributed as in* (21) *and independent of $\boldsymbol{Z}$. Then,*

$$|\mathbb{E}f(\boldsymbol{S}, \boldsymbol{Z}) - \mathbb{E}f(\boldsymbol{G}, \boldsymbol{Z})| = o\left(1/p\right). \quad (22)$$

Lemma 4. *Let $\boldsymbol{G}$ be as in Corollary 1 and $\boldsymbol{Z}$ as in Lemma 1, then*

$$\mathbb{E}\left[f(\boldsymbol{G}, \boldsymbol{Z})\right] = \tfrac{\alpha r^2}{p}\left(1 + (b-1)\xi\alpha\right) + o(1/p). \quad (23)$$

The proof of Lemma 1 is now a direct result of Lemmas 2, 3 and 4.

References

[1] Y. Zhang, J. Duchi, and M. Wainwright, "Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3299–3340, 2015.

[2] J. D. Lee, Y. Sun, Q. Liu, and J. E. Taylor, "Communication-efficient sparse regression: a one-shot approach," *arXiv preprint arXiv:1503.04337*, 2015.

[3] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[4] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[5] V. A. Marchenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Matematicheskii Sbornik*, vol. 114, no. 4, pp. 507–536, 1967.

[6] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, vol. 50, no. 2, pp. 949–986, 2022.

[7] P. Nakkiran, P. Venkat, S. M. Kakade, and T. Ma, "Optimal regularization can mitigate double descent," in *International Conference on Learning Representations*, 2021.

[8] S. S. Ioushua, I. Hasidim, O. Shayevitz, and M. Feder, "Batches stabilize the minimum norm risk in high dimensional overparameterized linear regression," *arXiv preprint arXiv:2306.08432*, 2023.

[9] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[10] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[11] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[12] C. Paquette, K. Lee, F. Pedregosa, and E. Paquette, "Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality," in *Conference on Learning Theory*, pp. 3548–3626, PMLR, 2021.

[13] W. W. Hager, "Updating the inverse of a matrix," *SIAM review*, vol. 31, no. 2, pp. 221–239, 1989.

International Zurich Seminar on Information and Communication (IZS), March 6 – 8, 2024

Testing Dependency of Unlabeled Databases

Vered Paslev and Wasim Huleihel
Tel Aviv University
Department of Electrical Engineering - Systems
Tel Aviv, Israel
email: {veredpaslev@mail,wasimh@tauex}.tau.ac.il

***Abstract*—In this paper, we study the problem of deciding whether two random databases $\mathsf{X} \in \mathbb{R}^{n\times d}$ and $\mathsf{Y} \in \mathbb{R}^{n\times d}$, each composed of n users with d features, are statistically dependent or not. This problem is relevant in computational biology, natural language processing, analysis of social media, etc. We formulate this decision task as the following hypothesis testing problem: under the null hypothesis, these two databases are statistically independent, while under the alternative, there exists an unknown row permutation σ, such that X and Y^{σ}, a permuted version of Y, are statistically dependent with some known joint distribution. For this problem, we characterize the thresholds at which optimal testing is information-theoretically impossible and possible, as a function of n, d, and the generative distributions of the datasets. Specifically, we prove that if a certain function of the eigenvalues of the likelihood function and d, is below a certain threshold, as $d \to \infty$, then weak detection (performing slightly better than random guessing) is statistically impossible, *irrespectively* of the value of n. This matches the performance of an efficient test that thresholds a centered version of the log-likelihood function of the observed matrices.**

I. Introduction

Prompted by practical scenarios, such as computational biology applications [1], [2], social network analysis [3], [4], computer vision [5], [6], and data anonymization/privacy-focused systems, there has been a recent focus on exploring the theoretical underpinnings and algorithmic solutions for database alignment under statistical frameworks. Indeed, quantifying relationships between disparate databases stand as fundamental undertakings in statistics. Modern databases present many challenges: they are high-dimensional, lack labels, contain noise, and appear scrambled. One concrete example of an inference problem, involving a pair of databases, is framed as the following hypothesis testing problem. Under the null hypothesis, no statistical dependency exists between the databases, whereas under the alternative hypothesis, there exists a permutation that reorganizes one database in such a way that the two become dependent. Then, given these databases, under what conditions can we discern whether they are dependent or not?

As a tangible folklore example, consider the following scenario: envision two distinct data sources, such as Netflix and IMDb, each providing feature lists for a set of entities, like users. These features encompass diverse user attributes, such as names, user IDs, and ratings. Frequently, feature labels are either unavailable or intentionally removed to safeguard sensitive personally identifiable information. Consequently, straightforwardly matching feature pairs between the two sources, corresponding to the same user, becomes challenging. Nevertheless, there is optimism that when a substantial correlation exists between the two databases, it becomes feasible to establish connections between them and create a coherent alignment for their respective feature lists [3], [4].

Recently, there has been a focus on what is known as the *data alignment problem.* This problem can be seen as a straightforward probabilistic model that encapsulates the scenario described above. It was introduced and explored in, e.g., [7], [8]. In essence, this problem involves two databases, denoted as $\mathsf{X} \in \mathbb{R}^{n\times d}$ and $\mathsf{Y}^{n\times d}$, each comprising n users, each with d features. The key challenge lies in uncovering an unknown permutation or correspondence that matches users in X with those in Y. When a pair of entries from these databases is matched, their features are dependent according to a known distribution, whereas for unmatched entries, the features are independent. The primary objective is to *recover* the unknown permutation and establish statistical assurances regarding the feasibility and impossibility of this recovery. The feasibility of recovery is contingent upon factors such as the correlation level, n, and d. The statistical limits of this recovery problem are understood for some specific probability distributions that generate the databases. For instance, in the Gaussian case, where the two databases have independent standard normal entries, with correlation coefficient ρ between the entries of matched rows, it has been demonstrated in [8] that perfect recovery is attainable if $\rho^2 = 1 - o(n^{-4/d})$, while it becomes impossible if $\rho^2 = 1 - \omega(n^{-4/d})$ as both $n, d \to \infty$.

The *detection* counterpart of the recovery problem discussed above has also undergone extensive investigation in [9]–[12], in the Gaussian case. As mentioned earlier, the central question here revolves around determining the correlation level needed to decide whether the two databases are correlated or not. It has been established in [10] that when $\rho^2 d \to \infty$, efficient detection becomes feasible (with an exceedingly small error probability) using a straightforward thresholding of the sum of entries in $\mathsf{X}^T\mathsf{Y}$. Conversely, it has also been demonstrated that when $\rho^2 d\sqrt{n} \to 0$ and $d = \Omega(\log n)$, detection is information-theoretically impossible. Most recently, in [12], the gap between the lower and upper bound above has been conclusively addressed, proving that the upper bound is tight, i.e., if $\rho^2 d \to 0$ as $d \to \infty$, weak detection (performing slightly better than random guessing) is information-theoretically impossible, regardless of the specific value of n. In fact, while the main focus in related literature was exclusively on the asymptotic regime where both n and d tend to infinity, [12] determine sharp thresholds for all possible asymptotic regimes of n and d.

In this paper, we continue the investigation of the database alignment detection problem. While the results above are neat and interesting, they apply for the Gaussian case only. This is in fact true not just for the detection but also for the recovery problem. This case is, however, restricted in the sense that for jointly Gaussian random variables, the notion of uncorrelatedness and independence are equivalent; the information-theoretic limits depend on the distribution through the correlation parameter solely. Accordingly, for general distributions, it is a priori unclear how these thresholds depend on the distribution. Furthermore, as it turns out, many of the techniques in [9]–[12] are tailored to the Gaussian case, and the tight analysis of general distributions becomes challenging. To the best of our knowledge, our paper is the first to provide characterization for the thresholds at which optimal testing is information-theoretically impossible and possible, for two general distributions. Specifically, given two generative distributions, we prove that weak (strong) detection is information-theoretically impossible if a certain function of the eigenvalues of the likelihood function kernel (see, (6)) is below some threshold; we consider all possible asymptotic regimes as a function of n and d. We then complement out lower bounds by algorithmic upper bounds, and show that these are tight in a few classical examples. To that end, we propose three detection algorithms and analyze their associated risks and sample complexities.

We will now provide a brief overview of other related works. In [13], the problem of partial recovery of the hidden alignment was investigated. In [14], necessary and sufficient conditions for successful recovery through a typicality-based framework were established. Additionally, [15] and [16] explored the problem of alignment recovery in cases involving feature deletions and repetitions, respectively. A recent development in this area involved the joint analysis of correlation detection and alignment in Gaussian databases, as presented in [11]. Finally, it is worth noting that the challenges in database alignment and detection closely relate to various planted matching problems, specifically the graph alignment problem, which has yielded numerous intriguing results and valuable mathematical techniques. Roughly speaking, in this problem the goal is to detect the edge correlation between two random graphs with unlabeled nodes. For further exploration, one can refer to works such as [17]–[24], and their associated references. In particular, as we mention later on in the outlook of our paper, an interesting open problem mentioned in [21] is to investigate the case of general edge weight distributions, which inspired our work for the random databases case.

For any $n \in \mathbb{N}$, the set of integers $\{1, 2, \ldots, n\}$ is denoted by $[n]$, and $a_1^n = \{a_1, a_2, \ldots, a_n\}$. Let $\mathbb{S}_n$ denotes the set of all permutations on $[n]$. For a given permutation $\sigma \in \mathbb{S}_n$, let σ_i denote the value to which σ maps $i \in [n]$. Random vectors are denoted by capital letters such as X with transpose X^T. A collection of n random vectors is written as $\mathsf{X} = (X_1, \ldots, X_n)$. The notation $(X_1, \ldots, X_n) \sim P_X^{\otimes n}$ means that the random vectors $(X_1, \ldots, X_n)$ are independent and identically distributed (i.i.d.) according to P_X. We use $\mathcal{N}(\eta, \Sigma)$ to represent the multivariate normal distribution with mean vector η and covariance matrix Σ. Let $\mathsf{Poisson}(\lambda)$ denote the Poisson distribution with parameter λ. The $n \times n$ identity matrix is denoted by $I_{n\times n}$. For probability measures $\mathbb{P}$ and $\mathbb{Q}$, let $d_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\int |\mathrm{d}\mathbb{P} - \mathrm{d}\mathbb{Q}|$ denote the total variation distance. For a probability measure μ on a space Ω, we use $\mu^{\otimes d}$ for the product measure of μ (d times) on the product space Ω^d. For a measure $\nu \ll \mu$ (that is, a measure absolutely continuous with respect to μ), we denote (by abuse of notation) the Randon-Nikodym derivative ν with respect to μ by $\frac{\nu}{\mu}$. For functions $f, g : \mathbb{N} \to \mathbb{R}$, we say that $f = O(g)$ (and $f = \Omega(g)$) if there exists $c > 0$ such that $f(n) \leq cg(n)$ (and $f(n) \geq cg(n)$) for all n. We say that $f = o(g)$ if $\lim_{n\to\infty} f(n)/g(n) = 0$, and that $f = \omega(g)$ if $g = o(f)$. **Due to space limitation, our proofs appear in [25].**

II. Setup and Learning Problem

Probabilistic model. As mentioned in the introduction, in this paper we will investigate the following decision problem. In general, we deal with two databases $\mathsf{X} = (X_1, \ldots, X_n)$ and $\mathsf{Y} = (Y_1, \ldots, Y_n)$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}^d$, for $i = 1, 2, \ldots, n$, with n being the number of entities (say, users), and d is the number of features. Now, under the null hypothesis $\mathcal{H}_0$, the databases X and Y are generated *independently* at random, where $X_1, \ldots, X_n \sim P_X^{\otimes d}$ and $Y_1, \ldots, Y_n \sim P_Y^{\otimes d}$, with $P_X = P_Y$. We denote by $\mathbb{P}_0$ the null distribution, i.e., the joint distribution of (X, Y) under $\mathcal{H}_0$. Also, for simplicity of notation, we denote the product measure $Q_{XY} \triangleq P_X \times P_Y$. Under the alternate hypothesis $\mathcal{H}_1$, the databases X and Y are dependent under an unknown alignment/permutation $\sigma \in \mathbb{S}_n$, namely, given $\sigma \in \mathbb{S}_n$, a permutation over $1, 2, \ldots, n$, we have $(X_1, Y_{\sigma_1}), (X_2, Y_{\sigma_2}), \ldots, (X_n, Y_{\sigma_n}) \overset{\text{i.i.d}}{\sim} P_{XY}^{\otimes d}$, with the same marginals $P_X = P_Y$. For a fixed $\sigma \in \mathbb{S}_n$, we denote the joint distribution measure of (X, Y) under the hypothesis $\mathcal{H}_1$ by $\mathbb{P}_{\mathcal{H}_1|\sigma}$. To summarize,

$$\begin{aligned} &\mathcal{H}_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d}}{\sim} P_X^{\otimes d} \times P_Y^{\otimes d} \\ &\mathcal{H}_1 : (X_1, Y_{\sigma_1}), \ldots, (X_n, Y_{\sigma_n}) \overset{\text{i.i.d}}{\sim} P_{XY}^{\otimes d}, \quad \text{for } \sigma \in \mathbb{S}_n. \end{aligned} \tag{1}$$

Thinking about X as an $n \times d$ matrix, we denote its (i, j) element by X_{ij} (and similarly for Y). We remark here that the distributions Q_{XY} and P_{XY} are allowed to be functions of n and d. In fact, as we will see later on, this is the interesting non-trivial case.

Learning problem. Given the databases X and Y, a test/detection algorithm $\phi_{n,d} : \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \to \{0, 1\}$, for the hypothesis testing problem above, is tasked with outputting a decision in $\{0, 1\}$. We define the risk of a detection algorithm $\phi_{n,d}$ as the sum of its Type-I and (worst-case) Type-II error probabilities, namely,

$$\mathsf{R}(\phi_{n,d}) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi_{n,d}(\mathsf{X}, \mathsf{Y}) = 1] + \max_{\sigma\in\mathbb{S}_n} \mathbb{P}_{\mathcal{H}_1|\sigma}[\phi_{n,d}(\mathsf{X}, \mathsf{Y}) = 0]. \tag{2}$$

The *minimax* risk for our hypothesis detection problem is

$$\mathsf{R}^\star \triangleq \inf_{\phi_{n,d}:\mathbb{R}^{n\times d}\times\mathbb{R}^{n\times d}\to\{0,1\}} \mathsf{R}(\phi_{n,d}). \tag{3}$$

We remark that R (and $\mathsf{R}^\star$) is in general a function of n, d, Q_{XY}, and P_{XY}. However, we omit them from our notation for the benefit of readability, as we shall do for our detection algorithms as well.

As in [12], we study the information-theoretic limits (i.e., impossibility and possibility lower and upper bounds) of the hypothesis testing problem in (1), for several possible asymptotic regimes of n and d. To be more precise, let $\mathcal{D}$ denote the pair of distributions Q_{XY}, and P_{XY}. Then, the regimes we investigate are characterized by sequences of the parameters $(\mathcal{D},d,n) = (\mathcal{D}_\ell,d_\ell,n_\ell)_{\ell\in\mathbb{N}}$. For example, if $P_X = P_Y = \mathcal{N}(0,1)$ and P_{XY} denote the joint distribution of two standard normal random variables with correlation coefficient ρ, then the triplet $(\mathcal{D},d,n)$ is equivalent to (ρ,d,n). In this paper, the asymptotic regimes we consider correspond to the scenarios where d_ℓ and n_ℓ are either fixed or tend to infinity. Accordingly, asymptotic notations, such as, $f(\mathcal{D}) = o(\cdot)$, $f(\mathcal{D}) = \Omega(\cdot)$, etc., where $f(\cdot)$ is some one-dimensional function of $\mathcal{D}$, should be interpreted in terms of the sequences above. For example, the condition $f(\mathcal{D}) = o(d^{-1})$ means that the sequence $(\mathcal{D},d,n)$ satisfies $f(\mathcal{D}_\ell)d_\ell \to 0$, as $\ell\to\infty$. Later on, we will give concrete examples which elucidate the notations above. With the above in mind, we are in a position to define the notions of strong and weak detection.

Definition 1. *A sequence $(\mathcal{D},d,n) = (\mathcal{D}_k,d_k,n_k)_k$ is said to be:*

1) *Admissible for strong detection if* $\lim_{k\to\infty}\mathsf{R}^\star = 0$.
2) *Admissible for weak detection if* $\limsup_{k\to\infty}\mathsf{R}^\star < 1$.

A few comments are in order. First, it should be clear that admissibility of strong detection implies the admissibility of weak detection. More concretely, if a test ϕ achieves strong detection, i.e., $\mathsf{R}(\phi)\to 0$, then ϕ achieves weak detection as well. Also, note that strong detection requires the test statistic to determine with high probability whether (X,Y) is is drawn from $\mathcal{H}_0$ or $\mathcal{H}_1$, while weak detection only aims at strictly outperforming random guessing of the underlying hypothesis. Finally, recall that the optimal average risk $\bar{\mathsf{R}}^\star$, which corresponds to the case where σ is uniformly drawn over $\mathbb{S}_n$ (rather being unknown), and which lower bounds the worst-case risk $\mathsf{R}^\star$, is given by $1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1})$. This is achieved by the likelihood ratio (or, Neyman-Pearson) test. Accordingly, weak and strong detection are equivalent to $\liminf d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1}) > 0$ and $\lim d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1}) = 0$, respectively. In the same vein, to rule out the possibility of weak/strong detection, we will use the well-known facts that,

$$d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1}) = o(1) \implies \lim_{k\to\infty}\mathsf{R}^\star = 1, \tag{4}$$

and

$$d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1}) \le 1-\Omega(1) \implies \liminf_{k\to\infty}\mathsf{R}^\star > 0. \tag{5}$$

To wit, the implication in (4) correspond to the impossibility of weak detection, while (5) corresponds to the impossibility of strong detection, respectively.

III. Main Results

In this section, we present our main results concerning the thresholds for admissibility and impossibility of weak and strong detection. We differentiate between the various possible asymptotic regimes: (I) both $n,d\to\infty$; the standard regime analyzed in most related past literature, (II) n is a constant and $d\to\infty$, and (III) d is a constant and $n\to\infty$. We begin with our impossibility lower-bounds. To that end, we introduce a few important notations.

Let $\mathcal{L}(x,y) = \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$, for any $x\in\mathcal{X}$ and $y\in\mathcal{Y}$. This kernel defines an operator as follows: for any square-integrable function f under Q_{XY},

$$(\mathcal{L}f)(x) \triangleq \mathbb{E}_{Y\sim Q_Y}\left[\mathcal{L}(x,Y)f(Y)\right], \tag{6}$$

In addition, $\mathcal{L}^2 = \mathcal{L}\circ\mathcal{L}$ is given by $\mathcal{L}^2(x,y) = \mathbb{E}_{Z\sim Q}[\mathcal{L}(x,Z)\mathcal{L}(Z,y)]$, and $\mathcal{L}^k$ is similarly defined. Assume that $\mathcal{L}(x,y) = \mathcal{L}(y,x)$, and hence $\mathcal{L}$ is self-adjoint. Furthermore, if we assume that $\int\int\mathcal{L}^2(x,y)Q_X(\mathrm{d}x)Q_Y(\mathrm{d}y) < \infty$, then $\mathcal{L}$ is Hilbert-Schmidt. Thus $\mathcal{L}$ is diagonazable with eigenvalues λ_i's and the trace of $\mathcal{L}$ is given by $\mathsf{trace}(\mathcal{L}) = \mathbb{E}_{Y\sim Q_Y}[\mathcal{L}(Y,Y)] = \sum_{i\in\mathbb{N}}\lambda_i$. Without loss of generality, we assume that the sequence of eigenvalues $\{\lambda_i\}_{i\ge 0}$ decreasing, namely, $\lambda_i \ge \lambda_{i+1}$, for all $i\in\mathbb{N}$. As we show in the proofs, the largest eigenvalue of $\mathcal{L}$ is one, i.e., $\lambda_0 = 1$. We are now in a position to state our main results.

Theorem 1 (Weak detection lower bound)**.** *Weak detection is impossible as long as*

$$\sum_{i\ge 1}\frac{\lambda_i^2}{1-\lambda_i^2} = o(d^{-1}). \tag{7}$$

That is, for a sequence $(\mathcal{D},d,n) = (\mathcal{D}_\ell,d_\ell,n_\ell)_{\ell\in\mathbb{N}}$ such that (7) holds:

- *If d is any function of k, and $n\to\infty$ then* $\lim_{k\to\infty}\mathsf{R}^\star = 1$.
- *If n is constant and $d\to\infty$ then* $\lim_{k\to\infty}\mathsf{R}^\star = 1$.

Namely, (ρ,d,n) is not admissible for weak detection.

Based on Theorem 1, we see that weak detection is impossible when (7) holds, for all different asymptotic regimes of n and d; however, if d is fixed then the right-hand-side of (7) should be understood as $o(1)$, as $n\to\infty$. Next, we move forward to our strong detection lower bounds.

Theorem 2 (Strong detection lower bounds)**.** *A sequence $(\mathcal{D},d,n) = (\mathcal{D}_\ell,d_\ell,n_\ell)_{\ell\in\mathbb{N}}$ is not admissible for strong detection if:*

1) *$d\in\mathbb{N}$ and $\{\lambda_i\}_{i\ge 1}$ are constants, such that*

$$d < -\frac{\log\lambda_1^2}{\log\sum_{i\in\mathbb{N}}\lambda_i^2}, \tag{8}$$

and $n\to\infty$.

2) *$n, d \to \infty$, and*

$$\sum_{i\geq 1}\frac{\lambda_i^2}{1-\lambda_i^2} < (1-\varepsilon)d^{-1}, \tag{9}$$

for some $\varepsilon > 0$, independent of n and d.

3) *$d \to \infty$, and $n \in \mathbb{N}$,*

$$\sum_{i\geq 1}\frac{\lambda_i^2}{1-\lambda_i^2} = O(d^{-1}). \tag{10}$$

It is evident that for strong detection, we can prove slightly stronger results; for example, when $n, d \to \infty$, comparing (7) and (9) we see that the barriers are the same up to a constant factor. We next consider two canonical special cases for which we find explicit formulas for the lower bounds in Theorems 1 and 2.

Example 1 (Gaussian databases)**.** *In the Gaussian case, we assume that $Q_{XY} = P_X \times P_Y$, with P_X and P_Y correspond to the densities of a Gaussian random variable with zero mean and unit variance, while P_{XY} is the joint density of two correlated zero mean Gaussian random variables with unit variance, i.e.,*

$$P_{XY} \equiv \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right), \tag{11}$$

for some known correlation coefficient $\rho \in [-1,1]\setminus\{0\}$. The Gaussian case above was analyzed in [9]–[12]. It can be shown that the eigenvalues of $\mathcal{L}$ in this case are given by $\lambda_\ell = \rho^\ell$, for $\ell \geq 0$. Accordingly, a little bit of algebra reveals that (7) holds if $\rho^2 = o(d^{-1})$. This condition, coincides with [12]. Similarly, in the regime where d is fixed, strong detection is impossible when (8) holds, which in the Gaussian setting boils down to $d < \frac{\log\rho^2}{\log(1-\rho^2)}$, which again coincides with the results of [12]. Therefore, our general lower bounds in Theorems 1 and 2 recover the known bounds in the literature.

Example 2 (Bernoulli databases)**.** *In the Bernoulli case, we assume that $Q_{XY} = P_X \times P_Y$, with $P_X = P_Y = \mathsf{Bernoulli}(\tau p)$, for some $p \in (0,1)$ and $\tau \in [0,1]$, and P_{XY} denotes the joint distribution of two correlated Bernoulli random variables. Specifically, under P_{XY}, we have $X \sim \mathsf{Bernoulli}(\tau p)$, and*

$$Y|X \sim \begin{cases}\mathsf{Bernoulli}(\tau), & \text{if } X = 1\\ \mathsf{Bernoulli}\left(\frac{\tau p(1-\tau)}{1-\tau p}\right), & \text{if } X = 0.\end{cases} \tag{12}$$

Here, Pearson correlation coefficient is given by,

$$\rho \triangleq \frac{\mathsf{cov}(X,Y)}{\sqrt{\mathsf{var}(X)}\sqrt{\mathsf{var}(Y)}} = \frac{\tau(1-p)}{1-\tau p}. \tag{13}$$

The Bernoulli case was not studied in the literature in the context of the database alignment detection problem, but it is the focus in related work on testing correlation of unlabeled random graphs, e.g., [21], where the edges are modelled as Bernoulli random variables. Here, it can be shown that the eigenvalues of $\mathcal{L}$ are 1 and ρ. Thus, the lower bound in Theorem 1 for weak detection, boils down to $\rho^2 = o(d^{-1})$, as in the Gaussian case. Similarly, in the regime where d is fixed, strong detection is impossible when (8) holds, which in the Bernoulli setting, translates to $d < \frac{\log(1/\rho^2)}{\log(1+\rho^2)}$, similarly to the Gaussian case.

Next, we present our detection algorithms and the corresponding upper bounds. We start with the following sum test,

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\frac{1}{dn^2}\sum_{i,j=1}^{n}\sum_{\ell=1}^{d}\mathcal{K}(X_{i\ell},Y_{j\ell}) \geq \tau_{\mathsf{sum}}\right\}, \tag{14}$$

where

$$\begin{aligned}\mathcal{K}(X_{i\ell},Y_{j\ell}) \triangleq{}& \log\frac{P_{XY}(X_{i\ell},Y_{j\ell})}{Q_{XY}(X_{i\ell},Y_{j\ell})} - d_{\mathsf{KL}}(Q_{XY}||P_{XY})\\ &- \mathbb{E}_{A\sim P_X}\left[\log\frac{P_{XY}(A,Y_{i\ell})}{Q_{XY}(A,Y_{i\ell})}\right]\\ &- \mathbb{E}_{B\sim P_X}\left[\log\frac{P_{XY}(X_{i\ell},B)}{Q_{XY}(X_{i\ell},B)}\right],\end{aligned} \tag{15}$$

is the centered likelihood function, and $\tau_{\mathsf{sum}} \in \mathbb{R}$. In the Gaussian case (as well as in the Bernoulli case) it can be shown that (14) is equivalent to thresholding the sum of all entries of the inner product $\mathsf{X}^T\mathsf{Y}$, i.e., $\sum_{i,j=1}^n X_{ij}Y_{ij}$. This test was analyzed in [10], in the Gaussian case. For the general case, we have the following result. Below $\mathsf{Var}_Q(X)$ denotes the variance of X distributed according to Q.

Theorem 3 (Sum test)**.** *Consider the sum test in (14), and let*

$$\tau_{\mathsf{sum}} = dn\frac{d_{\mathsf{KL}}(P_{XY}||Q_{XY}) + d_{\mathsf{KL}}(Q_{XY}||P_{XY})}{2}. \tag{16}$$

Then,

$$\mathsf{R}(\phi_{\mathsf{sum}}) \leq \frac{16\cdot\mathsf{Var}_{Q_{XY}}(\mathcal{K}(A,B))}{d\cdot\left(d_{\mathsf{KL}}(P_{XY}||Q_{XY}) + d_{\mathsf{KL}}(Q_{XY}||P_{XY})\right)^2}. \tag{17}$$

In particular, if

$$d\cdot\frac{\left(d_{\mathsf{KL}}(P_{XY}||Q_{XY}) + d_{\mathsf{KL}}(Q_{XY}||P_{XY})\right)^2}{\mathsf{Var}_{Q_{XY}}(\mathcal{K}(A,B))} = \omega(1), \tag{18}$$

then, $\mathsf{R}(\phi_{\mathsf{sum}}) \to 0$, as $d \to \infty$.

Note that Theorem 3 implies also that weak detection is possible using the sum test if

$$d\cdot\frac{\left(d_{\mathsf{KL}}(P_{XY}||Q_{XY}) + d_{\mathsf{KL}}(Q_{XY}||P_{XY})\right)^2}{\mathsf{Var}_{Q_{XY}}(\mathcal{K}(A,B))} = \Omega(1). \tag{19}$$

At this point, it is evident that based on the *bound* in Theorem 3, the sum test can achieve strong detection (vanishing risk) only if $d \to \infty$; in fact it can be shown that this is also necessary. Accordingly, we propose the following testing procedure,

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n}\mathbb{1}\left\{\frac{1}{d}\log\frac{P_{XY}^{\otimes d}(X_i,Y_j)}{Q_{XY}^{\otimes d}(X_i,Y_j)} \geq \tau_{\mathsf{count}}\right\} \geq \frac{1}{2}n\mathcal{P}_d\right\}, \tag{20}$$

where $\mathcal{P}_d \triangleq \mathbb{P}_{P_{XY}^{\otimes d}}\left[\sum_{\ell=1}^{d} \mathscr{L}(A_\ell, B_\ell) \geq d \cdot \tau_{\mathsf{count}}\right]$, where $\mathscr{L}(x,y) \triangleq \log \mathcal{L}(x,y) = \log \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$, for $(x,y) \in \mathcal{X} \times \mathcal{Y}$, and $\tau_{\mathsf{count}} \in \mathbb{R}$. Roughly speaking, ϕ_{count} counts the number of pairs whose likelihood individually exceed a certain threshold. This is similar (but not exactly the same) to a test proposed in [12] in the Gaussian setting, which counts the number of inner products between all possible (normalized) rows in X and Y who individually exceed a certain threshold. We mention here that our result holds for any natural $d \geq 1$, while in [12] it is assumed that $d \geq d_0$, for some fixed $d_0 \in \mathbb{N}$ (most notably, excluding the $d = 1$ case). We define the Chernoff's exponents $E_P, E_Q : \mathbb{R} \to [-\infty, \infty)$ as the Legendre transforms of the log-moment generating functions, namely,

$$E_Q(\theta) \triangleq \sup_{\lambda \in \mathbb{R}} \lambda\theta - \psi_Q(\lambda); \; E_P(\theta) \triangleq \sup_{\lambda \in \mathbb{R}} \lambda\theta - \psi_P(\lambda), \quad (21)$$

where $\psi_Q(\lambda) \triangleq \log \mathbb{E}_Q[\exp(\lambda \mathscr{L})]$ and $\psi_P(\lambda) \triangleq \log \mathbb{E}_P[\exp(\lambda \mathscr{L})]$. We have the following result.

Theorem 4 (Count test)**.** *Fix $d \in \mathbb{N}$, and consider the count test in (20). Suppose there is a $\tau_{\mathsf{count}} \in (-d_{\mathsf{KL}}(Q_{XY}||P_{XY}), d_{\mathsf{KL}}(P_{XY}||Q_{XY}))$ with*

$$E_Q(\tau_{\mathsf{count}}) = \omega(\log n^{1/d}), \; E_P(\tau_{\mathsf{count}}) = \omega(n^{-1} d^{-1}). \quad (22)$$

Then, $\mathsf{R}(\phi_{\mathsf{count}}) \to 0$, as $n \to \infty$.

As for the lower bounds, we provide examples for which we derive explicit formulas for the upper bounds above.

Example 3 (Gaussian databases)**.** *Consider the same setting as in Example 1. For the sum test in Theorem 3, a straightforward calculation shows that*

$$d_{\mathsf{KL}}(P_{XY}||Q_{XY}) = -\frac{1}{2}\log(1-\rho^2), \quad (23)$$

$$d_{\mathsf{KL}}(Q_{XY}||P_{XY}) = \frac{1}{2}\log(1-\rho^2) + \frac{\rho^2}{1-\rho^2}. \quad (24)$$

Furthermore,

$$\mathcal{K}(A,B) = \frac{\rho}{1-\rho^2} \cdot AB, \quad (25)$$

and thus $\mathsf{Var}_{Q_{XY}}(\mathcal{K}(A,B)) = \frac{\rho^2}{(1-\rho^2)^2}$. Therefore, the condition in (18) boils down to $d\rho^2 = \omega(1)$, which in light of Example 1 is tight up to a constant term. Finally, for fixed $d \in \mathbb{N}$, it can be shown that a sufficient condition for the count test to achieve strong detection is $\rho^2 = 1 - o(n^{-4/d})$, as $n \to \infty$. Interestingly, this bound coincided with the threshold for the recovery problem [8], achieved by the exhaustive maximum-likelihood estimator, while the count test is efficient.

Acknowledgment

This work is supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1734/21)

References

[1] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 12 763–8, Sep 2008.

[2] U. Kang, M. Hebert, and S. Park, "Fast and scalable approximate spectral graph matching for correspondence problems," *Information Sciences*, 2012.

[3] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111–125.

[4] ——, "De-anonymizing social networks," in *2009 30th IEEE Symposium on Security and Privacy*, 2009, pp. 173–187.

[5] A. Berg, T. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences." in *Proc. Computer Vision and Pattern Recognition*, 2005.

[6] T. Cour, P. Srinivasan, and J. Shi, "Balanced graph matching," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, p. 313–320.

[7] D. Cullina, P. Mittal, and N. Kiyavash, "Fundamental limits of database alignment," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE Press, 2018, p. 651–655.

[8] O. E. Dai, D. Cullina, and N. Kiyavash, "Database alignment with gaussian features," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 3225–3233.

[9] Z. K and B. Nazer, "Detecting correlated gaussian databases," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 2064–2069.

[10] ——, "Detecting correlated gaussian databases," *arXiv preprint arXiv:2206.12011*, 2022.

[11] R. Tamir, "On correlation detection and alignment recovery of Gaussian databases," *https://arxiv.org/pdf/2211.01069.pdf*, 2023.

[12] D. Elimelech and W. Huleihel, "Phase transitions in the detection of correlated databases," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 9246–9266.

[13] O. E. Dai, D. Cullina, and N. Kiyavash, "Achievability of nearly-exact alignment for correlated gaussian databases," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1230–1235.

[14] F. Shirani, S. Garg, and E. Erkip, "A concentration of measure approach to database de-anonymization," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE Press, 2019, p. 2748–2752.

[15] S. Bakirtas and E. Erkip, "Database matching under column deletions," *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 2720–2725, 2021.

[16] ——, "Database matching under column repetitions," *ArXiv*, vol. abs/2202.01730, 2022.

[17] M. Moharrami, C. Moore, and J. Xu, "The planted matching problem: Phase transitions and exact results," *The Annals of Applied Probability*, vol. 31, no. 6, pp. 2663 – 2720, 2021.

[18] J. Ding, Y. Wu, J. Xu, and D. Yang, "The planted matching problem: Sharp threshold and infinite-order phase transition," *ArXiv*, vol. abs/2103.09383, 2021.

[19] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Knowledge Discovery and Data Mining*, 2011.

[20] J. Ding, Z. Ma, Y. Wu, and J. Xu, "Efficient random graph matching via degree profiles," *Probability Theory and Related Fields*, vol. 179, pp. 29–115, 2018.

[21] Y. Wu, J. Xu, and S. H. Yu, "Testing correlation of unlabeled random graphs," *The Annals of Applied Probability*, vol. 33, no. 4, pp. 2519 – 2558, 2023.

[22] C. Mao, Y. Wu, J. Xu, and S. H. Yu, "Testing network correlation efficiently via counting trees," 2021.

[23] Y. Wu, J. Xu, and S. H. Yu, "Settling the sharp reconstruction thresholds of random graph matching," *IEEE Transactions on Information Theory*, vol. 68, no. 8, pp. 5391–5417, 2022.

[24] L. Ganassali, "Sharp threshold for alignment of graph databases with gaussian weights," in *MSML*, 2020.

[25] V. Paslev and W. Huleihel, "Testing dependency of unlabeled databases," *arXiv preprint arXiv:2311.05874*, 2023.

A Unifyied Framework to Generalization Error via Variable-size Compressibility

Milad Sefidgaran[†] and Abdellatif Zaidi [† ‡]
[†] Paris Research Center, Huawei Technologies France [‡]Université Gustave Eiffel, France
Emails: milad.sefidgaran2@huawei.com and abdellatif.zaidi@univ-eiffel.fr

***Abstract*—In this paper, we establish novel *data-dependent* upper bounds on the generalization error through the lens of a "variable-size compressibility" framework that we introduce newly here. In this framework, the generalization error of an algorithm is linked to a variable-size 'compression rate' of its input data. This is shown to yield bounds that depend on the empirical measure of the given input data at hand, rather than its unknown distribution. Our new generalization bounds that we establish are tail bounds, tail bounds on the expectation, and in-expectations bounds. Moreover, it is shown that our framework also allows to derive general bounds on *any* function of the input data and output hypothesis random variables. In particular, these general bounds are shown to subsume and possibly improve over several existing PAC-Bayes and data-dependent intrinsic dimension-based bounds that are recovered as special cases, thus unveiling a unifying character of our approach.**

I. Introduction and problem setup

Let $Z \in \mathcal{Z}$ be some *input data* distributed according to an unknown distribution μ, where $\mathcal{Z}$ is the *data space*. A major problem in statistical learning is to find a *hypothesis* (model) w in the *hypothesis space* $\mathcal{W}$ that minimizes the *population risk* defined as [1]

$$\mathcal{L}(w) := \mathbb{E}_{Z\sim\mu}[\ell(Z,w)], \quad w \in \mathcal{W}, \tag{1}$$

where $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ is a loss function that measures the quality of the prediction of the hypothesis $w \in \mathcal{W}$. The distribution μ is assumed to be unknown, however; and one has only access to n (training) samples $S = \{Z_1, \ldots, Z_n\} \sim P_S = \mu^{\otimes n}$ of the input data. Let $\mathcal{A}\colon \mathcal{S} \to \mathcal{W}$, $\mathcal{S} = \mathcal{Z}^n$, be a possibly stochastic algorithm which, for a given input data $s = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$, picks the hypothesis $\mathcal{A}(s) = W \in \mathcal{W}$. This induces a conditional distribution $P_{W|S}$ on the hypothesis space $\mathcal{W}$. Instead of the population risk minimization problem (1) one can consider minimizing the *empirical risk*, given by

$$\hat{\mathcal{L}}(s,w) := \frac{1}{n}\sum\nolimits_{i=1}^{m} \ell(z_i, w). \tag{2}$$

Nonetheless, the minimization of the empirical risk (or a regularized version of it) is meaningful only if the difference between the population and empirical risks is small enough. This difference is known as the *generalization error* of the learning algorithm and is given by

$$\operatorname{gen}(s,\mathcal{A}(s)) := \mathcal{L}(\mathcal{A}(s)) - \hat{\mathcal{L}}(s,\mathcal{A}(s)). \tag{3}$$

An exact analysis of the statistical properties of the generalization error (3) is out-of-reach, however, except in very few special cases; and, often, one resort to bounding the generalization error from the above, instead. The last two decades have witnessed the development of various such upper bounds, from different perspectives and by undertaking approaches that often appear unrelated. Common approaches include information-theoretic, compression-based, fractal-based, or intrinsic-dimension-based, and PAC-Bayes ones. Initiated by Russo and Zou [2]and Xu and Raginsky [3], the information-theoretic approach measures the complexity of the hypothesis space by the Shannon mutual information between the input data and the algorithm output. See also the follow-up works [4]–[7]. The roots of compression-based approaches perhaps date back to Littlestone and Warmuth [8] who studied the predictability of the training data labels using only part of the dataset. This compressibility approach has been extended in various ways in several works to elaborate *data-dependent* bounds. Closer to our work is another popular compressibility approach that studies the compressibility of the *hypothesis space*, see, e.g., [9], [10]. The fractal-based approach is a recently initiated line of work that hinges on that when the algorithm has a recursive nature, e.g., it involves an iterative optimization procedure, it might generate a fractal structure either in the model trajectories [11]–[14] or in its distribution [15]. These works show that, in that case, the generalization error is controlled by the intrinsic dimension of the generated fractal structure. The original PAC-Bayes bounds were stated for classification [16]; and, it has then become clear that the results could be extended to any bounded loss, resulting in many variants and extensions of them.

The aforementioned approaches have evolved independently of each other; and the bounds obtained with them differ in many ways that it is generally difficult to compare them. Arguably, however, the most useful bounds must be *computable*. This means that the bound should depend on the particular sample of the input data at hand, rather than just on the unknown data distribution. Such bounds are called *data-dependent*; they are preferred and are generally of bigger utility in practice. In this sense, most existing information-theoretic and rate-distortion theoretic-based bounds on the generalization error are data-independent. This includes the mutual information bounds of Russo and Zou [2] and Xu and Raginsky [3] whose computation requires knowledge of the joint distribution of the input data and output hypothesis; and, as such, they are not computable with just one sample training dataset at hand.

Contributions. In this paper, we establish novel *data-dependent* generalization bounds through the lens of a "variable-size compressibility" framework that we introduce here. In this framework, the generalization error of an algorithm is linked to a variable-size 'compression rate' of the input data. This allows us to derive bounds that depend on the particular empirical measure of the input data, rather than its unknown distribution. The novel generalization bounds that we establish are tail bounds, tail bounds on the expectation, and in-expectations bounds. Moreover, we show that our variable-size compressibility approach is somewhat generic and it can be used to derive general bounds on *any* function of the input data and output hypothesis random variables – In particular, see our general tail bound of Theorem 3. In fact, as we show, the framework can be accommodated easily to encompass various forms of tail bounds, tail bounds on the expectation, and in-expectation bounds through judicious choices of the distortion measure. In particular, in Section IV, by specializing them we show that our general variable-size compressibility bounds subsume various existing data-dependent PAC-Bayes and intrinsic-dimension-based bounds and recover them as special cases. Hence, another advantage of our approach is that it builds a unifying framework that allows formal connections with the aforementioned, seemingly unrelated, Rate-distortion theoretic, PAC-Bayes, and dimension-based approaches. In the extended version of this work [17], we show how using this framework, we can establish new PAC-Bayes bounds as well as new data-dependent intrinsic dimension-based bounds.

The rest of this paper is organized as follows. In Section II, we introduce our variable-size compressibility framework and provide a data-dependent tail bound on the generalization error. Section III contains our general bounds. Section IV provides various applications of our main results, in particular, to establish rate-distortion based, PAC-Bayes and dimension-based bounds. The proofs, as well as further results can be found in the extended version of this work [17].

Notations: We denote random variables, their realizations, and their alphabets by upper-case letters, lower-case letters, and calligraphy fonts; *e.g.,* X, x, and $\mathcal{X}$. The distribution, the expected value, and the support set of a random variable X are denoted as P_X, $\mathbb{E}[X]$, and $\operatorname{supp}(P_X)$. A random variable X is called σ-subgaussian, if $\log \mathbb{E}[\exp(\lambda(X-\mathbb{E}[X]))] \leqslant \lambda^2\sigma^2/2$, $\forall \lambda \in \mathbb{R}$.[1] A collection of $m \in \mathbb{N}$ random variables $(X_1,\ldots,X_m)$ is denoted as X^m or $\mathbf{X}$, when m is known by the context. The notation $\{x_i\}_{i=1}^m$ is used to represent m real numbers; used also similarly for sets or functions. We use the shorthand notation $[m]$ to denote integer ranges $1,\ldots,m \in \mathbb{N}$. Finally, the non-negative real numbers are denoted by $\mathbb{R}^+$.

Throughout we will use the following sets, defined for a random variable $X \in \mathcal{X}$ with distribution P_X and a real-valued function $g(X)\colon \mathcal{X} \to \mathbb{R}$ as

$$\mathcal{G}_X^\delta := \{\nu_X \in \mathcal{P}_\mathcal{X} \colon D_{KL}(\nu_X \| P_X) \leqslant \log(1/\delta)\}, \quad (4)$$

$$\mathcal{S}_X(g(X)) := \big\{\nu_X \in \mathcal{P}_\mathcal{X} \,\big|\, \forall x \in \operatorname{supp}(\nu_X) \colon g(x) > 0\big\}, \quad (5)$$

where $\mathcal{P}_\mathcal{X}$ is the set of all distributions defined over $\mathcal{X}$.

II. Variable-size compressibility

As we already mentioned, the approach of Sefidgaran et al. [10] is based on a fixed-size compressibility framework; and, for this reason, it only accommodates bounds on the generalization error that are independent of the data. In this work, we develop a "variable-size" compressibility framework, which is more general and allows us to establish new data-dependent bounds on the generalization error. As it will become clearer throughout, in particular, this allows us to build formal connections with seemingly-unrelated approaches such as PAC-Bayes and data-dependent intrinsic dimension bounds.

We start by recalling the aforementioned fixed-size compressibility framework, which itself can be seen as an extension of the classic compressibility framework found in source coding literature.

Consider a learning algorithm $\mathcal{A}(S)\colon \mathcal{S} \to \mathcal{W}$. The goal of the compression for the generalization error problem is to find a suitable *compressed* learning algorithm $\hat{\mathcal{A}}(S,W) = \hat{W} \in \hat{\mathcal{W}} \subseteq \mathcal{W}$ which has a smaller *complexity* than that of the original algorithm $\mathcal{A}(S)$ and whose generalization error is close enough to that of $\mathcal{A}(S)$. Define the distortion function $\tilde{d} := \mathcal{S} \times \mathcal{W} \times \hat{\mathcal{W}} \to \mathbb{R}$ as $\tilde{d}(w,\hat{w};s) := \operatorname{gen}(s,w) - \operatorname{gen}(s,\hat{w})$. In order to guarantee that $\tilde{d}(\mathcal{A}(S), \hat{\mathcal{A}}(S,\mathcal{A}(S)))$ does not exceed some desired threshold one needs to consider the *worst-case* scenario; and, in general, this results in looser bounds. Instead, they considered an adaptation of the *block-coding* technique, previously introduced in the source coding literature, for the learning algorithms. Consider a block of $m \in \mathbb{N}$ datasets $s^m = (s_1,\ldots,s_m)$ and one realization of the associated hypotheses $w^m = (w_1,\ldots,w_m)$, with $w_i = \mathcal{A}(s_i)$ for $i \in [m]$, which we denote in the rest of this paper with a slight abuse of notation as $\mathcal{A}(s^m) = w^m$. In this technique, the compressed learning algorithm $\mathcal{A}(s^m,w^m)\colon \mathcal{S}^m \times \mathcal{W}^m \to \hat{\mathcal{W}}^m$ is allowed to *jointly* compress these m instances, to produce $\hat{\mathcal{W}}^m$. Let, for given s^m the distortion between the output hypothesis of algorithm $\mathcal{A}(\cdot)$ applied on the vector s^m, i.e., $w^m = \mathcal{A}(s^m)$, and its compressed version $\hat{A}(\cdot,\cdot)$ applied on the vector (s^m,w^m), i.e., $\hat{w}^m = \hat{\mathcal{A}}(s^m,w^m)$, be the average of the element-wise distortions $\tilde{d}(\cdot,\cdot,\cdot)$ between their components, $\tilde{d}_m(w^m,\hat{w}^m;s^m) := \frac{1}{m}\sum_{i\in[m]} \tilde{d}(w_i,\hat{w}_i;s_i)$. As is easily seen, this block-coding approach with average distortion enables possibly smaller distortion levels, in comparison with those allowed by worst-case distortion over the components.

Sefidgaran et al. introduced the following definition of (exponential) compressibility, which they then used to establish *data-independent* tail and in-expectation bounds on the generalization error. Denote by $\mathbb{P}_{(S,W)^{\otimes m}}$ the probability with respect to the m-times product measure of the joint distribution of S and W.

Definition 1 ([10, Definition 8]): The learning algorithm $\mathcal{A}$ is called $(R,\epsilon,\delta;\tilde{d}_m)$-compressible[2] for some $R,\delta \in \mathbb{R}^+$

[1] All log are considered with base e in this paper.

[2] Similar to the previous work, we drop the dependence of the definition on μ, n, and $P_{W|S}$.

and $\epsilon \in \mathbb{R}$, if there exists a sequence of hypothesis books $\{\mathcal{H}_m\}_{m\in\mathbb{N}}$, $\mathcal{H}_m = \{\hat{\mathbf{w}}[j], j \in [l_m]\} \subseteq \hat{\mathcal{W}}^m$ such that $l_m \leqslant e^{mR}$ and

$$\lim_{m\to\infty}\left[-\frac{1}{m}\log\mathbb{P}_{(S,W)^{\otimes m}}\left(\min_{j\in[l_m]}\tilde{d}_m(W^m,\hat{\mathbf{w}}[j];S^m)>\epsilon\right)\right] \geqslant \log(1/\delta). \tag{6}$$

The inequality (6) expresses the condition that, for large m, the probability (over (S^m, W^m)) of finding *no* $\hat{\mathbf{w}}[j]$ that is within a distance less than ϵ from W^m vanishes faster than δ^m. Equivalently, the probability that the distance from W^m of any element $\hat{\mathbf{w}}[j]$ of the book exceeds ϵ (sometimes called probability of "excess distortion" or "covering failure") is smaller than δ^m for large m.

A result of [10, Theorem 9] states that if $\mathcal{A}$ is $(R, \epsilon, \delta; \tilde{d}_m)$-compressible in the sense of Definition 1 and the loss $\ell(Z, w)$ is σ-subgaussian for every $w \in \mathcal{W}$ then with probability $(1-\delta)$ it holds that

$$\mathrm{gen}(S, W) \leqslant \sqrt{2\sigma^2(R + \log(1/\delta))/n} + \epsilon. \tag{7}$$

Also, let $R(\delta, \epsilon) := \sup_{Q\in\mathcal{G}^{\delta}_{S,W}} \mathfrak{RD}(\epsilon; Q)$ where

$$\mathfrak{RD}(\epsilon; Q) := \inf\nolimits_{P_{\hat{W}|S}} I(S; \hat{W}), \tag{8}$$
$$\text{s.t.}\quad \mathbb{E}[\mathrm{gen}(S, W) - \mathrm{gen}(S, \hat{W})] \leqslant \epsilon,$$

the supremum is over all distributions Q over $\mathcal{S} \times \mathcal{W}$ that are in the δ-vicinity of the joint $P_{S,W}$ in the sense of (4), i.e., $Q \in \mathcal{G}^{\delta}_{S,W}$; and, in (8), the Shannon mutual information and the expectation are computed with respect to $QP_{\hat{W}|S}$. In the case in which $\mathcal{S}\times\mathcal{W}$ is discrete, a result of [10, Theorem 10] states that every algorithm $\mathcal{A}$ that induces $P_{S,W}$ is $(R(\delta, \epsilon) + \nu_1, \epsilon + \nu_2, \delta; \tilde{d}_m)$-compressible, for every $\nu_1, \nu_2 > 0$. Combined, the mentioned two results yield the following tail bound on the generalization error for the case of discrete $\mathcal{S} \times \mathcal{W}$,

$$\mathrm{gen}(S, W) \leqslant \sqrt{2\sigma^2(R(\delta, \epsilon) + \log(1/\delta))/n} + \epsilon. \tag{9}$$

It is important to note that the dependence of the tail-bound (9) on the input data S is only through the joint distribution $P_{S,W}$, not the particular realization at hand. Because of this, the approach of [10] falls short of accommodating any meaningful connection between their framework and ones that achieve data-dependent bounds such as PAC-Bayes bounds and data-dependent intrinsic dimension-based bounds. In fact, in the terminology of information-theoretic rate-distortion, the described framework can be thought of as being one for fixed-size compressibility, whereas one would here need a framework that allows *variable-size* compressibility. It is precisely such a framework that we develop in this paper. For the ease of the exposition, hereafter we first illustrate our approach and its utility for a simple case. More general results enabled by our approach will be given in the next section. To this end, define

$$d(w, \hat{w}; s) := \mathrm{gen}(s, w)^2 - \mathrm{gen}(s, \hat{w})^2,$$
$$d_m(w^m, \hat{w}^m; s^m) := \frac{1}{m}\sum\nolimits_{i\in[m]} d(w_i, \hat{w}_i; s_i). \tag{10}$$

Definition 2 (Variable-size compressibility): The learning algorithm $\mathcal{A}$ is called $(R_{S,W},\epsilon,\delta; d_m)$-compressible for some $\{R_{s,w}\}_{(s,w)\in\mathcal{S}\times\mathcal{W}}$, where $R_{s,w} \in \mathbb{R}^+$ and $R_{\max} := \sup_{s,w} R_{s,w} < \infty$, $\epsilon \in \mathbb{R}$, and $\delta \in \mathbb{R}^+$, if there exists a sequence of hypothesis books $\{\mathcal{H}_m\}_{m\in\mathbb{N}}$, $\mathcal{H}_m := \{\hat{\mathbf{w}}[j], j \in [\lfloor e^{mR_{\max}} \rfloor]\}$, such that

$$\lim_{m\to\infty}\left[-\frac{1}{m}\log\mathbb{P}_{(S,W)^{\otimes m}}\left(\min_{j} d_m(W^m, \hat{\mathbf{w}}[j]; S^m) > \epsilon\right)\right] \geqslant \log(1/\delta), \tag{11}$$

where the minimum over j is taken over $j \leqslant e^{\sum_{i\in[m]} R_{S_i,W_i}}$. The former compressibility definition (Definition 1) corresponds to $R_{s,w} := R$ for all (s, w). Comparatively, our Definition 2 here accommodates *variable-size* hypothesis books. That is, the number of hypothesis outputs of $\mathcal{H}_m$ among which one searches for a suitable *covering* of (s^m, w^m) depends on (s^m, w^m). The dependency is not only through $P_{S,W}$ but, more importantly, via the quantity $\sum_{i\in[m]} R_{S_i,W_i}$. The theorem that follows shows how this framework can be used to obtain a data-dependent tail bound on the generalization error.

Theorem 1: If the algorithm $\mathcal{A}$ is $(R_{S,W},\epsilon,\delta; d_m)$-compressible and $\forall w\in\mathcal{W}$, $\ell(Z, w)$ is σ-subgaussian, then with probability at least $(1 - \delta)$,

$$\mathrm{gen}(S, W) \leqslant \sqrt{4\sigma^2(R_{S,W} + \log(2n/\delta))/(2n - 1)} + \epsilon.$$

Note that the seemingly benign generalization to variable-size compressibility has far-reaching consequences for the tail bound itself as well as its proof. For example, notice the difference with the associated bound (7) allowed by fixed-size compressibility, especially in terms of the evolution of the bound with the size n of the training dataset. Also, investigating the proof and contrasting it with that of (7) for the fixed-size compressibility setting, it is easily seen that while for the latter it is sufficient to consider the union bound over all hypothesis vectors in $\mathcal{H}_m$, among which there exists a suitable *covering* of (S^m, W^m) with probability at least $(1-\delta)$, in our variable-size compressibility case this proof technique does not apply and falls short of producing the desired bound as the *effective* size of the hypothesis book depends on each (S^m, W^m).

Next, we establish a bound on the degree of compressibility of each learning algorithm.

Theorem 2: Suppose that the algorithm $\mathcal{A}(S) = W$ induces $P_{S,W}$ and $\mathcal{S}\times\mathcal{W}$ is a finite set. Then, for any arbitrary $\nu_1, \nu_2 > 0$, $\mathcal{A}$ is $(R_{S,W}+\nu_1, \epsilon+\nu_2, \sigma; d_m)$-compressible if the following sufficient condition holds: for any $\nu_{S,W} \in \mathcal{G}^{\delta}_{S,W}$,

$$\inf_{p_{\hat{W}|S}} \inf_{q_{\hat{W}}} \left\{ D_{KL}\left(p_{\hat{W}|S}\nu_S \| q_{\hat{W}}\nu_S\right) - D_{KL}(\nu_{S,W} \| P_{W|S}\nu_S)\right\} \leqslant \mathbb{E}_{\nu_{S,W}}[R_{S,W}], \tag{12}$$

where the first infimum is taken over all $p_{\hat{W}|S}$ that satisfy

$$\mathbb{E}_{\nu_{S,W}p_{\hat{W}|S}}\left[\mathrm{gen}(S, W)^2 - \mathrm{gen}(S, \hat{W})^2\right] \leqslant \epsilon. \tag{13}$$

Combining Theorems 1 and 2 one readily gets a potentially data-dependent tail bound on the generalization error. Because the result is in fact a special case of the more general Theorem 3 that will follow in the next section, we do not state it here. Instead, we elaborate on a useful connection with the PAC-Bayes bound of [16]. Let P be a fixed prior on $\mathcal{W}$. It is not difficult to see that the choice $R_{S,W} := \log \frac{\mathrm{d}P_{W|S}}{\mathrm{d}P}(W)$ satisfies the condition (12) for $\epsilon = 0$. The resulting tail bound recovers the PAC-Bayes bound of [18], [19], which is a *disintegrated* version of that of [16].

We hasten to mention that an appreciable feature of our approach here, which is rate-distortion theoretic in nature, is its *flexibility*, in the sense that it can be accommodated easily to encompass various forms of tail bounds, by replacing (10) with a suitable choice of the distortion measure. For example, if instead of a tail bound on the generalization error itself, ones seeks a tail bound on the expected generalization error relative to $W \sim \pi$, it suffices to consider $(R_{S,\pi}, \epsilon, \delta; d_m)$-compressibility, for some $R_{S,\pi} \in \mathbb{R}^+$, to hold when in the inequality (11) the left hand side (LHS) is substituted with

$$\lim_{m\to\infty} \left[\frac{-1}{m} \log \mathbb{P}_{S^{\otimes m}} \left(\min_j \frac{1}{m} \sum_{i\in[m]} \big(\mathbb{E}_{W_i \sim \pi_{S_i}}[\mathrm{gen}(S_i, W_i)^2] \right.\right.$$
$$\left.\left. - \,\mathrm{gen}(S_i, \hat{w}_i[j])^2\big) > \epsilon \right) \right]$$

and the inequality should hold for any choice of distributions π_S (indexed by S) over W and any distribution $\nu_S \in \mathcal{G}_S^\delta$ – Note the change of distortion measure (10) which now involves an expectation w.r.t. $W \sim \pi_S$. Using this, we obtain that with probability at least $(1-\delta)$, the following holds:

$$\forall \pi : \mathbb{E}_{W\sim\pi}[\mathrm{gen}(S,W)] \leqslant \sqrt{4\sigma^2(R_{S,\pi} + \log(2n/\delta)/(2n-1))}.$$

In the next section, we discuss how to derive a general form of this tail bound which, in particular, recovers as a special case the PAC-Bayes bound of [16].

III. General data-dependent generalization bound

In this section, we take a bigger view. We provide generic bounds, as well as proof techniques to establishing them, that are general enough and apply not only to the generalization error but also to any arbitrary function of the pair (S, W). Specifically, let $f \colon \mathcal{S} \times \mathcal{W} \to \mathbb{R}$ be a given function. We establish tail bounds on the random variable $f(S, W)$ that are *in general* [3] data-dependent. The extension of this result to generalization bounds expressed in terms of the Rényi information divergence term instead of the KL-divergence term, and also similar tail bounds on the expectation and in-expectation bounds can be found in [17]. We insist that by "data-dependent" we here mean that the bound can be computed using just one sample $S = (Z_1, \ldots, Z_n)$ and does not require knowledge of P_S. For instance, bounds that depend on (S, W) through its distribution $P_{S,W}$, such as those of [3], [10], are, in this sense, *data-independent*. Also, as it is shown in Section IV, many existing data-dependent PAC-Bayes and intrinsic dimension-based bounds can be recovered as special cases of our bounds, through judicious choices of $f(S, W)$, e.g., $f(S, W) = (\mathrm{gen}(S, W))^2$.

Theorem 3: Let $f(S, W) \colon \mathcal{S} \times \mathcal{W} \to \mathbb{R}$ and $\Delta(S, W) \colon \mathcal{S} \times \mathcal{W} \to \mathbb{R}^+$. Fix arbitrarily the set $\hat{\mathcal{W}}$ and define arbitrarily $g(S, \hat{W}) \colon \mathcal{S} \times \hat{\mathcal{W}} \to \mathbb{R}$. Then, for any $\delta \in \mathbb{R}^+$, with probability at least $1-\delta$,

$$f(S, W) \leqslant \Delta(S, W), \tag{14}$$

if for some $\epsilon \in \mathbb{R}$[4] and any $\nu_{S,W} \in \mathcal{F}_{S,W}^\delta := \mathcal{G}_{S,W}^\delta \bigcap \mathcal{S}_{S,W}(f(s,w) - \Delta(s,w))$, it holds that

$$\begin{aligned}\inf_{p_{\hat{W}|S}} \inf_{\lambda>0, q_{\hat{W}|S}} \Big\{ &\mathbb{E}_{\nu_S}\Big[D_{KL}\big(p_{\hat{W}|S} \| q_{\hat{W}|S}\big)\Big] \\ &+ \log \mathbb{E}_{P_S q_{\hat{W}|S}}\Big[e^{g(S,\hat{W})}\Big] \\ &- D_{KL}(\nu_{S,W} \| P_{W|S}\nu_S) \\ &- \lambda\big(\mathbb{E}_{\nu_{S,W}}[\Delta(S,W)] - \epsilon\big)\Big\} \leqslant \log(\delta), \end{aligned} \tag{15}$$

where ν_S is the marginal distribution of S under $\nu_{S,W}$, the first infimum is taken over all $p_{\hat{W}|S}$ that satisfy

$$\mathbb{E}_{\nu_{S,W} p_{\hat{W}|S}}\big[\Delta(S, W) - g(S, \hat{W})\big] \leqslant \epsilon. \tag{16}$$

The bound of Theorem 3 requires a condition to hold for every $\nu_{s,w} \in \mathcal{F}_{S,W}^\delta \subseteq \mathcal{G}_{S,W}^\delta$. Intuitively, this is equivalent to *covering* all sequences (S^m, W^m) whose empirical distributions Q are in the vicinity of $P_{S,W}$ in the sense of (4), using the $\hat{W}$ defined by $p_{\hat{W}|S}$. Furthermore, the distribution $q_{\hat{W}|S}$ is the one used to build (part of) the hypothesis book $\mathcal{H}_{m,Q}$ (see the proof of Theorem 2 in [17] for a definition of $\mathcal{H}_{m,Q}$).

Furthermore, in our framework of compressibility, ϵ stands for the allowed level of average distortion in (16). The specific case $\epsilon = 0$ corresponds to lossless compression; and, it is clear that allowing a non-zero average distortion level, i.e., $\epsilon \neq 0$, can yield a tighter bound (14). In fact, as will be shown in the subsequent sections, many known data-dependent PAC-Bayes and intrinsic dimension-based bounds can be recovered from the "lossless compression" case. Also, for $\epsilon = 0$ the condition (15) of the part (i.) of the theorem with the choices $\hat{\mathcal{W}} := \mathcal{W}$, $g(s, \hat{w}) := f(s, \hat{w})$, and $p_{\hat{W}|S} := \nu_{W|S}$ reduces to

$$\begin{aligned}\inf_{q_{W|S}, \lambda>0} \Bigg\{ \mathbb{E}_{\nu_{S,W}}\Bigg[\log\left(\frac{\mathrm{d}P_{W|S}}{\mathrm{d}q_{W|S}}\right) + \log \mathbb{E}_{P_S q_{W|S}}\Big[e^{\lambda f(S,W)}\Big] \\ -\lambda\Delta(S,W)\Bigg]\Bigg\} \leqslant \log(\delta). \end{aligned} \tag{17}$$

Besides, since $\mathcal{F}_{S,W}^\delta \subseteq \mathcal{G}_{S,W}^\delta$, the theorem holds if we consider all $\nu_{S,W} \in \mathcal{G}_{S,W}^\delta$. The latter set, although possibly larger, seems to be more suitable for analytical investigations. Furthermore, for any $\nu_{S,W} \in \mathcal{F}_{S,W}^\delta$, the distortion criterion

[3] However, we hasten to mention that since our approach and the resulting general-purpose bounds of this section are meant to unify several distinct approaches, some of which are data-independent, special instances of our bounds obtained by specialization to those settings can be data-independent.

[4] Although for simplicity ϵ is assumed to take a fixed value here, in general, it can be chosen to depend on $\nu_{S,W}$.

(16) is satisfied whenever

$$\mathbb{E}_{\nu_{S,W} p_{\hat{W}|S}}\big[f(S,W) - g(S,\hat{W})\big] \leqslant \epsilon. \tag{18}$$

This condition is often easier to consider, as we will see in the next section. In particular, (18) can be further simplified under the Lipschitz assumption, *i.e.*, when $\forall w, \hat{w}, s\colon |f(s,w) - g(s,\hat{w})| \leqslant \mathfrak{L}\rho(w,\hat{w})$, where $\rho\colon \mathcal{W} \times \hat{\mathcal{W}} \to \mathbb{R}^+$ is a distortion measure over $\mathcal{W} \times \hat{\mathcal{W}}$. In this case, a sufficient condition to meet the distortion criterion (16) is

$$\mathbb{E}_{\nu_{S,W} p_{\hat{W}|S}}\big[\rho(W,\hat{W})\big] \leqslant \epsilon/(2\mathfrak{L}). \tag{19}$$

IV. Applications

In this section, we show breifly how the general bound of Section III unifies various existing approaches, including rate-distortion theoretic, PAC-Bayes, and dimension-based approaches. As these have so far been thought of, and developed, largely independently of each other in the related literature, in particular, this unveils the strength and unifying character of our variable-size compression framework.

Rate-distortion theoretic bounds. As shown in [17], the tail bound of [10, Theorem 10] can be recovered and *extended* using our Theorem 3 with the choices $f(S,W) := \operatorname{gen}(S,W)$, $\hat{\mathcal{W}} \subseteq \mathcal{W}$, and $\Delta(S,W) := \Delta := \sqrt{2\sigma^2\Big(\sup_{\nu_{S,W} \in \mathcal{G}^{\delta}_{S,W}} \mathfrak{RD}(\epsilon;\nu_{S,W}) + \log(1/\delta)\Big)/n} + \epsilon$.

PAC-Bayes bounds. As shown in [17], using and extension of Theorem 3, we derive that with probability at least $(1-\delta)$,

$$\begin{aligned}\forall \pi\colon \mathbb{E}_{W\sim\pi}[f(S,W)] \leqslant & D_{KL}\big(\pi \| q_{W|S}\big) \\ & + \log \mathbb{E}_{P_S q_{W|S}}\Big[e^{f(S,W)}\Big] + \log(1/\delta).\end{aligned} \tag{20}$$

The obtained bound (20) equals that of [20, Theorem 1.ii]. Similarly, derivations using our Theorem 3 and the condition (17) allow to recover the result of [20, Theorem 1.i]. As observed by Clerico et al. [21], these recovered bounds are themselves general enough to subsume most of other existing PAC-Bayes bounds.

Our variable-size compressibility framework also allow to establish novel PAC-Bayes type bounds, presented in [17].

Dimension-based bounds. Prior to this work, the connection between compressibility and intrinsic dimension-based approaches has been established in [10]. However, as the framework introduced therein is of a “fixed-size” compressibility type and only allows establishing data-independent bounds, the connection was made only to the intrinsic dimensions of the *marginal* distributions introduced by the algorithm. This departs from most of the proposed dimension-based bounds in the related literature, which are data-dependent, i.e., they depend on a particular dimension arising for a given $S = s$. See, e.g., [11], [13], [15].

In the extended version of this work [17], we show how using our variable-size compressibility one can recover the main generalization error bound of [11], which is in terms of the *Hausdorff* dimensions of the optimization trajectories. The approach can be extended similarly to derive [15, Theorem 1]. Finally, we establish a new data-dependent intrinsic dimension-based bounds in [17], which connects the generalization error to the optimization trajectories and reveals various interesting connections with rate-distortion dimension of process, Rényi information dimension of process, and metric mean dimension.

References

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[2] D. Russo and J. Zou, “Controlling bias in adaptive data analysis using information theory,” in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Gretton and C. C. Robert, Eds., vol. 51. Cadiz, Spain: PMLR, 09–11 May 2016, pp. 1232–1240.

[3] A. Xu and M. Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] T. Steinke and L. Zakynthinou, “Reasoning about generalization via conditional mutual information,” in *Proceedings of Thirty Third Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, J. Abernethy and S. Agarwal, Eds., vol. 125. PMLR, 09–12 Jul 2020, pp. 3437–3452.

[5] A. R. Esposito, M. Gastpar, and I. Issa, “Generalization error bounds via Rényi-, f-divergences and maximal leakage,” 2020.

[6] R. Zhou, C. Tian, and T. Liu, “Individually conditional individual mutual information bound on generalization error,” *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3304–3316, 2022.

[7] S. Masiha, A. Gohari, and M. H. Yassaee, “f-divergences and their applications in lossy compression and bounding generalization error,” *IEEE Transactions on Information Theory*, 2023.

[8] N. Littlestone and M. Warmuth, “Relating data compression and learnability,” *Citeseer*, 1986.

[9] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, “Stronger generalization bounds for deep nets via a compression approach,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 254–263.

[10] M. Sefidgaran, A. Gohari, G. Richard, and U. Simsekli, “Rate-distortion theoretic generalization bounds for stochastic learning algorithms,” in *Conference on Learning Theory*. PMLR, 2022, pp. 4416–4463.

[11] U. Şimşekli, O. Sener, G. Deligiannidis, and M. A. Erdogdu, “Hausdorff dimension, heavy tails, and generalization in neural networks,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 5138–5151.

[12] T. Birdal, A. Lou, L. Guibas, and U. Şimşekli, “Intrinsic dimension, persistent homology and generalization in neural networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[13] L. Hodgkinson, U. Simsekli, R. Khanna, and M. Mahoney, “Generalization bounds using lower tail exponents in stochastic optimizers,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 8774–8795.

[14] S. H. Lim, Y. Wan, and U. Simsekli, “Chaotic regularization and heavy-tailed limits for deterministic gradient descent,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 590–26 602, 2022.

[15] A. Camuto, G. Deligiannidis, M. A. Erdogdu, M. Gurbuzbalaban, U. Simsekli, and L. Zhu, “Fractal structure and generalization properties of stochastic optimization algorithms,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 774–18 788, 2021.

[16] D. A. McAllester, “Some pac-bayesian theorems,” in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 230–234.

[17] M. Sefidgaran and A. Zaidi, “Data-dependent generalization bounds via variable-size compressibility,” 2023.

[18] G. Blanchard and F. Fleuret, “Occam’s hammer,” in *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings 20*. Springer, 2007, pp. 112–126.

[19] O. Catoni, “Pac-bayesian supervised classification,” *Lecture Notes-Monograph Series. IMS*, vol. 1277, 2007.

[20] O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor, “Pac-bayes analysis beyond the usual bounds,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 833–16 845, 2020.

[21] E. Clerico, G. Deligiannidis, B. Guedj, and A. Doucet, “A pac-bayes bound for deterministic classifiers,” *arXiv preprint arXiv:2209.02525*, 2022.

Rate-Loss Regions for Polynomial Regression with Side Information

Jiahui Wei[1,2], Philippe Mary[2], and Elsa Dupraz[1]
[1] IMT Atlantique, CNRS UMR 6285, Lab-STICC, Brest, France
[2] Univ. Rennes, INSA, IETR, UMR CNRS, Rennes, France

***Abstract*—In the context of goal-oriented communications, this paper addresses the achievable rate versus generalization error region of a learning task applied on compressed data. The study focuses on the distributed setup where a source is compressed and transmitted through a noiseless channel to a receiver performing polynomial regression, aided by side information available at the decoder. The paper provides the asymptotic rate generalization error region, and extends the analysis to the non-asymptotic regime. Additionally, it investigates the asymptotic trade-off between polynomial regression and data reconstruction under communication constraints. The proposed achievable scheme is shown to achieve the minimum generalization error as well as the optimal rate-distortion region.**

***Index Terms*—Information theory, source coding, statistical learning, rate-distortion theory, generalization error**

I. Introduction

Learning under communication constraints has received increased attention recently, for instance for distributed learning and sensor networks applications [1]. When considering a rate-limited channel, one key question is whether the design principles for the encoder and decoder for a learning task still align with those of traditional communication systems, where the main goal is data reconstruction.

To address this issue, researchers have explored simple distributed learning problems involving two correlated sources X and Y, where X is the source to be encoded and Y serves as side information at the decoder. Distributed hypothesis testing has been extensively studied for specific hypothesis tests on the joint distribution P_{XY}, and asymptotic limits on the Type-II error exponent have been determined in [2]–[4]. Furthermore, [5] demonstrated that the rate required for estimating a parameter θ from the joint distribution P_{XY} is less than the rate necessary for reconstructing the source. Finally, [6] developed a universal achievable bound on the learning generalization error, applicable to a wide range of distributed learning problems involving two sources. However, it was later shown in [7] that this bound is quite loose when applied to linear regression. Building upon [7], this paper focuses on the wider problem of polynomial regression and aims to establish achievable generalization error bounds that improve over the ones presented in [6]. Despite its simplicity, polynomial regression, captures essential learning theory concepts and is widely applied in signal and image processing, *e.g.*, [8], [9].

Morever, this paper investigates a secondary, yet significant concern, which is the trade-off between data reconstruction and learning under communication constraints. In this matter, [10] demonstrated that there indeed exists a tradeoff between data reconstruction and visual perception. Similar tradeoffs have been observed for other problems, such as hypothesis testing in [2], or identifying noisy data in a database in [11]. All previous works utilize distortion as the figure of merit for data reconstruction and employ distinct measures for the learning aspect; like a divergence between two distributions in [10], and the type-II error exponent in [2]. Unfortunately, none of these metrics are applicable to polynomial regression, underlining the need for a different analysis in our case.

Least squares regression, a fundamental statistical prediction problem, has been extensively investigated in literature. The ordinary least squares (OLS) estimator is a popular regression method, and its generalization error with k predictors and n samples is known to scale as $\frac{k}{n-k+1}$ [12]. However, this result does not take into account the communication constraint, which is an important consideration in many practical scenarios. In the context of polynomial regression, this paper determines the minimum achievable source coding rate under a constraint on the generalization error for both asymptotic and non-asymptotic regimes. The regions are derived using both standard asymptotic information theory tools [13], [14] and finite-length tools [15], and they improve over the bounds established by [6]. Additionally, the analysis reveals that no trade-off exists between data reconstruction and polynomial regression in terms of coding rate.

The outline of the paper is as follows. Section II defines the problem of coding for polynomial regression. Section III introduces the asymptotic rate-loss bounds. Section IV provides the rate-loss bounds in finite blocklength. Section V shows numerical results.

II. Problem statement

A. Notation

Throughout this article, random variables and their realizations are denoted with capital and lower-case letters, respectively, *e.g.*, X and x. Random vectors of length n are denoted $\boldsymbol{X} = [X_1, ..., X_n]^T$, and $\mathbb{E}[\boldsymbol{X}]$ and $\mathbb{C}\left[\boldsymbol{X}\right]$ are the expected value and the covariance matrix of $\boldsymbol{X}$, respectively.

This work has received a French government support granted to the Cominlabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01. This work was also funded by the Brittany region.

Next, $\underline{\boldsymbol{X}} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p]$ is a matrix gathering a p-length sequence of random vectors $\boldsymbol{X}_i$, $i \in [\![1,p]\!]$. We use Tr($\underline{\boldsymbol{X}}$) to denote the trace of matrix $\underline{\boldsymbol{X}}$, while $\lambda_{\max}(\underline{\boldsymbol{X}})$ and $\lambda_{\min}(\underline{\boldsymbol{X}})$ are the maximum and minimum eigenvalues of matrix $\underline{\boldsymbol{X}}$, respectively. We further denote $||\underline{\boldsymbol{X}}||$ as the norm-2 of a matrix $\underline{\boldsymbol{X}}$. Sets are denoted with calligraphic fonts, and if $f : \mathcal{X} \to \mathcal{Y}$ is a mapping then $|f|$ denotes the cardinality of $\mathcal{Y}$. Finally $\log(\cdot)$ denotes the base-2 logarithm.

B. Source definitions

Let $(X, Y) \sim P_{XY}$ be a pair of jointly distributed random variables, where X is the source to be encoded and Y is the side information only available at the decoder, see Figure 1. For simplicity and without loss of generality, we consider $\mathbb{E}[Y] = 0$. We define $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_{k-1}]^T \in \mathbb{R}^k$, and $\boldsymbol{Y}^\star = [Y^0, Y^1, ..., Y^{k-1}]^T \in \mathbb{R}^k$, where Y^i is the variable Y raised to power i. We assume that X follows a polynomial model of order k defined as

$$X = \sum_{i=0}^{k-1} \beta_i Y^i + N = \boldsymbol{\beta}^T \boldsymbol{Y}^\star + N, \tag{1}$$

where $N \sim \mathcal{N}(0, \sigma^2)$ follows a Gaussian distribution with mean 0 and variance σ^2. The vector $\boldsymbol{\beta}$ is constant and unknown at the transmitter.

C. Polynomial Regression

Polynomial regression aims at estimating the parameter vector $\hat{\boldsymbol{\beta}}$ from realizations, or noisy realizations, of X and Y. As a standard supervised learning problem, polynomial regression consists of two phases. We use X, Y to denote symbols generated at the training phase, and $\tilde{X}$, $\tilde{Y}$ for symbols generated at the inference phase. The training phase consists of estimating $\boldsymbol{\beta}$ from a training sequence composed by the available side information $\boldsymbol{Y}$ and by a coded version of $\boldsymbol{X}$ which is denoted $\boldsymbol{U}$. The inference phase consists of calculating estimates of the symbols $\tilde{X}$ as $\hat{X} = \hat{\boldsymbol{\beta}}\tilde{\boldsymbol{Y}}^\star$, where $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ from the training phase. Note that the inference phase does not need any data transmission, since the side information $\tilde{Y}$ is directly available to the decoder.

Following the notation introduced by Raginsky in [6], we next formalize the problem as follows. Let $\mathcal{F}$ be the set of polynomial functions $f : \mathbb{R} \to \mathbb{R}$ of the form $f(y) = \boldsymbol{\alpha}^T \boldsymbol{y}^\star$, where $\boldsymbol{\alpha} \in \mathbb{R}^k$. Polynomial regression outputs a sequence of functions $\widehat{f}^{(n)} \in \mathcal{F}$, called predictors, such that $\widehat{f}^{(n)} : \mathcal{Z}^n \times \mathbb{R} \to \mathbb{R}$, where $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{Y}) \in \mathcal{Z}^n$ is a training sequence in which $\boldsymbol{U}$ and $\boldsymbol{Y}$ are sequences of length n. Given that $\widehat{f}^{(n)} \in \mathcal{F}$, we can equivalently write

$$\widehat{f}^{(n)}(\boldsymbol{Z}, y) = \boldsymbol{\alpha}(\boldsymbol{Z})^T \boldsymbol{y}^\star, \tag{2}$$

where $\boldsymbol{\alpha} : \mathcal{Z}^n \to \mathbb{R}^k$.

Consider the quadratic loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$ defined as $\ell(x, \hat{x}) = (x - \hat{x})^2$. The minimum expected loss is defined as in [6], [7] as[1]

$$L^\star(\mathcal{F}, \boldsymbol{\beta}) = \inf_{f \in \mathcal{F}} \mathbb{E}\left[\ell(X, f(Y))\right]. \tag{3}$$

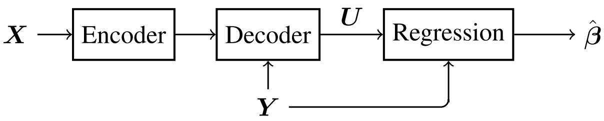


Fig. 1. Coding scheme for regression

The generalization error is defined as

$$G(\widehat{f}^{(n)}, \boldsymbol{\beta}) = \mathbb{E}_{\tilde{X}\tilde{Y}}\left[\ell\left(\tilde{X}, \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})\right) | \boldsymbol{Z}\right]. \tag{4}$$

where $(\tilde{X}, \tilde{Y}) \sim P_{XY}$ is independent from $\boldsymbol{Z}$, the training sequence. The generalization error being a random variable due to the conditioning on $\boldsymbol{Z}$, the quantity $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\widehat{f}^{(n)}, \boldsymbol{\beta}\right)\right]$ is referred to as the expected generalization error.

In the previous expressions, the minimum expected loss (3) simply expresses the average gap between X and $f(Y)$, for the function f that minimizes the quantity $\mathbb{E}\left[\ell(X, f(Y))\right]$ over the space of polynomial functions $\mathcal{F}$. However, there is no guarantee that this optimal function f can be obtained from training. On the opposite, the generalization error measures the learning performance as the expected loss for a certain training sequence $\boldsymbol{Z}$. This training sequence allows to produce an estimated function $\widehat{f}^{(n)}(\boldsymbol{Z}, \cdot)$ which can then be used to evaluate new samples $\hat{\tilde{X}} = \widehat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})$ at the inference phase. Especially, it is easy to show that $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\widehat{f}^{(n)}, \boldsymbol{\beta}\right)\right] \geq L^\star(\mathcal{F}, \boldsymbol{\beta})$. Therefore, the gap $\mathbb{E}_{\boldsymbol{Z}}\left[G\left(\widehat{f}^{(n)}, \boldsymbol{\beta}\right)\right] - L^\star(\mathcal{F}, \boldsymbol{\beta})$ is a key quantity to characterize the performance of a coding scheme dedicated to learning, and this is why our rate-learning regions will be expressed from this quantity.

D. Coding scheme

The coding scheme is analogue to the one for linear regression in [7]. However, the theoretical analysis differs and becomes more complex, as will be described in the next sections.

Definition 1. *A polynomial regression scheme at rate R is defined by a sequence $\{(e_n, d_n, R, \hat{f}^{(n)}))\}$ with an encoder $e_n : \mathcal{X}^n \to [\![1, M_n]\!]$ a decoder $d_n : \mathcal{Y}^n \times [\![1, M_n]\!] \to \mathcal{U}^n$ and the learner $t_n : \mathcal{Y}^n \times \mathcal{U}^n \to \mathcal{F}$ such that*

$$\limsup_{n \to \infty} \frac{\log M_n}{n} \leq R.$$

Definition 2. *An (n, M, l, ε) code for the sequence $\{(e_n, d_n, R, \hat{f}^{(n)})\}$ and $\varepsilon \in (0, 1)$ is a code with $|e_n| = M$ such that*

$$\mathbb{P}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta}) \geq l\right] \leq \varepsilon \text{ and } \frac{\log M}{n} \leq R. \tag{5}$$

Definition 3. *For fixed l and blocklength n, the finite blocklength rate-loss functions with excess loss ε is defined by:*

$$R(n, l, \varepsilon) = \inf_R \{\exists \;\; (n, M, l, \varepsilon) \text{ code}\} \tag{6}$$

[1] One may also define a loss over a sequence. However, since the samples from the training and inference phases are i.i.d. it does not change the analysis.

Definition 4. *A pair (R, δ) is said to be achievable if there exists a sequence $\{(e_n, d_n, R, \hat{f}^{(n)})\}$ such that*

$$\limsup_{n\to\infty} \mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{\beta})\right] \leq L^*(\mathcal{F}, \boldsymbol{\beta}) + \delta \tag{7}$$

As discussed in Section II-C, the achievable region is defined in terms of gap between $\mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^{(n)}, \boldsymbol{\beta})\right]$ and $L^*(\mathcal{F}, \boldsymbol{\beta})$. Although the regions defined in this section pertain to rate-generalization error regions, for the sake of simplicity and with a minor deviation in terminology, we refer to them as rate-loss regions in the subsequent discussions.

III. Asymptotic Bound on the Rate-Loss function

In [6, Theorem 3.3], it is shown that, for a quadratic loss function, the generalization error can be bounded as:

$$\begin{aligned} L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{\beta}) \leq \limsup_{n\to\infty} \mathbb{E}\left[G(\hat{f}^{(n)}, \boldsymbol{\beta})^{\frac{1}{2}}\right] \leq L^{\star\frac{1}{2}}(\mathcal{F}, \boldsymbol{\beta}) \\ + 2\mathbb{D}_{X|Y}(R)^{1/2} \end{aligned} \tag{8}$$

where $\mathbb{D}_{X|Y}(R)$ is the conditional distortion-rate function. It can be shown that for the polynomial regression, the minimum expected loss in (3) is $L^\star(\mathcal{F}, \boldsymbol{\beta}) = \sigma^2$. In this section, we build a coding scheme that allows to improve the upper bound in (8) for the polynomial model.

A. Rate-loss region

Theorem 1. *Given any rate $R > 0$, the pair $(R, 0)$ is achievable for the polynomial regression scheme with squared loss, for sources (X, Y) following the polynomial model* (1).

This result states that the minimum generalization error which is given by the loss function $L^\star(\mathcal{F}, \boldsymbol{\beta})$ in (8) can be achieved with any arbitrary rate R, as long as the training sequence is long enough. The proof of the Theorem is based on an achievability scheme built on a Gaussian test channel. This test channel is known for being optimal for joint Gaussian sources when considering data reconstruction [16], although it may be suboptimal for other models like the one we consider in this paper. However, in our case, we show that this test channel achieves the optimal rate-loss region $(R, 0)$ for polynomial regression, and we further discuss its optimality for data reconstruction in Section III-C.

B. Proof of Theorem 1 : Achievability scheme

Let us consider the test channel $U = \alpha(X + \Phi)$, where $\Phi \sim \mathcal{N}(0, \sigma_\Phi^2)$ is independent of X, and α and σ_Φ^2 are two parameters which depend on the distribution of X and Y.

The parameters $\boldsymbol{\beta}$ and the joint distribution P_{XY} are unknown to the encoder and decoder but the noise variance of the model, i.e. σ^2, is assumed to be known at the encoder. Hence, the transmission rate is perfectly known at the encoder and the variable-rate scheme in [14] becomes a fixed-rate coding scheme in our setup. The same idea of binning is used and the de-binning is performed based on the empirical mutual information between $\boldsymbol{x}$ and $\boldsymbol{u}$ evaluated thanks to the type of $\boldsymbol{x}$ transmitted in a prefix transmission [14]. Given that $D < \sigma_x^2$ and $(X \mid Y)$ is Gaussian, we show that the rate-distortion function $R_b(D) = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{\sigma_\Phi^2}\right)$ is achievable for $\mathbb{E}_{XU}\left[d(X, U)\right] \leq D$, where D is a function of σ_Φ^2.

Then, for a training sequences $(\boldsymbol{y}, \boldsymbol{u})$, the OLS estimator $\hat{\boldsymbol{\beta}}$ is given by [17, Chapter 7]

$$\hat{\boldsymbol{\beta}} = \alpha^{-1}(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T})^{-1} \underline{\boldsymbol{Y}}^\star \boldsymbol{u}. \tag{9}$$

where $\underline{\boldsymbol{Y}}^\star = [\boldsymbol{Y}_{\boldsymbol{1}}^\star, ..., \boldsymbol{Y}_{\boldsymbol{n}}^\star] \in \mathbb{R}^{k\times n}$ and this estimator has the following statistical properties :

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} \;\text{ and }\; \mathbb{C}\left[\hat{\boldsymbol{\beta}}|\boldsymbol{Y}\right] = \frac{1}{\alpha^2}\sigma^2_{U|Y}(\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T})^{-1} \tag{10}$$

where $\mathbb{C}\left[\hat{\boldsymbol{\beta}}|\boldsymbol{Y}\right]$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{Y}$. Hence, the generalization error (4) can be rewritten as

$$\begin{aligned} G(\widehat{f}^{(n)}, \boldsymbol{\beta}) &= \mathbb{E}_{\tilde{X}\tilde{Y}}\left[[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}]^T \tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star^T}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \boldsymbol{N}^T\boldsymbol{N}|\boldsymbol{Z}\right] \\ &= [\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}]^T \mathbb{E}_{\tilde{Y}}\left[\tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star^T}\right][\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \sigma^2. \end{aligned} \tag{11}$$

Let $\tilde{\underline{\boldsymbol{\Sigma}}} = \mathbb{E}_{\tilde{Y}}\left[\tilde{\boldsymbol{Y}}^\star \tilde{\boldsymbol{Y}}^{\star^T}\right]$ and $\underline{\boldsymbol{\Sigma}} = \frac{1}{n}\underline{\boldsymbol{Y}}^\star \underline{\boldsymbol{Y}}^{\star T}$. Then, the expected generalization error is

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta})\right] \\ &= \sigma^2 + \mathbb{E}\left[\frac{1}{n}(\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Y}}^\star(\boldsymbol{N} + \boldsymbol{\Phi}))^T \tilde{\underline{\boldsymbol{\Sigma}}} \frac{1}{n}(\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Y}}^\star(\boldsymbol{N} + \boldsymbol{\Phi}))\right] \\ &= \sigma^2 + \frac{\sigma^2 + \sigma_\Phi^2}{n}\mathbb{E}\left[\mathrm{Tr}\left(\tilde{\underline{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right)\right]. \end{aligned} \tag{12}$$

The next step is to show that $\mathbb{E}\left[\mathrm{Tr}\left(\tilde{\underline{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right)\right]$ is bounded by some constant C for n large enough. The following proposition bounds the trace of a product of two matrices by their eigenvalues.

Proposition 1. *[18, p 340] (Ruhe's trace inequality). If $\underline{\boldsymbol{U}}$ and $\underline{\boldsymbol{V}}$ are $k \times k$ positive semidefinite Hermitian matrices with eigenvalues $\lambda_i(\underline{\boldsymbol{U}}), \lambda_i(\underline{\boldsymbol{V}})$, $i \in \{1, \cdots, k\}$ then*

$$\mathrm{Tr}(\underline{\boldsymbol{U}}\underline{\boldsymbol{V}}) \leq \sum_{i=1}^{k} \lambda_i(\underline{\boldsymbol{U}})\lambda_i(\underline{\boldsymbol{V}}) \tag{13}$$

Lemma 1. *If $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{B}}$ are real symmetric matrices, then:*

$$\lambda_{\min}(\underline{\boldsymbol{A}}) \geq \lambda_{\min}(\underline{\boldsymbol{B}}) - ||\underline{\boldsymbol{A}} - \underline{\boldsymbol{B}}|| \tag{14}$$

Proof: Let $\boldsymbol{x}$ be a vector such that $||x||_2 = 1$, by Cauchy-Schwartz inequality, for a real symmetric matrix $\underline{\boldsymbol{M}}$, we have

$$-||\underline{\boldsymbol{M}}|| \leq \boldsymbol{x}^T\underline{\boldsymbol{M}}\boldsymbol{x} \leq ||\underline{\boldsymbol{M}}||. \tag{15}$$

With the properties of eigenvalues, we have

$$\lambda_{\min}(\underline{\boldsymbol{M}}) \leq \boldsymbol{x}^T\underline{\boldsymbol{M}}\boldsymbol{x} \leq \lambda_{\max}(\underline{\boldsymbol{M}}). \tag{16}$$

For real symmetric matrices $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{B}}$, we have

$$\boldsymbol{x}^T\underline{\boldsymbol{A}}\boldsymbol{x} = \boldsymbol{x}^T\underline{\boldsymbol{B}}\boldsymbol{x} + \boldsymbol{x}^T(\underline{\boldsymbol{A}} - \underline{\boldsymbol{B}})\boldsymbol{x}. \tag{17}$$

Applying the above inequalities shows the desired result. ■

We remark that $\underline{\boldsymbol{\Sigma}}$ is an estimator of the covariance matrix of $\boldsymbol{Y}$. Then, from Proposition 1 and Lemma 1, for n large enough, Tr $\left(\underline{\tilde{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right)$ is bounded almost surely by:

$$\text{Tr}\left(\underline{\tilde{\boldsymbol{\Sigma}}}\underline{\boldsymbol{\Sigma}}^{-1}\right) \leq k\frac{\lambda_{\max}(\underline{\tilde{\boldsymbol{\Sigma}}})}{\lambda_{\min}(\underline{\tilde{\boldsymbol{\Sigma}}}) - ||\underline{\tilde{\boldsymbol{\Sigma}}} - \underline{\boldsymbol{\Sigma}}||}. \tag{18}$$

Substituting this into (12) with some constant $C = \frac{\lambda_{\max}(\underline{\tilde{\boldsymbol{\Sigma}}})}{\lambda_{\min}(\underline{\tilde{\boldsymbol{\Sigma}}})}$ and the fact that $||\underline{\tilde{\boldsymbol{\Sigma}}} - \underline{\boldsymbol{\Sigma}}|| \to 0$ almost surely, shows that the expected generalization error is upper bounded by

$$\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta})\right] \leq \sigma^2 + \frac{(\sigma^2 + \sigma_\Phi^2)}{n}kC \tag{19}$$

Thus $\mathbb{E}_{\boldsymbol{Z}}\left[G(\widehat{f}^{(n)}, \boldsymbol{\beta})\right] \to \sigma^2$ as $n \to \infty$, which completes the proof.

Our result closes the gap between the lower bound and the upper bound from [6] (see equation (8)). In order to provide a bound applicable to a wide range of problems, the upper bound from [6] considered both the observation noise between $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the distortion between $\boldsymbol{X}$ and $\boldsymbol{U}$. While in our result, by the Gaussian test channel and OLS estimation from $\boldsymbol{U}$ and $\boldsymbol{Y}$, we show that the quantification error term in (19), and hence the distortion term, is vanishing with the blocklength n.

C. Trade-off between data reconstruction and polynomial regression

In this section, we show that the previous achievability scheme considered for polynomial regression also achieves the optimal Wyner-Ziv rate-distortion function for data reconstruction, for sources modeled by (1).

Corollary 1. *For a pair of sources (X, Y) modeled from* (1)*, there is no trade-off in terms of coding rate between distortion and polynomial regression generalization error.*

Proof: We first investigate the conditional setup in which the side information Y is also available at the encoder. Since the random variable $(X|Y) \sim \mathcal{N}(0, \sigma^2)$, the following conditional rate-distortion function can be achieved [19]

$$R_{X|Y}(D) = \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right), \tag{20}$$

where $D = \mathbb{E}\left[(X - \hat{X})^2\right]$ is the distortion. We now show that in the Wyner-Ziv setup where Y is only available at the decoder, the rate-distortion function $R_{\text{WZ}}(D)$ is equal to $R_{X|Y}(D)$ when considering the same test channel $U = \alpha(X + \Phi)$ as in the proof of Theorem 1, with $\alpha = \frac{\sigma^2 - D}{\sigma^2}$, and $\sigma_\Phi^2 = \frac{D\sigma^2}{\sigma^2 - D}$. By using the proposed achievability scheme, the random variable U can be recovered perfectly at the decoder, and then produces $\hat{X} = U + (1-\alpha)\boldsymbol{\beta}^T\boldsymbol{Y}^\star$. This allows us to evaluate $\mathbb{E}\left[(X - \hat{X})^2\right] = (\alpha - 1)^2\sigma^2 + \alpha^2\sigma_\Phi^2$. Replacing α and σ_Φ^2 by their expressions leads to $\mathbb{E}\left[(X - \hat{X})^2\right] = D$. Second, the Wyner-Ziv rate-distortion function has expression [16]

$$I(X;U) - I(Y;U) = \frac{1}{2}\log_2\left(\frac{\sigma^2 + \sigma_\Phi^2}{\sigma_\Phi^2}\right)$$

where the equality comes from the fact that N and Φ are Gaussian random variables. Replacing σ_Φ^2 by its expression gives that $R_{\text{WZ}}(D) = R_{X|Y}(D)$ in (20), which shows that the Gaussian test channel is optimal when considering our polynomial source model. Note that in the previous derivation, we considered that $\boldsymbol{\beta}$ is perfectly known. If this is not the case, $\hat{X}$ is computed from $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$, and following the same derivation as for the generalization error permits to show that $\mathbb{E}\left[(X - \hat{X})^2\right] \to D$ as $n \to \infty$.

This result differs from the other ones in literature that show that there is a tradeoff between reconstruction and learning, such as for the hypothesis testing problem for instance [2].

■

IV. Rate-Loss non-asymptotic bound

In the finite-blocklength regime, not all codewords satisfy the generalization error constraint, and hence the excess probability, defined in Definition 2, has to be taken into account. The characterization of the non-asymptotic achievable bound for the rate-generalization error region is built from the rate-distortion problem in finite blocklength regime, studied in [15]. Similarly, we define the information-loss density vector as follows:

$$\boldsymbol{i}(U, X, Y, \tilde{X}, \tilde{Y}) := \begin{bmatrix} -\log\dfrac{P_{U|Y}(U|Y)}{P_U(U)} \\ \log\dfrac{P_{U|X}(U|X)}{P_U(U)} \\ \ell(\tilde{X}, \hat{f}^{(n)}(\boldsymbol{Z}, \tilde{Y})) \end{bmatrix} \tag{21}$$

where the third term is specific to our non-linear regression problem. The expectation of $\boldsymbol{i}$ over the distribution $P_{UXY\tilde{X}\tilde{Y}}$ is $\boldsymbol{J} = \left[-I(U;Y), I(U;X), \mathbb{E}_{\boldsymbol{Z}}\left[G(\hat{f}^n, \boldsymbol{\beta})\right]\right]^T$, where the sum of the first two components gives the Wyner-Ziv coding rate. The covariance matrix of (21) is

$$\boldsymbol{V} = \mathbb{C}\left[\boldsymbol{i}(U, X, Y, \tilde{X}, \tilde{Y})\right]. \tag{22}$$

Let k be a positive integer and $\boldsymbol{V} \in \mathcal{R}^{k\times k}$ be a positive-semi-definite matrix. Given a Gaussian random vector $\boldsymbol{B} \sim \mathcal{N}(0, \boldsymbol{V})$, the dispersion region is [20]

$$\mathscr{S}(\boldsymbol{V}, \varepsilon) := \{\boldsymbol{b} \in \mathbb{R}^k : \Pr(\boldsymbol{B} \leq \boldsymbol{b}) \geq 1 - \varepsilon\}. \tag{23}$$

By replacing the distortion measure by the generalization error, and adapting some steps of the analysis, we can obtain a similar result as Theorem 2 in [15]. Finally, by applying this theorem in conjunction with the multidimensional Berry-Esséen Theorem, we show that for all $0 < \varepsilon < 1$ and n sufficiently large, the (n, ε)-rate-generalization error function satisfies:

$$R_{\text{b}}(n, \varepsilon, l) \leq \inf\left\{\boldsymbol{M}^T\left(\boldsymbol{J} + \frac{\mathscr{S}(\boldsymbol{V}, \varepsilon)}{\sqrt{n}} + \frac{2\log n}{n}\mathbf{1}_3\right)\right\} \tag{24}$$

with $\boldsymbol{M} = [1 \quad 1 \quad 0]^T$ and $\mathbf{1}_3 = [1 \;\; 1 \;\; 1]^T$.

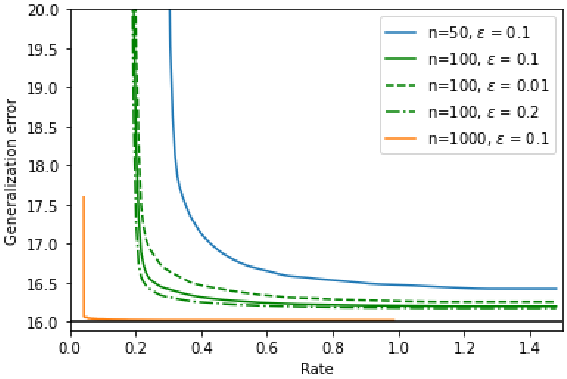


Fig. 2. Non-asymptotic rate-generalization error region labeled on the blocklength n and the excess loss probability ε.

V. Numerical results

Let us consider $X = \beta_0 + \beta_1 Y + \beta_2 Y^2 + N$, and assume that Y is uniform over $[-1, 1]$. We also set $\boldsymbol{\beta} = [2, 3, 1]^T$ and $\sigma^2 = 16$. From the theorem of change variable, for $\beta_2 > 0$ and $\beta_1^2 + 4\beta_2(v - \beta_0) \geq 0$, the distribution of $V = \boldsymbol{\beta}^T \mathbf{Y}^\star$ is:

$$P_V(v) = \begin{cases} \frac{1}{\sqrt{\beta_1^2+4\beta_2(v-\beta_0)}} & |y_1(v)| \leq 1 \text{ and } |y_2(v)| \leq 1 \\ \frac{1}{2\sqrt{\beta_1^2+4\beta_2(v-\beta_0)}} & |y_1(v)| \leq 1 \text{ or } |y_2(v)| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $y_1 = \frac{-\beta_1 - \sqrt{\beta_1^2+4\beta_2(v-\beta_0)}}{2\beta_2}, y_2 = \frac{-\beta_1 + \sqrt{\beta_1^2+4\beta_2(v-\beta_0)}}{2\beta_2}$. The probability density function of $U = \alpha(V + N + \Phi)$ can then be expressed as

$$P_U(u) = \frac{1}{\alpha\sqrt{2\pi(\sigma^2 + \sigma_\Phi^2)}} \int_{-\infty}^{\infty} P_V(v) e^{-\frac{(\frac{u}{\alpha} - v)^2}{2(\sigma^2+\sigma_\Phi^2)}} dv \quad (25)$$

which can be evaluated numerically. Using (24) with $(U|Y) \sim \mathcal{N}(0, \alpha^2(\sigma^2 + \sigma_\Phi^2))$ and $(U|X) \sim \mathcal{N}(0, \alpha^2\sigma_\Phi^2)$, we can estimate the information-density-loss vector by generating a large number of samples, and thus estimate the dispersion region in (23). Figure 2 shows the boundaries of the achievable rate-loss region for different parameters n and ε. The black line represents the best achievable generalization error, i.e. σ^2. We observe that the achievable region enlarges when the source size, n, or the excess probability increases. Indeed, when the excess probability is larger, the proportion of codewords which exceeds the generalization error constraint is larger, and this situation occurs for smaller rate. Moreover, for a fixed excess probability, increasing n allows to reduce the rate since the poorly reconstructed U is compensated by the large number of samples for estimating the regression parameters. These results do not deal with an outer bound at finite blocklength, i.e. a rate-loss region that *cannot* be exceeded, and the region outside the boundary needs further investigation.

VI. Conclusion

This paper provided achievable rate-generalization error regions for the polynomial regression problem in both asymptotic and non-asymptotic regimes. An important result of our study states that asymptotically there is no trade-off between data reconstruction and polynomial regression under communication constraints. The characterization of the outer bound (converse) for the rate-generalization error region is also of great interest and would allow to refine the analysis. The developed framework could be extended to more complex learning taks, such as non-parametric estimation, in the future.

References

[1] E. C. Strinati and S. Barbarossa, “6g networks: Beyond shannon towards semantic and goal-oriented communications,” *Computer Networks*, vol. 190, p. 107930, 2021.

[2] G. Katz, P. Piantanida, and M. Debbah, “Distributed binary detection with lossy data compression,” *IEEE Transactions on information Theory*, vol. 63, no. 8, pp. 5207–5227, 2017.

[3] S. Salehkalaibar, M. Wigger, and L. Wang, “Hypothesis testing over the two-hop relay network,” *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4411–4433, 2019.

[4] S. Sreekumar and D. Gündüz, “Distributed hypothesis testing over discrete memoryless channels,” *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2044–2066, 2020.

[5] M. El Gamal and L. Lai, “Are Slepian-Wolf rates necessary for distributed parameter estimation?” in *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015, pp. 1249–1255.

[6] M. Raginsky, “Learning from compressed observations,” in *2007 IEEE Information Theory Workshop*, 2007, pp. 420–425.

[7] J. Wei, E. Dupraz, and P. Mary, “Asymptotic and non-asymptotic rate-loss bounds for linear regression with side information,” in *31st European Signal Processing Conference, EUSIPCO*, 2023.

[8] E. Siggiridou and D. Kugiumtzis, “Dimension reduction of polynomial regression models for the estimation of granger causality in high-dimensional time series,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 5638–5650, 2021.

[9] G. D. Finlayson, M. Mackiewicz, and A. Hurlbert, “Color correction using root-polynomial regression,” *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1460–1470, 2015.

[10] Y. Blau and T. Michaeli, “Rethinking lossy compression: The rate-distortion-perception tradeoff,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 675–685.

[11] E. Tuncel and D. Gündüz, “Identification and lossy reconstruction in noisy databases,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 822–831, 2014.

[12] J. Mourtada, “Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices,” *The Annals of Statistics*, vol. 50, no. 4, 2022.

[13] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Transactions on information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[14] S. C. Draper, “Universal incremental Slepian-Wolf coding,” in *42nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Citeseer, 2004, pp. 1332–1341.

[15] S. Watanabe, S. Kuzuoka, and V. Y. Tan, “Nonasymptotic and second-order achievability bounds for coding with side-information,” *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1574–1605, 2015.

[16] A. D. Wyner, “The rate-distortion function for source coding with side information at the decoder-ii: General sources,” *Information and control*, vol. 38, pp. 60–80, 1978.

[17] A. C. Rencher and G. B. Schaalje, *Linear models in statistics*. John Wiley & Sons, 2008.

[18] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and its Applications*, 2nd ed. Springer, 2011, vol. 143.

[19] R. M. Gray, “Conditional rate-distortion theory,” Stanford Univ CA Stanford Electronic Labs, Tech. Rep., 1972.

[20] V. Y. F. Tan and O. Kosut, “On the dispersions of three network information theory problems,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 881–903, 2014.

An Information-Spectrum Approach to Distributed Hypothesis Testing for General Sources

Ismaila Salihou Adamou[1], Elsa Dupraz[1], and Tad Matsumoto[1,2]
[1] IMT Atlantique, CNRS UMR 6285, Lab-STICC, Brest, France
[2] JAIST and University of Oulu (Emeritus)

***Abstract*—This paper investigates Distributed Hypothesis testing (DHT), in which a source $\mathbf{X}$ is encoded given that side information $\mathbf{Y}$ is available at the decoder only. Based on the received coded data, the receiver aims to decide on the two hypotheses H_0 or H_1 related to the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. While most existing contributions in the literature on DHT consider i.i.d. assumptions, this paper assumes more generic, non-i.i.d., non-stationary, and non-ergodic sources models. It relies on information-spectrum tools to provide general formulas on the achievable Type-II error exponent under a constraint on the Type-I error. The achievability proof is based on a quantize-and-binning scheme. It is shown that with the quantize-and-binning approach, the error exponent boils down to a trade-off between a binning error and a decision error, as already observed for the i.i.d. sources. The last part of the paper provides error exponents for particular source models, *e.g.*, Gaussian, stationary, and ergodic models.**

I. Introduction

In distributed communication networks, data is gathered from various remote nodes and then sent to a server for further processing. Often, the primary objective of the server is not to reconstruct the data, but instead to make a decision based on the collected data. This type of setup is known as distributed hypothesis testing (DHT), and it was first investigated from an information-theoretic perspective in [1], [2].

In DHT, a source $\mathbf{X}$ is encoded using side information $\mathbf{Y}$ available only to the decoder, as shown in Figure 1. The receiver aims to make a decision between two hypotheses: H_0, where the joint probability distribution of $(\mathbf{X},\mathbf{Y})$ is $P_{\mathbf{XY}}$, and H_1, where the joint distribution is $P_{\overline{\mathbf{XY}}}$. Hypothesis testing involves two types of errors, called the Type-I error and the Type-II error [3]. The information-theoretic analysis of DHT aims to determine the achievable error exponent for the Type-II error while keeping the Type-I error below a fixed threshold [1], [2].

Previous contributions on DHT typically assume that the sources $\mathbf{X}$ and $\mathbf{Y}$ generate independent and identically distributed (i.i.d.) pairs of symbols (X_t, Y_t) [4]–[8]. For example, [7] and [8] provide the error exponent achieved by a quantize-and-binning scheme for i.i.d. sources. Some more complex source models have been investigated in [9], [10], which assume that the sources $\mathbf{X}$ and $\mathbf{Y}$ generate pairs of Gaussian vectors $(\mathbf{X}_t^M, \mathbf{Y}_t^N)$ with auto-correlations in each vector $\mathbf{X}_t^M$ and $\mathbf{Y}_t^N$, as well as cross-correlation between them. However, the models of [9], [10] are block-i.i.d. in the sense that the successive pairs $(\mathbf{X}_t^M, \mathbf{Y}_t^N)$ are assumed to be i.i.d. with t.

Nevertheless, i.i.d. and block-i.i.d. models are often inadequate for capturing the statistics of signals like time series or videos, which cannot be decomposed into fixed-length independent blocks and are frequently non-stationary and/or non-ergodic. As a result, the objective of this paper is to consider a more general source model that is non-i.i.d. and can account for non-stationary and non-ergodic signals, while still encompassing the previous models as particular instances. To investigate DHT under these conditions, we utilize information spectrum tools, which were first introduced in [11] and generally provide information theory results that are applicable to a broad range of source models. It should be noted that information spectrum has been previously used for hypothesis testing in [12], but only for the encoding of a source $\mathbf{X}$ alone, without the use of side information $\mathbf{Y}$.

In this paper, we investigate DHT using general source models for $\mathbf{X}$ and $\mathbf{Y}$ and provide an achievability scheme that yields a general expression for the Type-II error exponent. Our approach to the achievability scheme builds upon the quantize-and-binning techniques presented in [8], while taking into account the use of side information for more complex source models. As in [8], the resulting error exponent consists of two terms: one for the binning error and the other for the decision error. We then specialize our error-exponent to source models of interest, including (i) i.i.d. sources, for which we recover the error exponent reported in [8]; (ii) non-i.i.d. stationary and ergodic sources in general; and (iii) non-i.i.d. Gaussian stationary and ergodic sources.

The outline of the paper is as follows. Section II describes the general sources model and restates the DHT problem. Section III provides the achievable error exponent for general sources, and Section IV derives the proof. Section V considers some examples of source models.

II. Problem statement

In what follows, $[\![1,n]\!]$ denotes the set of integers between 1 and n. We also use upper-case letters, *e.g*, X, to denote random variables (RVs) and lower-case letters, *e.g*, x, to denote their realizations. Random sequences of length n are denoted $\mathbf{X}^n = (X_1, X_2, \cdots, X_n)$.

This work has received a French government support granted to the Cominlabs excellence laboratory and managed by the National Research Agency in the “Investing for the Future” program under reference ANR-10-LABX-07-01.

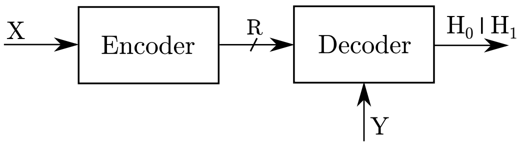


Figure 1. Distributed Hypothesis Testing coding scheme

A. General Sources

In the DHT problem shown in Fig.1, the encoder observes a source sequence $\mathbf{X}$, and the decoder receives a coded version of $\mathbf{X}$ as well as a side information sequence $\mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are correlated. We consider that the sequences $\mathbf{X}$ and $\mathbf{Y}$ are produced from two general sources which are not necessarily i.i.d., and not even stationary or ergodic. As in [12], we define general sources $\mathbf{X}$ and $\mathbf{Y}$ as two infinite sequences :

$$\begin{aligned}\mathbf{X} &= \{\mathbf{X}^n = (X_1, X_2, \cdots, X_n)\}_{n=1}^{\infty},\\ \mathbf{Y} &= \{\mathbf{Y}^n = (Y_1, Y_2, \cdots, Y_n)\}_{n=1}^{\infty}\end{aligned} \tag{1}$$

of n-dimensional random variables $\mathbf{X}^n, \mathbf{Y}^n$, respectively. Each component random variable X_i, Y_i, $i \in [\![1, n]\!]$, takes values in a finite source alphabet $\mathcal{X}, \mathcal{Y}$, respectively. Next, $P_{\mathbf{X}^n}$ is the probability distribution of the length-n vector $\mathbf{X}^n$, and $P_{\mathbf{X}} = \{P_{\mathbf{X}^n}\}_{n=1}^{\infty}$ is the collection of all probability distributions $P_{\mathbf{X}^n}$. The same holds for the source $\mathbf{Y}$.

We now describe two particular cases of (1). The first one consists of a scalar i.i.d. model in which the sequences $\mathbf{X}^n$ and $\mathbf{Y}^n$ come from two i.i.d. sources, *i.e.*, the successive pairs of symbols (X_n, Y_n) are independent and distributed according to the same joint distribution P_{XY}. This model was considered for DHT in [7], [8]. The second case still relies on an i.i.d. model but for source vectors. In this case, the source sequences $\mathbf{X}^n$ and $\mathbf{Y}^n$ are defined as

$$\mathbf{X}^n = \left\{\mathbf{X}_t^M\right\}_{t=1}^{n}, \; \mathbf{Y}^n = \left\{\mathbf{Y}_t^M\right\}_{t=1}^{n}, \tag{2}$$

where $\{\mathbf{X}_t^M\}_{t=1}^n$ and $\{\mathbf{Y}_t^M\}_{t=1}^n$ are sequences of i.i.d. M-dimensional random vectors and the successive pairs $\left(\mathbf{X}_t^M, \mathbf{Y}_t^M\right)$ are distributed according to the same joint distribution $P_{\mathbf{X}^M\mathbf{Y}^M}$. The i.i.d. property of the successive M-length vectors simplifies the DHT analysis by allowing for an orthogonal transform to be applied onto the successive independent blocks $\mathbf{X}_t^M$ and $\mathbf{Y}_t^M$ [9], [10]. Our model described in (1) is more general since it considers infinite sequences without the i.i.d. assumption.

B. Distributed Hypothesis Testing

In what follows, we consider that the joint distribution of the sequence pair $\{(\mathbf{X}^n, \mathbf{Y}^n)\}_{n=1}^{\infty}$ depends on the underlying hypotheses H_0 and H_1 defined as

$$H_0 : (\mathbf{X}^n, \mathbf{Y}^n) \sim P_{\mathbf{X}^n\mathbf{Y}^n}, \tag{3}$$

$$H_1 : (\mathbf{X}^n, \mathbf{Y}^n) \sim P_{\overline{\mathbf{X}}^n\overline{\mathbf{Y}}^n}. \tag{4}$$

where the marginal probability distributions $P_{\mathbf{X}^n}$ and $P_{\mathbf{Y}^n}$ do not depend on the hypothesis.

We consider the following usual coding scheme defined in the literature on DHT [1], [8].

Definition 1: The encoding function $f^{(n)}$ and decoding function $g^{(n)}$ are defined as

$$f^{(n)} : \mathcal{X}^n \longrightarrow \mathcal{M}_n = [\![1, \mathrm{M}_2]\!], \tag{5}$$

$$g^{(n)} : \mathcal{M}_n \times \mathcal{Y}^n \longrightarrow \mathcal{H} = \{H_0, H_1\}, \tag{6}$$

such that $\limsup_{n\to\infty} \frac{1}{n}\log \mathrm{M}_2 \leq \mathrm{R}$, where R is the rate and M_2 is the cardinality of the alphabet set $\mathcal{M}_n$.

Definition 2: The Type-I and Type-II error probabilities α_n and β_n are defined as

$$\alpha_n = \mathbb{P}\left[g^{(n)}\left(f^{(n)}\left(\mathbf{X}^n\right), \mathbf{Y}^n\right) = H_1 \mid H_0 \text{ is true}\right], \tag{7}$$

$$\beta_n = \mathbb{P}\left[g^{(n)}\left(f^{(n)}\left(\mathbf{X}^n\right), \mathbf{Y}^n\right) = H_0 \mid H_1 \text{ is true}\right]. \tag{8}$$

Definition 3: For given The Type-II error exponent θ is said to be achievable for a given rate R, if for large blocklength n, there exists encoding and decoding functions $\left(f^{(n)}, g^{(n)}\right)$ such that the Type-I and Type-II error probabilities α_n and β_n satisfy

$$\alpha_n \leq \epsilon, \tag{9}$$

and

$$\limsup_{n\to\infty} \frac{1}{n}\log\frac{1}{\beta_n} \geq \theta \tag{10}$$

for any $\epsilon > 0$.
In the following, we aim to determine the achievable Type-II error exponent θ for general sources.

III. Main result: Error exponent

A. Definitions

We first provide some definitions which will be useful to express our main result. The $\limsup$ and $\liminf$ in probability of a sequence $\{Z_n\}_{n=1}^{\infty}$ are, respectively, defined as [11]

$$\mathrm{p} - \limsup_{n\to\infty} Z_n = \inf\left\{\alpha \mid \lim_{n\to+\infty} \mathbb{P}\left(Z_n > \alpha\right) = 0\right\}, \tag{11}$$

$$\mathrm{p} - \liminf_{n\to\infty} Z_n = \sup\left\{\alpha \mid \lim_{n\to+\infty} \mathbb{P}\left(Z_n < \alpha\right) = 0\right\}. \tag{12}$$

The spectral sup-mutual information $\overline{I}(\mathbf{X};\mathbf{U})$, the spectral inf-mutual information $\underline{I}(\mathbf{U};\mathbf{Y})$, the spectral inf-divergence rate $\underline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right)$, and the spectral sup-divergence rate $\overline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right)$ are, respectively, defined as [11]

$$\overline{I}(\mathbf{X};\mathbf{U}) = \mathrm{p} - \limsup_{n\to\infty} \frac{1}{n}\log\frac{P_{\mathbf{U}^n|\mathbf{X}^n}\left(\mathbf{U}^n \mid \mathbf{X}^n\right)}{P_{\mathbf{U}^n}\left(\mathbf{U}^n\right)}, \tag{13}$$

$$\underline{I}(\mathbf{U};\mathbf{Y}) = \mathrm{p} - \liminf_{n\to\infty} \frac{1}{n}\log\frac{P_{\mathbf{U}^n|\mathbf{Y}^n}\left(\mathbf{U}^n \mid \mathbf{Y}^n\right)}{P_{\mathbf{U}^n}\left(\mathbf{U}^n\right)}, \tag{14}$$

$$\underline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right) = \mathrm{p} - \liminf_{n\to\infty} \frac{1}{n}\log\frac{P_{\mathbf{U}^n\mathbf{Y}^n}\left(\mathbf{U}^n, \mathbf{Y}^n\right)}{P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}\left(\mathbf{U}^n, \mathbf{Y}^n\right)}, \tag{15}$$

$$\overline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right) = \mathrm{p} - \limsup_{n\to\infty} \frac{1}{n}\log\frac{P_{\mathbf{U}^n\mathbf{Y}^n}\left(\mathbf{U}^n, \mathbf{Y}^n\right)}{P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}\left(\mathbf{U}^n, \mathbf{Y}^n\right)}. \tag{16}$$

B. Achievable error-exponent for general sources

Theorem 1: For the coding scheme of Definition 1, the following error exponent θ is achievable for general sources defined by (1):

$$\theta \leq \min \Bigg\{ r - \left(\overline{I}(\mathbf{X};\mathbf{U}) - \underline{I}(\mathbf{U};\mathbf{Y})\right), \\ \underline{D}\left(P_{\mathbf{UY}} \| P_{\overline{\mathbf{UY}}}\right) + \left(\underline{I}(\mathbf{X};\mathbf{U}) - \overline{I}(\mathbf{X};\mathbf{U})\right) \Bigg\}, \quad (17)$$

where $\mathbf{U}$ is an auxiliary random variable with same conditional distribution $P_{\mathbf{U}|\mathbf{X}} = P_{\overline{\mathbf{U}}|\overline{\mathbf{X}}}$ under H_0 and H_1 and such that the Markov chain $\mathbf{U} \to \mathbf{X} \to \mathbf{Y}$ is satisfied under both H_0 and H_1. In addition, $P_{\mathbf{UY}}$, and $P_{\overline{\mathbf{UY}}}$ are the joint distributions of $(\mathbf{U}^n, \mathbf{Y}^n)$ under H_0 and H_1, respectively, and $r \leq \mathrm{R}$.
As expected, we find that our error exponent is consistent with that shown in [8] for the i.i.d. case. The error exponent (17) is the result of a trade-off between the binning error and the decision error, as in the i.i.d. case [7], [8]. The binning strategy introduces a new type of error event that does not appear in the DHT scheme without binning for general sources of [12]. In addition, the decision error, *e.g.*, the second term in (17), not only contains a divergence term that appears in [7], [8] and related works, but also the difference $\underline{I}(\mathbf{X};\mathbf{U}) - \overline{I}(\mathbf{X};\mathbf{U})$ between the spectral inf-mutual information and the spectral sup-mutual information of $\mathbf{X}$ and $\mathbf{U}$. Especially, if the term $\frac{1}{n}\log\frac{P_{\mathbf{U}^n|\mathbf{X}^n}(\mathbf{U}^n|\mathbf{X}^n)}{P_{\mathbf{U}^n}(\mathbf{U}^n)}$ does not converge in probability, then the two mutual information terms differ, inducing a penalty in the error exponent. For stationary and ergodic sources, this term converges and there is no such penalty.

IV. Proof of Theorem 1

We first restate the following lemma from [13], which will be useful in the proof.

Lemma 1 ([13]): Let $\mathbf{Z}^n, \mathbf{X}^n, \mathbf{U}^n$, be random sequences which take values in finite sets $\mathcal{Z}^n, \mathcal{X}^n, \mathcal{U}^n$, respectively, and satisfy the Markov condition $\mathbf{U}^n \to \mathbf{X}^n \to \mathbf{Z}^n$. Let $\{\Psi_n\}_{n=1}^{\infty}$ be a sequence of mappings such that $\Psi_n : \mathcal{Z}^n \times \mathcal{U}^n \to \{0,1\}$, and

$$\lim_{n\to\infty} \mathbb{P}\left(\Psi_n(\mathbf{Z}^n, \mathbf{U}^n) = 1\right) = 0. \quad (18)$$

Then, $\forall \varepsilon > 0$, there exists a sequence $\{f_n\}_{n=1}^{\infty}$ of mappings $f_n : \mathcal{X}^n \to \{\mathbf{u}_i^n\}_{i=1}^{\mathrm{M}} \subset \mathcal{U}^n$ such that $\mathrm{M} = \lceil e^{n(\overline{I}(\mathbf{U};\mathbf{X})+\varepsilon)} \rceil$ and

$$\lim_{n\to\infty} \mathbb{P}\left(\Psi_n(\mathbf{Z}^n, f_n(\mathbf{X}^n)) = 1\right) = 0. \quad (19)$$

A. Coding scheme

Random codebook generation: Generate $\mathrm{M}_1 = e^{n\overline{r}_0}$ sequences $\mathbf{u}_i^n$ randomly according to a fixed distribution $P_{\mathbf{U}^n|\mathbf{X}^n}$. Assign randomly each $\mathbf{u}_i^n$ to one of $\mathrm{M}_2 = e^{nr}$ bins according to a uniform distribution over $[\![1, \mathrm{M}_2]\!]$. Let $\boldsymbol{B}(\mathbf{u}_i^n) \in [\![1, \mathrm{M}_2]\!]$ denote the index of the bin to which $\mathbf{u}_i^n$ belongs to.

Encoder : Given the sequence $\mathbf{x}^n$, the encoder uses a pre-defined mapping $f_n : \mathcal{X}^n \to \{\mathbf{u}_i^n\}_{i=1}^{\mathrm{M}_1}$ to output a certain sequence $\mathbf{u}_i^n = f_n(\mathbf{x}^n)$ and checks if the condition $(\mathbf{x}^n, \mathbf{u}_\mathbf{i}^\mathbf{n}) \in T_n^{(1)}$ is satisfied, where

$$T_n^{(1)} = \quad (20) \\ \left\{ (\mathbf{x}^n, \mathbf{u}^n) \text{ s.t. } \underline{r}_0 - \epsilon < \frac{1}{n}\log\frac{P_{\mathbf{U}^n|\mathbf{X}^n}(\mathbf{u}^n \mid \mathbf{x}^n)}{P_{\mathbf{U}^n}(\mathbf{u}^n)} < \overline{r}_0 + \epsilon \right\}$$

where $\underline{r}_0, \overline{r}_0 \in \mathbb{R}$. If such a sequence is found, the encoder sends the bin index $\boldsymbol{B}(\mathbf{u}_i^n)$. Otherwise, it sends an error message.

Decoder : The decoder first looks for a sequence in the bin according to the joint distribution $P_{\mathbf{U}^n\mathbf{Y}^n}$ under H_0. Given the received bin index and the side information $\mathbf{y}^n$, going over the sequences $\mathbf{u}^n$ in the bin one by one, the decoder checks whether $(\mathbf{y}^n, \mathbf{u}^n) \in T_n^{(2)}$ with

$$T_n^{(2)} = \left\{ (\mathbf{y}^n, \mathbf{u}^n) \text{ s.t. } \frac{1}{n}\log\frac{P_{\mathbf{U}^n|\mathbf{Y}^n}(\mathbf{u}^n \mid \mathbf{y}^n)}{P_{\mathbf{U}^n}(\mathbf{u}^n)} > r^{'} - \epsilon \right\}, \quad (21)$$

with $r' \in \mathbb{R}$. The decoder declares H_1 if no such sequence is found in the bin or if it receives an error message from the encoder. Otherwise, it declares H_0 if the sequence $\mathbf{u}^n$ extracted from the bin belongs to the acceptance region $\mathcal{A}_n$ defined as

$$\mathcal{A}_n = \left\{ (\mathbf{y}^n, \mathbf{u}^n) \text{ s.t. } \frac{1}{n}\log\frac{P_{\mathbf{U}^n\mathbf{Y}^n}(\mathbf{u}^n, \mathbf{y}^n)}{P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}(\mathbf{u}^n, \mathbf{y}^n)} > S - \epsilon \right\}, \quad (22)$$

where $S \in \mathbb{R}$ is the decision threshold; if otherwise, it declares H_1. The sets $T_n^{(1)}$, $T_n^{(2)}$, and $\mathcal{A}_n$ can be seen as decision regions depending on threshold values $\overline{r}_0, \overline{r}_0, r'$ and S. Those parameters will be chosen such that $\alpha_n \leq \epsilon$, for any $\epsilon > 0$.

B. Error probability analysis

Type-I error α_n : The error events with which the decoder declares H_1 under H_0 are as follows:

$$E_{11} = \Bigg\{ \nexists \mathbf{u}^n \text{ s.t. } (\mathbf{X}^n, \mathbf{u}^n) \in T_n^{(1)}, (\mathbf{Y}^n, \mathbf{u}^n) \in T_n^{(2)}, \\ (\mathbf{Y}^n, \mathbf{u}^n) \in \mathcal{A}_n \Bigg\}, \quad (23)$$

$$E_{12} = \Bigg\{ \exists \mathbf{u'}^n \neq \mathbf{u}^n \text{ s.t. } \boldsymbol{B}(\mathbf{u'}^n) = \boldsymbol{B}(\mathbf{u}^n), \left(\mathbf{Y}^n, \mathbf{u'}^n\right) \in T_n^{(2)}, \\ \text{but } \left(\mathbf{Y}^n, \mathbf{u'}^n\right) \notin \mathcal{A}_n \Bigg\}. \quad (24)$$

The first event E_{11} is when there is an error either in the encoding, during debinning, or when taking the decision. The second event E_{12} corresponds to a debinning error, where a wrong sequence is extracted from the bin. By the union-bound, the Type-I error probability α_n can be upper bounded as

$$\alpha_n \leq \mathbb{P}(E_{11}) + \mathbb{P}(E_{12}). \quad (25)$$

Regarding the first error event, for $\underline{r}_0 = \underline{I}(\mathbf{X};\mathbf{U})$, $\overline{r}_0 = \overline{I}(\mathbf{X};\mathbf{U})$, and from the definitions of $\underline{I}(\mathbf{X};\mathbf{U})$ and $\overline{I}(\mathbf{X};\mathbf{U})$ in (13) and (14), we have

$$\lim_{n\to\infty} \mathbb{P}\left((\mathbf{X}^n, \mathbf{U}^n) \notin T_n^{(1)} \right) = 0.$$

In addition, according to the definition of $\underline{I}(\mathbf{Y};\mathbf{U})$ in (14), and setting $r^{'}=\underline{I}(\mathbf{Y};\mathbf{U})$, we also have

$$\lim_{n\to\infty}\mathbb{P}\left((\mathbf{Y}^n,\mathbf{U}^n)\notin T_n^{(2)}\right)=0. \tag{26}$$

Finally, when $S=\underline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right)$ and from the definition of $\underline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right)$, we have

$$\lim_{n\to\infty}\mathbb{P}\left((\mathbf{Y}^n,\mathbf{U}^n)\notin\mathcal{A}_n\right)=0.$$

Thus, by defining

$$\Psi_n(\mathbf{x}^n,\mathbf{y}^n,\mathbf{u}^n)= \tag{27}$$
$$\begin{cases} 0, & \text{if } (\mathbf{x}^n,\mathbf{u}^n)\in T_n^{(1)}, (\mathbf{y}^n,\mathbf{u}^n)\in T_n^{(2)} \text{ and} \\ & (\mathbf{y}^n,\mathbf{u}^n)\in\mathcal{A}_n, \\ 1, & \text{otherwise.}\end{cases}$$

we get that $\mathbb{P}(\Psi_n(\mathbf{X}^n,\mathbf{Y}^n,\mathbf{U}^n)=1)\to 0$ as $n\to\infty$. Then, given that $\mathbf{U}^n\to\mathbf{X}^n\to\mathbf{Y}^n$ forms a Markov chain, applying Lemma 1 allows to show that there exists a sequence of functions f_n such that $\mathbb{P}(E_{11})\to 0$ as $n\to\infty$.

Then, the error probability $\mathbb{P}\left(E_{12}\right)$ can be expressed as

$$\begin{aligned}\mathbb{P}\left(E_{12}\right)&\leq\sum_{\mathbf{y}^n}P_{\mathbf{Y}^n}\left(\mathbf{y}^n\right)\sum_{\substack{\mathbf{u}'^n:\mathbf{u}'^n\neq\mathbf{u}^n\\(\mathbf{y}^n,\mathbf{u}'^n)\in\mathcal{T}_n^{(2)}\cap\overline{\mathcal{A}_n}}}\mathbb{P}\left(\boldsymbol{B}(\mathbf{u}'^n)=\boldsymbol{B}(\mathbf{u}^n)\right)\\&\leq\sum_{\mathbf{y}^n}P_{\mathbf{Y}^n}\left(\mathbf{y}^n\right)\sum_{\substack{\mathbf{u}'^n:\mathbf{u}'^n\neq\mathbf{u}^n\\(\mathbf{y}^n,\mathbf{u}'^n)\in\mathcal{T}_n^{(2)}}}e^{-nr}\end{aligned} \tag{28}$$

From (21), for $\left(\mathbf{y}^n,\mathbf{u}'^n\right)\in T_n^{(2)}$ we get

$$P_{\mathbf{Y}^n}\left(\mathbf{y}^n\right)<P_{\mathbf{Y}^n|\mathbf{U}^n}\left(\mathbf{y}^n\mid\mathbf{u}'^n\right)e^{-n(r'-\epsilon)},$$

which allows us to write

$$\begin{aligned}\mathbb{P}\left(E_{12}\right)&\leq\sum_{\mathbf{u}'^n}\sum_{\mathbf{y}^n:(\mathbf{y}^n,\mathbf{u}'^n)\in\mathcal{T}_n^{(2)}}P_{\mathbf{Y}^n|\mathbf{U}^n}\left(\mathbf{y}^n\mid\mathbf{u}'^n\right)e^{-n(r+r'-\epsilon)}\\&\leq e^{-n(r+r'-\overline{r}_0-\epsilon)}\end{aligned} \tag{29}$$

where $e^{n\overline{r}_0}$ is the number of sequences $\mathbf{u}^n$ in the codebook. Therefore, from the condition $r\geq\overline{r}_0-r'+\epsilon=\overline{I}(\mathbf{X};\mathbf{U})-\underline{I}(\mathbf{Y};\mathbf{U})+\epsilon$, we get that $\mathbb{P}\left(E_{21}\right)\to 0$ as $n\to\infty$.

Type-II error β_n : A Type-II error occurs when the decoder declares H_0 although H_1 is the true hypothesis. The corresponding error events are:

$$\begin{aligned}E_{21}&=\left\{\exists\tilde{\mathbf{u}}^n\neq\mathbf{u}^n:\boldsymbol{B}(\tilde{\mathbf{u}}^n)=\boldsymbol{B}(\mathbf{u}^n),\left(\overline{\mathbf{Y}}^n,\tilde{\mathbf{u}}^n\right)\in T_n^{(2)},\right.\\&\qquad\left.\text{and}\left(\overline{\mathbf{Y}}^n,\tilde{\mathbf{u}}^n\right)\in\mathcal{A}_n\right\},\\E_{22}&=\left\{(\overline{\mathbf{Y}}^n,\mathbf{u}^n)\in T_n^{(2)},(\overline{\mathbf{Y}}^n,\mathbf{u}^n)\in\mathcal{A}_n\right\}.\end{aligned} \tag{30}$$

The first event E_{21} is a debinning error and the second event E_{22} is the testing error. By the union bound, we get

$$\beta_n\leq\mathbb{P}\left(E_{21}\right)+\mathbb{P}\left(E_{22}\right). \tag{31}$$

Since the marginal probability distribution $P_{\mathbf{Y}^n}$ does not depend on the hypothesis, the probability $\mathbb{P}\left(E_{21}\right)$ can be expressed by following the same steps as for $\mathbb{P}\left(E_{12}\right)$. Given that $\overline{r}_0=\overline{I}(\mathbf{X};\mathbf{U})$ and $r^{'}=\underline{I}(\mathbf{Y};\mathbf{U})$, we get

$$\mathbb{P}\left(E_{21}\right)\leq e^{-n\left(r-\left(\overline{I}(\mathbf{X};\mathbf{U})-\underline{I}(\mathbf{Y};\mathbf{U})\right)-\varepsilon\right)}. \tag{32}$$

Next, the probability $\mathbb{P}\left(E_{22}\right)$ can be expressed as

$$\begin{aligned}\mathbb{P}\left(E_{22}\right)&\leq\sum_{(\mathbf{x}^n,\mathbf{y}^n)}P_{\overline{\mathbf{X}}^n\overline{\mathbf{Y}}^n}\left(\mathbf{x}^n,\mathbf{y}^n\right)\sum_{\substack{\mathbf{u}^n\in[\![1,\mathrm{M}_1]\!],\\(\mathbf{x}^n,\mathbf{u}^n)\in\mathcal{T}_n^{(1)}}}\mathbb{P}\left((\mathbf{y}^n,\mathbf{u}^n)\in\mathcal{A}_n\right)\\&\leq e^{n\overline{r}_0}\sum_{(\mathbf{x}^n,\mathbf{y}^n)}P_{\overline{\mathbf{X}}^n\overline{\mathbf{Y}}^n}\left(\mathbf{x}^n,\mathbf{y}^n\right)\sum_{\substack{\mathbf{u}^n:\\(\mathbf{x}^n,\mathbf{u}^n)\in\mathcal{T}_n^{(1)}\\(\mathbf{y}^n,\mathbf{u}^n)\in\mathcal{A}_n}}P_{\mathbf{U}^n}(\mathbf{u}^n)\end{aligned}$$

Since $(\mathbf{x}^n,\mathbf{u}^n)\in\mathcal{T}_n^{(1)}$,

$$P_{\mathbf{U}^n}(\mathbf{u}^n)<P_{\mathbf{U}^n|\mathbf{X}^n}(\mathbf{u}^n\mid\mathbf{x}^n)e^{-n(\underline{r}_0-\epsilon)}.$$

In addition, the conditional distributions $P_{\mathbf{U}^n|\mathbf{X}^n}$ and $P_{\overline{\mathbf{U}}^n|\overline{\mathbf{X}}^n}$ are the same, and the Markov chain $\mathbf{U}^n\to\mathbf{X}^n\to\mathbf{Y}^n$ is satisfied. Thus, $P_{\mathbf{U}^n|\mathbf{X}^n}=P_{\overline{\mathbf{U}}^n|\overline{\mathbf{X}}^n,\overline{Y}^n}$, and

$$\mathbb{P}\left(E_{22}\right)\leq e^{n(\overline{r}_0-\underline{r_0}+\epsilon)}\sum_{\mathbf{u}^n:(\mathbf{y}^n,\mathbf{u}^n)\in\mathcal{A}_n}P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}\left(\mathbf{u}^n,\mathbf{y}^n\right). \tag{33}$$

For $(\mathbf{y}^n,\mathbf{u}^n)\in\mathcal{A}_n$, we have

$$P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}\left(\mathbf{u}^n,\mathbf{y}^n\right)<P_{\mathbf{U}^n\mathbf{Y}^n}\left(\mathbf{u}^n,\mathbf{y}^n\right)e^{-n(S-\epsilon)}. \tag{34}$$

Combining this with (33) gives that

$$\mathbb{P}\left(E_{22}\right)\leq e^{-n(\underline{r}_0-\overline{r_0}+S-2\epsilon)} \tag{35}$$

Now, substituting (32) and (35) into (31), with $S=\underline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right)$, the Type-II error is upper-bounded as

$$\begin{aligned}\beta_n\leq&e^{-n\left(r-\left(\overline{I}(\mathbf{X};\mathbf{U})-\underline{I}(\mathbf{Y};\mathbf{U})\right)-\epsilon\right)}\\&+e^{-n\left(\underline{I}(\mathbf{X};\mathbf{U})-\overline{I}(\mathbf{X};\mathbf{U})+\underline{D}\left(P_{\mathbf{UY}}\|P_{\overline{\mathbf{UY}}}\right)-2\epsilon\right)}.\end{aligned} \tag{36}$$

Finally, from the definition of the error exponent θ given by (10), we show that (17) is achievable, which proves Theorem 1.

V. Examples

A. Stationary and ergodic sources

We now apply Theorem 1 to sources which are stationary and ergodic, but not necessarily i.i.d.

Proposition 1: If the sources $\mathbf{X}^n$ and $\mathbf{Y}^n$ are stationary and ergodic under both H_0 and H_1, the error exponent (17) becomes :

$$\begin{aligned}\theta\leq\min\Bigg\{&\lim_{n\to\infty}r-\left[\frac{1}{n}h\left(\mathbf{U}^n\mid\mathbf{Y}^n\right)-\frac{1}{n}h\left(\mathbf{U}^n\mid\mathbf{X}^n\right)\right],\\&\lim_{n\to\infty}\frac{1}{n}D\left(P_{\mathbf{U}^n\mathbf{Y}^n}\|P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}\right)\Bigg\}.\end{aligned} \tag{37}$$

This proposition is due to the *strong converse property* [11].

B. Stationary and ergodic Gaussian sources

Let $\mathbf{X}$ and $\mathbf{Y}$ be two stationary and ergodic sources distributed according to Gaussian distributions $\mathcal{N}\left(\mu_{\mathbf{X}}, \mathbf{K}_{\mathbf{X}}\right)$ and $\mathcal{N}\left(\mu_{\mathbf{Y}}, \mathbf{K}_{\mathbf{Y}}\right)$, with covariance matrices $\mathbf{K}_{\mathbf{X}}$ and $\mathbf{K}_{\mathbf{Y}}$, respectively. The two hypotheses are formulated as

$$H_0 : \left(\begin{array}{c} \mathbf{X}^n \\ \mathbf{Y}^n \end{array}\right) \sim \mathcal{N}(\mu_{\mathbf{XY}}, \mathbf{K}), \tag{38}$$

$$H_1 : \left(\begin{array}{c} \mathbf{X}^n \\ \mathbf{Y}^n \end{array}\right) \sim \mathcal{N}(\overline{\mu}_{\mathbf{XY}}, \overline{\mathbf{K}}). \tag{39}$$

In the expressions (38) and (39), $\mu_{\mathbf{XY}}$ is defined as a block vector $\left[\mu_{\mathbf{X}}, \mu_{\mathbf{Y}}\right]^T$. In addition, $\mathbf{K}$ and $\overline{\mathbf{K}}$ are the joint covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ defined as

$$\mathbf{K} = \left[\begin{array}{cc} \mathbf{K}_{\mathbf{X}} & \mathbf{K}_{\mathbf{XY}} \\ \mathbf{K}_{\mathbf{YX}} & \mathbf{K}_{\mathbf{Y}} \end{array}\right], \overline{\mathbf{K}} = \left[\begin{array}{cc} \mathbf{K}_{\mathbf{X}} & \overline{\mathbf{K}}_{\mathbf{XY}} \\ \overline{\mathbf{K}}_{\mathbf{YX}} & \mathbf{K}_{\mathbf{Y}} \end{array}\right], \tag{40}$$

We assume that all the matrices $\mathbf{K}_{\mathbf{X}}$, $\mathbf{K}_{\mathbf{Y}}$, $\overline{\mathbf{K}}_{\mathbf{Y}}$, $\mathbf{K}_{\mathbf{XY}}$, and $\overline{\mathbf{K}}_{\mathbf{XY}}$ are positive-definite. We also denote the conditional covariance matrix of $\mathbf{X}^n$ given $\mathbf{Y}^n$ by

$$\mathbf{K}_{\mathbf{X}|\mathbf{Y}} = \mathbf{K}_{\mathbf{X}} - \mathbf{K}_{\mathbf{XY}} \mathbf{K}_{\mathbf{Y}}^{-1} \mathbf{K}_{\mathbf{XY}}. \tag{41}$$

The eigenvalues of $\mathbf{K}_{\mathbf{X}|\mathbf{Y}}$ are further denoted by $\lambda_i^{(X|Y)}$.

Proposition 2: If the sources $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian, stationary, and ergodic, under both H_0 and H_1, the terms in (37) reduce to

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} h\left(\mathbf{U}^n \mid \mathbf{Y}^n\right) - \lim_{n\to\infty} \frac{1}{n} h\left(\mathbf{U}^n \mid \mathbf{X}^n\right) = \\ &\qquad \lim_{n\to\infty} \frac{1}{2n} \sum_{i=1}^{n} \log \frac{\lambda_i^{(X|Y)} + \kappa}{\kappa}, \end{aligned} \tag{42}$$

and

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} D\left(P_{\mathbf{U}^n\mathbf{Y}^n} \| P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}\right\} = \lim_{n\to\infty} \frac{1}{2n} \left[\log \frac{|\overline{\mathbf{\Sigma}}|}{|\mathbf{\Sigma}|} - 2n + \right. \\ &\left. \left(\overline{\mu}_{\mathbf{UY}} - \mu_{\mathbf{UY}}\right)^T \overline{\mathbf{\Sigma}}^{-1} \left(\overline{\mu}_{\mathbf{UY}} - \mu_{\mathbf{UY}}\right) + \operatorname{tr}\left\{\overline{\mathbf{\Sigma}}^{-1} \mathbf{\Sigma}\right\}\right], \end{aligned} \tag{43}$$

where $\mathbf{\Sigma}$ and $\overline{\mathbf{\Sigma}}$ are the joint covariance matrices of $\mathbf{U}$ and $\mathbf{Y}$ under H_0 and H_1, respectively.

The terms given by (42) and (43) are obtained by considering that the source $\mathbf{U}$ is Gaussian such that $\mathbf{U} = \mathbf{X} + \mathbf{Z}$, where $\mathbf{Z} \sim \mathcal{N}\left(0, \kappa \mathbf{I}_n\right)$ is independent of $\mathbf{X}$, and $\mathbf{I}_n$ is the identity matrix of dimension $n \times n$. The covariance matrices $\mathbf{\Sigma}$ and $\overline{\mathbf{\Sigma}}$ are then defined as

$$\mathbf{\Sigma} = \left[\begin{array}{cc} \mathbf{K}_{\mathbf{U}} & \mathbf{K}_{\mathbf{UY}} \\ \mathbf{K}_{\mathbf{YU}} & \mathrm{K}_{\mathbf{Y}} \end{array}\right], \overline{\mathbf{\Sigma}} = \left[\begin{array}{cc} \mathbf{K}_{\mathbf{U}} & \overline{\mathbf{K}}_{\mathbf{UY}} \\ \overline{\mathbf{K}}_{\mathbf{YU}} & \mathrm{K}_{\mathbf{Y}} \end{array}\right]. \tag{44}$$

We now consider the case where the pair $(\mathbf{U}, \mathbf{Y})$ has different covariance matrices, $\mathbf{\Sigma}$ under H_0 and $\overline{\mathbf{\Sigma}}$ under H_1. We also assume that all the Gaussian vectors are zero-centered. We then define H_0 and H_1 as

$$H_0 : \left(\begin{array}{c} \mathbf{X}^n \\ \mathbf{Y}^n \end{array}\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}),\ H_1 : \left(\begin{array}{c} \mathbf{X}^n \\ \mathbf{Y}^n \end{array}\right) \sim \mathcal{N}(\mathbf{0}, \overline{\mathbf{K}}).$$

In this case, it can be shown that the expression (42) remains the same, while the expression (43) reduces to

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} D\left(P_{\mathbf{U}^n\mathbf{Y}^n} \| P_{\overline{\mathbf{U}}^n\overline{\mathbf{Y}}^n}\right\} = \lim_{n\to\infty} \frac{1}{2n} &\left[\log \frac{|\overline{\mathbf{\Sigma}}|}{|\mathbf{\Sigma}|} \right. \\ &\left. -2n + \operatorname{tr}\left\{\overline{\mathbf{\Sigma}}^{-1} \mathbf{\Sigma}\right\}\right]. \end{aligned} \tag{45}$$

The matrices $\mathbf{\Sigma}$ and $\overline{\mathbf{\Sigma}}$ are of length $2n \times 2n$, where n tends to infinity. Therefore, for some specific Gaussian sources, one needs to study the convergence of the determinants $|\mathbf{\Sigma}|$ and $|\overline{\mathbf{\Sigma}}|$, and also of the trace $\operatorname{tr}\left\{\overline{\mathbf{\Sigma}}^{-1} \mathbf{\Sigma}\right\}$.

VI. Conclusion

We provided an information-spectrum approach to DHT for general non-i.i.d., non-stationary, and non-ergodic sources. The derived error exponent is achieved from a quantize-and-binning scheme, which, we found, boils down to a trade-off between a binning error and a decision error. Future works will focus on comparing our error exponent to state-of-the-art ones obtained very recently for the i.i.d. case [14]. Other future works will include designing practical coding schemes, as well as considering other applications of DHT such as synchronism identification in spread spectrum signal detectors [15].

References

[1] T. S. Han, “Hypothesis Testing with Multiterminal Data Compression,” *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[2] R. Ahlswede and I. Csiszár, “Hypothesis testing with communication constraints,” *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[3] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*. Springer, 2005, vol. 3.

[4] G. Katz, P. Piantanida, and M. Debbah, “Distributed Binary Detection with Lossy Data Compression,” *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5207–5227, 2017.

[5] S. Salehkalaibar and M. Wigger, “Distributed hypothesis testing over multi-access channels,” in *IEEE Global Communications Conference (Globecom)*, 2018, pp. 1–6.

[6] S. Sreekumar and D. Gunduz, “Distributed Hypothesis Testing over Discrete Memoryless Channels,” *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2044–2066, 2020.

[7] M. S. Rahman and A. B. Wagner, “On the optimality of binning for distributed hypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6282–6303, 2012.

[8] G. Katz, P. Piantanida, R. Couillet, and M. Debbah, “On the necessity of binning for the distributed hypothesis testing problem,” *IEEE Int. Symp. Inf. Theory - Proc.*, pp. 2797–2801, 2015.

[9] M. S. Rahman and A. B. Wagner, “Vector gaussian hypothesis testing and lossy one-helper problem,” *IEEE Int. Symp. Inf. Theory - Proc.*, pp. 968–972, 2009.

[10] P. Escamilla, A. Zaidi, and M. Wigger, “Some Results on the Vector Gaussian Hypothesis Testing Problem,” *IEEE Int. Symp. Inf. Theory - Proc.*, no. May, pp. 2421–2425, 2020.

[11] T. S. Han, *Information-Spectrum Methods in Information Theory*, Baifukan, Tokyo, 1998.

[12] ——, “Hypothesis testing with the general source,” *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2415–2427, 2000.

[13] K.-i. Iwata and J. Muramatsu, “An information-spectrum approach to rate-distortion function with side information,” *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 85, no. 6, pp. 1387–1395, 2002.

[14] Y. Kochman and L. Wang, “Improved random-binning exponent for distributed hypothesis testing,” *arXiv preprint arXiv:2306.14499*, 2023.

[15] J. Arribas, C. Fernandez-Prades, and P. Closas, “Antenna array based gnss signal acquisition for interference mitigation,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 49, no. 1, pp. 223–243, 2013.

Communication Over Entanglement-Breaking Channels With Unreliable Entanglement Assistance

Uzi Pereg
ECE Department & Helen Diller Quantum Center, Technion
uzipereg@technion.ac.il

***Abstract*—Entanglement assistance can improve communication rates significantly. Yet, its generation is susceptible to failure. The unreliable assistance model accounts for those challenges. Previous work provided an asymptotic formula that outlines the tradeoff between the unassisted and excess rates from entanglement assistance. We derive a full characterization for entanglement-breaking channels, and show that combining entanglement-assisted and unassisted coding is suboptimal. From a networking perspective, this finding is nontrivial and highlights a quantum behavior arising from superposition.**

I. Introduction

Quantum entanglement has the potential to revolutionize communication systems, as it could be used to transmit information at speeds far beyond what is possible classically [1, 2]. In optical communications, generating pre-shared entanglement between the transmitter and the receiver can be challenging due to photon absorption during transmission. Therefore, practical systems rely on a back channel to confirm successful entanglement generation [3]. However, this introduces delays and further degrades entanglement resources. The author, along with Deppe and Boche [4], proposed an alternative approach for communication with unreliable entanglement assistance. Our principle of operation provides reliability by design, by adapting the communication rate based on the availability of entanglement assistance, while eliminating the need for feedback, repetition, or distillation.

Suppose that Alice wishes to send two messages, at rates R and R'. She encodes both messages using her share of the entanglement resources, as she does not know whether Bob will have access to the entangled resources. Nevertheless, heralded entanglement generation guarantees that Bob knows whether the procedure was successful or not. Bob has two decoding procedures. If the entanglement assistance has failed to reach Bob's location, he performs a decoding operation to recover the first message alone. Hence, the communication system operates on a rate R. Whereas if Bob has entanglement assistance, he decodes both messages, hence the overall transmission rate is $R+R'$. In other words, R is a guaranteed rate, and R' is the excess rate of information that entanglement assistance provides.

The previous work [4] established an asymptotic regularized formula for the capacity region, i.e., the set of all rate pairs (R, R') that can be achieved with a vanishing probability of decoding error. The achievability scheme is inspired by the classical network technique of superposition coding (SPC). We refer to the quantum method as *quantum SPC*. The classical technique consists of layered codebooks, by which the codewords are divided into so-called cloud centers and satellites, representing the first and second layers, respectively. In analogy, quantum SPC uses conditional quantum operations that map quantum cloud centers to quantum satellite states. Decoding is performed in two stages. First, Bob recovers the cloud index, corresponding to the guaranteed information. If the entanglement assistance is absent, then Bob quits after the first step. Otherwise, if Bob has entanglement assistance, then he continues to decode the satellite. Until now, it has remained unclear whether quantum SPC is optimal.

Entanglement breaking is a fundamental property of a large class of quantum channels, mapping any entangled state to a separable state [5]. One example is the qubit depolarizing channel, which is entanglement breaking only when the depolarization parameter is $\geq 2/3$ [6]. From a Shannon-theoretic perspective, entanglement-breaking channels are much better understood, compared to general quantum channels [7, 8]. In particular, the unassisted capacity is characterized by the single-letter Holevo information [7]. While an entanglement-breaking channel cannot be used to generate entanglement, it may facilitate the transmission of classical messages, and entanglement assistance can increase the channel capacity for sending classical information substantially [2].

In this work, we establish full characterization of the capacity region with unreliable entanglement assistance for the class of entanglement-breaking channels. Our main contribution is thus a converse result that complements the previous achievability proof, and shows that quantum SPC is indeed optimal for the class of entanglement-breaking channels. The analysis relies on observations from another work by the author [9, Sec. III-D] along with the geometric properties of the rate region. To complete the characterization, we single-letterize our capacity formula and show that the auxiliary systems have bounded dimensions.

We also demonstrate our results for an entanglement-breaking depolarizing channel. We show that quantum SPC can outperform time division even in this simple point-to-point setting. This is surprising because SPC is typically useful in more complex network setups, and does not yield an advantage in point-to-point communication. For example, in a classical broadcast channel with degraded messages, where a transmitter communicates with two receivers, SPC is unnecessary when the receivers' outputs are identical, as the capacity region can be attained using a simpler approach of time division. That is, concatenating two single-user codes is optimal. In our context, the system can be regarded as a quantum broadcast channel with degraded messages where one receiver has entanglement assistance, and the other does not. Nevertheless, the output states of the receivers are identical (without violating the no-cloning theorem, as we consider two alternative scenarios, see (1)-(2) below). The expectation would be that time division, combining assisted and unassisted codes, achieves optimality. However, this expectation is proven false as quantum SPC can outperform time division, based on the combination of a superposition code with a superposition state.

The full version of this paper can be found in [10].

Illustrative Metaphor

Communication with unreliable entanglement assistance is not a mere combination of the entanglement-assisted and unassisted settings. The protocol poses a challenge as Alice must encode without knowledge of the availability of assistance. The availability of entanglement is not associated with a probabilistic model either. To illustrate the concept of reliability, consider the following metaphor.

Imagine there are N travelers embarking on a journey aboard a ship that may have a variable number of lifeboats. The total capacity of the lifeboats is L, which determines how many travelers can be accommodated in case of a shipwreck, $L \leq N$. The ship's speed is denoted as $V \equiv V(N, L)$, while the lifeboats' speed is v_0. If the ship does not sink, each traveler will travel at speed V. To avoid a morbid narrative, let us envision that in the event of an unforeseen shipwreck, $(N - L)$ travelers will be safely rescued and brought back to the starting point, while the journey continues with the remaining travelers aboard the lifeboats. The speed of travel in this scenario is calculated as the average, $R = (L/N)v_0$.

In our metaphor, R represents the guaranteed speed for the remaining travelers, while $R' = V - R$ indicates the excess speed that the ship would have provided. Increasing the number of lifeboats improves the guaranteed speed but reduces the excess speed, while decreasing the number of lifeboats has the opposite effect. When planning for the worst-case scenario, it is crucial to consider both speeds, rather than just the average.

One may consider the option of dividing the travelers among a heavy ship and a light ship. Figuratively, our findings show that if the journey is subject to a quantum evolution, then we may outperform the division plan by allowing travelers to be in a quantum superposition state between the two ships.

II. Coding with Unreliable Assistance

A. Notation, Information Measures, and Quantum Channels

We use standard notation for quantum channels and information measures, as in [11, Chap. 11]: $X, Y, Z, \ldots$ are discrete random variables, on finite sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$, respectively. We use $x^n = (x_i)_{i\in[n]}$ to denote a sequence of letters from $\mathcal{X}$.

The state of a quantum system A is given by a density operator, ρ_A. The quantum mutual information is denoted by $I(A; B)_\rho = H(\rho_A) + H(\rho_B) - H(\rho_{AB})$, where $H(\rho) \equiv -\text{Tr}[\rho \log(\rho)]$ is the von Neumann entropy. The conditional quantum entropy is defined by $H(A|B)_\rho = H(\rho_{AB}) - H(\rho_B)$, and $I(A; B|C)_\rho$ is defined accordingly.

A quantum channel $\mathcal{N}_{A\to B}$ is a cptp map. If $A^n = (A_1, \ldots, A_n)$ is sent through n channel uses, then the input state undergoes the product map $\mathcal{N}^{\otimes n}_{A\to B}$. The channel is called entanglement breaking if for every input state $\rho_{AA'}$, where A' is an arbitrary reference system, the channel output $(\mathcal{N}_{A\to B} \otimes \mathbb{1})(\rho_{AA'})$ is separable. Every entanglement-breaking channel can be represented as a serial concatenation of a measurement channel followed by a classical-quantum channel [11, Corollary 4.6.1].

B. Coding and Channel Capacity

We define a code for communication with unreliable entanglement resources. Alice and Bob's entangled systems are denoted T_A and T_B, respectively.

Definition 1. A $(2^{nR}, 2^{nR'}, n)$ code with unreliable entanglement assistance consists of the following: Two message sets $[2^{nR}]$ and $[2^{nR'}]$, where 2^{nR}, $2^{nR'}$ are integers, an entangled state Ψ_{T_A,T_B}, a collection of encoding maps $\mathcal{F}^{m,m'}_{T_A\to A^n}$ for $m \in [2^{nR}]$ and $m' \in [2^{nR'}]$, and two decoding POVMs, $\mathcal{D}_{B^nT_B} = \{D_{m,m'}\}$ and $\mathcal{D}^*_{B^n} = \{D^*_m\}$.

Alice chooses two messages, $m \in [2^{nR}]$ and $m' \in [2^{nR'}]$. She applies the encoding map to her share of the entangled state, and then transmits A^n over n channel uses of $\mathcal{N}_{A\to B}$. Bob receives B^n. If the entanglement assistance is present, i.e., Bob has access to the resource T_B, then he should recover both messages. He performs a joint measurement $\mathcal{D}_{B^nT_B}$ to obtain an estimate $(\hat{m}, \hat{m}')$.

Otherwise, if entanglement assistance is absent, Bob does not have T_B. Hence, he performs the measurement $\mathcal{D}^*_{B^n}$ to obtain an estimate $\hat{\hat{m}}$ of the first message alone. The error probability is

$$P^{(n)}_{e|m,m'} = 1 - \text{Tr}\big[\mathcal{D} \circ \mathcal{N}^{\otimes n}_{A\to B} \circ \mathcal{F}^{m,m'}(\Psi_{T_A,T_B})\big] \tag{1}$$

in the presence of entanglement assistance, and

$$P^{*(n)}_{e|m,m'} = 1 - \text{Tr}\Big[\mathcal{D}^* \circ \mathcal{N}^{\otimes n}_{A\to B} \circ \mathcal{F}^{m,m'}(\Psi_{T_A})\Big] \tag{2}$$

without assistance. Note that the same channel $\mathcal{N}_{A\to B}$ appears in both. Furthermore, the encoded input is the same in both scenarios, since Alice does not know whether entanglement is available or not. Therefore, the error depends on (m, m') in both cases. A rate pair (R, R') is achievable if there exists a sequence of $(2^{nR}, 2^{nR'}, n)$ codes with unreliable entanglement assistance, such that $\max(P^{(n)}_{e|m,m'}, P^{*(n)}_{e|m,m'}) \to 0$ as $n \to \infty$. The capacity region $\mathcal{C}_{\text{EA}^*}(\mathcal{N})$ with unreliable entanglement assistance is defined as the set of achievable rate pairs.

III. Results

Let $\mathcal{N}_{A\to B}$ be an entanglement-breaking channel (see Section II-A). Define the region

$$\mathcal{R}_{\text{EA}^*}(\mathcal{N}) = \bigcup \left\{ \begin{array}{rl} (R, R') : R \leq & I(X; B)_\omega \\ R' \leq & I(G_2; B|X)_\omega \end{array} \right\} \tag{3}$$

where the union is over all auxiliary variables $X \sim p_X$, all quantum states $\varphi_{G_1 G_2}$, and all encoding channels $\mathcal{F}^{(x)}_{G_1 \to A}$,

$$\begin{aligned}\omega_{XAG_2} &= \sum_{x \in \mathcal{X}} p_X(x) \, |x\rangle\langle x| \otimes (\mathcal{F}^{(x)}_{G_1 \to A} \otimes \mathrm{id})(\varphi_{G_1 G_2}) \,, \\ \omega_{XBG_2} &= (\mathrm{id} \otimes \mathcal{N}_{A \to B} \otimes \mathrm{id})(\omega_{XAG_2}) \,. \qquad (4)\end{aligned}$$

Intuitively, X represents the guaranteed information, and G_1 is Alice's resource. Since the entangled resources G_1 and G_2 are pre-shared, the state is uncorrelated with the messages. Alice encodes the excess information using the encoding channel $\mathcal{F}^{(x)}$.

A. Capacity Theorem

Our main results are stated below, characterizing the capacity region for communication over entanglement-breaking channels with unreliable entanglement assistance. Previous work [4] established a regularized characterization for the capacity region, i.e., an asymptotic *multi-letter formula* of the form $\bigcup_{K=1}^{\infty} \frac{1}{K} \mathcal{R}(\mathcal{N}^{\otimes K})$. Here, we provide a complete characterization in the form of a single-letter formula.

Theorem 1. The capacity region of an entanglement-breaking quantum channel $\mathcal{N}_{A \to B}$ with unreliable entanglement assistance is given by

$$\mathcal{C}_{\mathrm{EA}*}(\mathcal{N}) = \mathcal{R}_{\mathrm{EA}*}(\mathcal{N}) \qquad (5)$$

where $\mathcal{R}_{\mathrm{EA}*}(\mathcal{N})$ is as defined in (3).

The proof is given in Section IV-B.

Remark 1. Single-letterization is highly valued in Shannon theory for reasons of computability, uniqueness, and insights on optimal coding [12]. However, the result in Theorem 1 in itself is not enough to claim that this is truly a single-letter characterization, as the computation of a rate region requires specified dimensions. Thereby, we show in Section III-C that the auxiliary systems, X, G_1, and G_2, all have bounded dimensions. Together, the results in Theorem 1 and Section III-C complete the characterization.

B. Equivalent characterization

Before we prove the capacity theorem, we establish useful properties of the region $\mathcal{R}_{\mathrm{EA}*}(\mathcal{N})$, as defined in (3). We show an equivalence to the region below:

$$\mathcal{O}_{\mathrm{EA}*}(\mathcal{N}) = \bigcup \left\{ \begin{array}{rl} (R, R') \, : \, R \leq & I(X;B)_\omega \\ R + R' \leq & I(XG_2;B)_\omega \end{array} \right\} \qquad (6)$$

where the union is as in (3). This will be useful in the proof for our main theorem. We begin with the convexity of our original region.

Lemma 2. The rate region $\mathcal{R}_{\mathrm{EA}*}(\mathcal{N})$ is a convex set.

Corollary 3. For every $\lambda \in [0,1]$,

$$\begin{aligned}&\mathcal{R}_{\mathrm{EA}*}(\mathcal{N}) \supseteq \\ &\left\{ \begin{array}{rl} (R, R') \, : \, R \leq & (1-\lambda) I(X;B)_\omega \\ R' \leq & I(G_2;B|X)_\omega + \lambda I(X;B)_\omega \end{array} \right\} . \qquad (7)\end{aligned}$$

The proof for the convexity properties in Lemma 2 and Corollary 3 is given in [10, App. A]. Next, we use those properties to establish equivalence.

Lemma 4 (Equivalence). $\mathcal{R}_{\mathrm{EA}*}(\mathcal{N}) = \mathcal{O}_{\mathrm{EA}*}(\mathcal{N})$.

Proof. The inclusion $\mathcal{R}_{\mathrm{EA}*}(\mathcal{N}) \subseteq \mathcal{O}_{\mathrm{EA}*}(\mathcal{N})$ is immediate by the chain rule. It remains to show that every rate pair in the region $\mathcal{O}_{\mathrm{EA}*}(\mathcal{N})$, belongs to $\mathcal{R}_{\mathrm{EA}*}(\mathcal{N})$ as well.

Let $(R, R') \in \mathcal{O}_{\mathrm{EA}*}(\mathcal{N})$, hence

$$R \leq I(X;B)_\omega \,, \; R + R' \leq I(XG_2;B)_\omega \,. \qquad (8)$$

By the first inequality, there exists $0 \leq \lambda \leq 1$ such that

$$R = (1-\lambda) I(X;B)_\omega \,. \qquad (9)$$

By (8)-(9),

$$\begin{aligned} R' &\leq I(XG_2;B)_\omega - R \\ &= I(XG_2;B)_\omega - I(X;B)_\omega + \lambda I(X;B)_\omega \\ &= I(G_2;B|X)_\omega + \lambda I(X;B)_\omega \,. \qquad (10)\end{aligned}$$

Hence, by Corollary 3, $(R, R') \in \mathcal{R}_{\mathrm{EA}*}(\mathcal{N})$. □

C. Single-letterization

We establish that our characterization is a single-letter formula (see Remark 1). Denote $d_A \equiv \dim(\mathcal{H}_A)$.

Lemma 5. The union in (3) is exhausted by pure states $|\phi_{G_1 G_2}\rangle$, cardinality $|\mathcal{X}| \leq d_A^2 + 1$, and dimensions $\dim(\mathcal{H}_{G_1}) = \dim(\mathcal{H}_{G_2}) \leq d_A(d_A^2+1)$.

The first part has already been stated in [4]. The quantum dimension bound is new, see proof is in Section IV-A below.

IV. Analysis

A. Single-Letterization

The first part of Lemma 5 has already been established in our previous work [4, Lemma 4], using convex analysis. Bounding the quantum dimensions is more challenging.

Consider a pure state, $|\psi_{G_1 G_2}\rangle$. Since the Schmidt rank is bounded by each dimension, we may assume w.l.o.g. that G_1 and G_2 are qudits of the same dimension d_0, for some $d_0 > 0$. We would like to show that the union can be restricted such that encoded state $\omega^x_{G_2 A} \equiv (\mathrm{id} \otimes \mathcal{F}^{(x)}_{G_1 A})(|\psi_{G_2 G_1}\rangle\langle\psi_{G_2 G_1}|)$ remains pure.

Since every quantum channel has a Stinespring dilation, there exists a unitary $V^{(x)}$ such that $\mathcal{F}^{(x)}_{G_1 \to A}(\rho) = \mathrm{Tr}_{DE}\left[V^{(x)}(|0\rangle\langle 0|_D \otimes \rho) V^{(x)\dagger}\right]$, where $V^{(x)}$ maps from $\mathcal{H}_D \otimes \mathcal{H}_{G_1}$ to $\mathcal{H}_E \otimes \mathcal{H}_A$, while D, E are reference systems with appropriate dimensions. Since G_1 is an arbitrary ancilla, we may include the reference D within this ancilla, and simplify as $\mathcal{F}^{(x)}_{G_1 \to A}(\rho) = \mathrm{Tr}_E\left[U^{(x)} \rho U^{(x)\dagger}\right]$, where $U^{(x)}$ is a unitary from $\mathcal{H}_{G_1}$ to $\mathcal{H}_E \otimes \mathcal{H}_A$.

We would like the ancilla G_2 to absorb the reference E as well. Seemingly, this would contradict (3) as E could be correlated with x. To resolve this difficulty, we show that the encoding operation can be reflected to G_2. Fix $x \in \mathcal{X}$ and consider the purification $\left|\omega^{(x)}_{G_2 E A}\right\rangle \equiv (\mathbb{1} \otimes U^{(x)}) \, |\psi_{G_2 G_1}\rangle$.

Let $W_{i,j}$ denote the Weyl operators on $\mathcal{H}_{G_1} \equiv \mathcal{H}_{G_2}$, for $i, j \in \{0, \ldots, d_0-1\}$ [11, Sec. 3.7.2]. By plugging a decomposition of $|\psi_{G_2G_1}\rangle$ in the generalized Bell basis [11, Ex. 3.7.11], and applying the mirror lemma, by which $(\mathbb{1}\otimes U)\,|\Phi\rangle = (U^T\otimes \mathbb{1})\,|\Phi\rangle$ for every qudit operator U [11, Ex. 3.7.12], we obtain $\left|\omega^{(x)}_{G_2EA}\right\rangle = \sum_{i,j=0}^{d_0-1} \alpha_{i,j}\left(W_{i,j}F^{(x)}_{G_1\to G_2E} \otimes \mathbb{1}_A\right)|\Phi\rangle_{G_1A}$, with $F^{(x)}_{G_1\to G_2E} = (U^{(x)})^T$. We see that (3) can thus be represented as a union over all unitaries $F^{(x)}_{G_1\to G_2E}\otimes \mathbb{1}_A$.

In this formulation, both E and G_2 are encoded by an operation depending on x. Thus, we can extend the union to $\bar{G}_2 = (G_2, E)$. The bound on the guaranteed rate R remains. As for the excess rate, $I(\bar{G}_2; B|X)_\omega \geq I(G_2; B|X)_\omega$. Hence, it suffices to consider pure states $\left|\omega^{(x)}_{G_2A}\right\rangle$, the Schmidt rank of which is bounded by d_A. Thus, the region is exhausted with $d_0 \leq |\mathcal{X}|d_A$. □

B. Capacity Proof

The direct part was proved in our earlier work [4]. We now focus our attention on the converse. Suppose that Alice and Bob share an unreliable resource $\Psi_{T_AT_B}$. Alice first prepares classical correlation,

$$\pi_{KMK'M'} \equiv \left(\frac{1}{2^{nR}}\sum_{m=1}^{2^{nR}} |m\rangle\langle m| \otimes |m\rangle\langle m|\right) \otimes \left(\frac{1}{2^{nR'}}\sum_{m'=1}^{2^{nR'}} |m'\rangle\langle m'| \otimes |m'\rangle\langle m'|\right) \quad (11)$$

locally. She encodes by $\mathcal{F}_{MM'T_A\to A^n}$, and transmits A^n. Bob receives B^n in the state $\omega_{KK'T_BB^n} \equiv (\mathrm{id}\otimes\mathcal{N}^{\otimes n}\mathcal{F})(\pi\otimes\Psi)$. He decodes with either $\mathcal{D}_{B^nT_B\to\hat{M}\hat{M}'}$ or $\mathcal{D}^*_{B^n\to\tilde{M}}$, depending on the availability of entanglement assistance.

Consider a sequence of codes $(\mathcal{F}_n, \Psi_n, \mathcal{D}_n, \mathcal{D}^*_n)$ with vanishing errors. By continuity and data processing arguments,

$$nR \leq I(K; B^n)_\omega + n\varepsilon^*_n\,, \quad (12)$$

$$n(R+R') \leq I(KK'T_B; B^n)_\omega + n\varepsilon_n \quad (13)$$

where $\varepsilon_n, \varepsilon^*_n \to 0$ as $n\to\infty$ [4, App. C].

Since the channel is entanglement-breaking, it can be represented by a measurement channel $\mathcal{M}_{A\to Y}$, followed by a preparation channel $\mathcal{P}_{Y\to B}$, where Y is classical [9, Sec. III-D]. Define the sequence of classical variables, $X_i \equiv (K, Y^{i-1})$, for $i\in[n]$. By the chain rule and the data processing inequality, (12)-(13) imply

$$\begin{aligned} n(R-\varepsilon^*_n) &\leq \sum_{i=1}^{n} I(KB^{i-1}; B_i)_\omega \\ &\leq \sum_{i=1}^{n} I(KY^{i-1}; B_i)_\omega \\ &= \sum_{i=1}^{n} I(X_i; B_i)_\omega\,, \end{aligned} \quad (14)$$

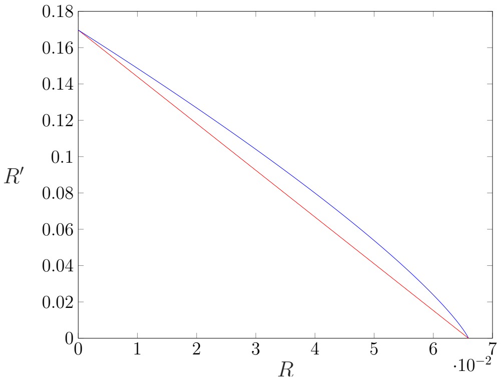


Fig. 1. Achievable rate regions.

and similarly,

$$n(R+R'-\varepsilon_n) \leq \sum_{i=1}^{n} I(K'T_BX_i; B_i)_\omega\,. \quad (15)$$

Letting J be uniformly distributed index in $[n]$, we have $R-\varepsilon^*_n \leq I(X_J; B_J|J)_\omega \leq I(JX_J; B_J)_\omega$ and $R+R'-\varepsilon_n \leq I(K'T_BJX_J; B_J)_\omega$ with respect to $\omega_{JK'T_BX_JB_J} \equiv \frac{1}{n}\sum_{i=1}^{n} |i\rangle\langle i|_J \otimes \omega_{K'T_BX_iB_i}$.

Taking $G_2 \equiv (K', T_B)$, $X \equiv (J, X_J)$, $A \equiv A_J$, hence $B\equiv B_J$, we deduce that $(R, R') \in \mathcal{O}_{EA*}(\mathcal{N})$. This, in turn, implies $(R, R') \in \mathcal{R}_{EA*}(\mathcal{N})$, by Lemma 4. □

V. Example

Consider the qubit depolarizing channel, $\mathcal{N}(\rho) = (1-\varepsilon)\rho + \varepsilon\frac{\mathbb{1}}{2}$, with $\varepsilon\in[0,1]$. The unassisted capacity, $C(\mathcal{N})$, is achieved with a symmetric distribution over $\{|0\rangle, |1\rangle\}$. On the other hand, the capacity with reliable entanglement assistance $C_{\mathrm{EA}}(\mathcal{N})$ is achieved with an EPR state (see [11]). A classical mixture of those strategies yields the time division region, $\mathcal{C}_{\mathrm{EA}^*}(\mathcal{N}) \supseteq \bigcup_{0\leq\lambda\leq1}\left\{\begin{array}{rl} (R,R') : R \leq & (1-\lambda)\,C(\mathcal{N}) \\ R' \leq & \lambda C_{\mathrm{EA}}(\mathcal{N}) \end{array}\right\}$. We claim that this is suboptimal.

Figure 1 depicts the capacity region for a parameter such that the channel is entanglement breaking, $\varepsilon = 0.7$ (as opposed to [4, Example 1]). The time-division bound is below the red line, whereas the blue curve indicates the capacity region that is achieved using a superposition state. Based on Theorem 1, we establish that the capacity region of an entanglement-breaking qubit depolarizing channel with unreliable entanglement assistance is given by

$$\mathcal{C}_{\mathrm{EA}^*}(\mathcal{N}) = \bigcup_{0\leq\alpha\leq\frac{1}{2}}\left\{\begin{array}{rl} (R,R') : R & \leq 1-h_2\left(\alpha*\frac{\varepsilon}{2}\right) \\ R' & \leq h_2(\alpha)+h_2\left(\alpha*\frac{\varepsilon}{2}\right) \\ & -H\left(\frac{\alpha\varepsilon}{2}, \frac{(1-\alpha)\varepsilon}{2}, \beta_+, \beta_-\right) \end{array}\right\} \quad (16)$$

with $\beta_\pm \equiv \frac{1}{2}-\frac{\varepsilon}{4}\pm\sqrt{\frac{\varepsilon^2}{16}-(1-\alpha)\alpha\varepsilon(1-\frac{3\varepsilon}{4})+\frac{1-\varepsilon}{4}}$, where $H(\mathbf{p}) \equiv -\sum_i p_i\log(p_i)$ is the Shannon entropy, the binary entropy function is $h_2(x) \equiv H(x, 1-x)$ for $x \in [0,1]$, and $\alpha * \beta = (1-\alpha)\beta + \alpha(1-\beta)$.

Proof. By Theorem 1, it suffices to evaluate the region $\mathcal{R}_{\text{EA*}}(\mathcal{N})$, as defined in (3).

We begin with the converse part and show that the set on the right-hand side of (16) is an outer bound on $\mathcal{R}_{\text{EA*}}(\mathcal{N})$. Consider a rate pair $(R, R') \in \mathcal{R}_{\text{EA*}}(\mathcal{N})$. Hence, $R \le I(X;B)_\omega$ and $R' \le I(G_2;B|X)_\omega$, or, equivalently,

$$R \le H(B)_\omega - H(B|X)_\omega\,, \tag{17a}$$

$$R' \le H(G_2|X)_\omega + H(B|X)_\omega - H(G_2B|X)_\omega\,, \tag{17b}$$

for some pure input state $|\phi_{G_1G_2}\rangle$, variable $X \sim p_X$, and encoder $\mathcal{F}^{(x)}_{G_1\to A}$ (see Lemma 5).

Based on the analysis in Section IV-A, it suffices to consider an encoder that produces a pure state $\left|\omega^{(x)}_{G_2A}\right\rangle$, for $x \in \mathcal{X}$. Consider a Schmidt decomposition,

$$\left|\omega^{(x)}_{G_2A}\right\rangle = \sqrt{1-\alpha_x}\,|\theta_{0x}\rangle \otimes |\psi_{0x}\rangle + \sqrt{\alpha_x}\,|\theta_{1x}\rangle \otimes |\psi_{1x}\rangle$$

with $\alpha_x \in [0,1]$. Since the encoding channel is applied to G_1 alone, the reduced state of G_2 remains unchanged. Thereby, the eigenvalues $(1-\alpha_x, \alpha_x)$ must be independent of x. That is, $\alpha_x \equiv \alpha$ for $x \in \mathcal{X}$, hence

$$H(G_2|X)_\omega = h_2(\alpha)\,. \tag{18}$$

Furthermore, the depolarizing channel is unitarily covariant, i.e., $\mathcal{N}(U\rho U^\dagger) = U\mathcal{N}(\rho)U^\dagger$ for every unitary U on $\mathcal{H}_A$. Thus,

$$H(B|X)_\omega = H(\mathcal{N}(\widetilde{\phi}_A)) = h_2\left(\alpha * \frac{\varepsilon}{2}\right) \tag{19}$$

where $\left|\widetilde{\phi}_{G_2A}\right\rangle = (1-\alpha)\,|00\rangle + \alpha\,|11\rangle$, and similarly,

$$H(G_2B|X)_\omega = H\left((\text{id} \otimes \mathcal{N})(\widetilde{\phi}_{G_2A})\right) \tag{20}$$

$$= H\left(\frac{\alpha\varepsilon}{2}, \frac{(1-\alpha)\varepsilon}{2}, \beta_\pm\right) \tag{21}$$

(see [13]). As $H(B)_\omega \le 1$, the converse follows.

Achievability follows as in [4, Example 1]. Instead of a classical mixture, we now use quantum superposition. Set $|\phi_{G_1G_2}\rangle \equiv \sqrt{1-\alpha}\,|00\rangle + \sqrt{\alpha}\,|11\rangle$, $p_X = \left(\frac{1}{2}, \frac{1}{2}\right)$, $\mathcal{F}^{(x)}(\rho) \equiv \mathsf{X}^x \rho \mathsf{X}^x$, where X is the bitflip Pauli operator. Thus, $\alpha = 0$ and $\alpha = \frac{1}{2}$ achieve the unassisted capacity and entanglement-assisted capacity, respectively. This results in (16). □

VI. Summary

We address communication over an entanglement-breaking quantum channel, given *unreliable* entanglement assistance. Previous work established a multi-letter formula and presented the quantum SPC achievable region [4]. Here, we show that the region is optimal for entanglement-breaking channels, and we single-letterize the formula, providing a complete characterization of the capacity region. Furthermore, we derive a closed-form expression for the qubit depolarizing channel, with a parameter $\varepsilon \ge \frac{2}{3}$. The capacity region is strictly larger than the time-division rate region. From a networking perspective, this finding is nontrivial and highlights a quantum behavior arising from superposition.

Acknowledgments

The author thanks Christoph Becher (Universität des Saarlandes) for his observations on the practical implementation of heralded entanglement, as well as Quntao Zhuang (University of Southern California) and Saikat Guha (University of Arizona) for useful discussions. This work was supported by the Israel VATAT Junior Faculty Program for Quantum Science and Technology through Grant 86636903, the Chaya Career Advancement Chair, Grant 8776026, the German-Israeli Project Cooperation (DIP), Grant 2032991, and the Helen Diller Quantum Center at the Technion.

References

[1] C. Wang and A. Rahman, "Quantum-enabled 6G wireless networks: Opportunities and challenges," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 58–69, 2022.

[2] S. Hao, H. Shi, W. Li, J. H. Shapiro, Q. Zhuang, and Z. Zhang, "Entanglement-assisted communication surpassing the ultimate classical capacity," *Phys. Rev. Lett.*, vol. 126, no. 25, p. 250501, 2021.

[3] E. Shchukin, F. Schmidt, and P. van Loock, "Waiting time in quantum repeaters with probabilistic entanglement swapping," *Phys. Rev. A*, vol. 100, no. 3, p. 032322, 2019.

[4] U. Pereg, C. Deppe, and H. Boche, "Communication with unreliable entanglement assistance," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4579–4599, 2023.

[5] M. Horodecki, P. W. Shor, and M. B. Ruskai, "Entanglement breaking channels," *Rev. Math. Phys.*, vol. 15, no. 06, pp. 629–641, 2003.

[6] L. Moravčíková and M. Ziman, "Entanglement-annihilating and entanglement-breaking channels," *J. Phys. A: Math. Theor.*, vol. 43, no. 27, p. 275306, 2010.

[7] P. W. Shor, "Additivity of the classical capacity of entanglement-breaking quantum channels," *J. Math. Phys.*, vol. 43, no. 9, pp. 4334–4340, May 2002.

[8] M. M. Wilde, A. Winter, and D. Yang, "Strong converse for the classical capacity of entanglement-breaking and hadamard channels via a sandwiched Rényi relative entropy," *Commun. Math. Phys.*, vol. 331, pp. 593–622, 2014.

[9] U. Pereg, "Communication over quantum channels with parameter estimation," *IEEE Trans. Inf. Theory*, vol. 68, no. 1, pp. 359–383, 2022.

[10] ——, "Communication over entanglement-breaking channels with unreliable entanglement assistance," *Phys. Rev. A*, vol. 108, p. 042616, Oct 2023. [Online]. Available: https://arxiv.org/pdf/2305.17692.pdf

[11] M. M. Wilde, *Quantum information theory*, 2nd ed. Cambridge University Press, 2017.

[12] J. Körner, "The concept of single-letterization in information theory," in *Open Prob. Commun. Comp.* Springer, 1987, pp. 35–36.

[13] D. Leung and J. Watrous, "On the complementary quantum capacity of the depolarizing channel," *Quantum*, vol. 1, p. 28, 2017.

Semantic Communications with Privacy Constraints

Amirreza Zamani[†], Sajad Daei[†], Tobias J. Oechtering[†], Deniz Gündüz[‡], Mikael Skoglund[†]

[†]Division of Information Science and Engineering, KTH Royal Institute of Technology
[‡]Dept. of Electrical and Electronic Engineering, Imperial College London
Email: amizam@kth.se, sajado@kth.se, oech@kth.se, d.gunduz@imperial.ac.uk, skoglund@kth.se

***Abstract*—We study a semantic communication problem with privacy constraints where an encoder has access to information source $X = (X_1, \ldots, X_N)$ which is arbitrarily correlated with private data S. A user asks for a task $h(X)$ and the encoder designs the semantic of the information source $f(X)$ to disclose. Here, $h(X)$ denotes the goal or task of the receiver and $f(X)$ corresponds to the semantic. Due to the privacy constraints $f(X)$ can not be revealed directly and the encoder applies a statistical privacy mechanism to produce disclosed data U. The goal is to design U based on $(f(X), h(X), S)$ that maximizes the revealed information about the task $h(X)$ while satisfying a privacy criterion.**

In this work, we propose a novel approach where U is produced by adding artificial noise M to the semantic $f(X)$. We design M utilizing different methods such as using extended versions of the Functional Representation Lemma, Strong Functional Representation Lemma, and separation technique. Lower and upper bounds on privacy-utility trade-off are derived and we study the obtained bounds in different scenarios to evaluate them.

I. Introduction

In this paper, random variable (RV) $X = (X_1, \ldots, X_N)$ denotes the information source and is correlated with the private data denoted by RV S. As shown in Fig. 1 a user asks an encoder about a task denoted by a function of X, i.e., $h(X)$. In this work $h(X)$ describes the task or goal of the communication. The encoder designs a message which is a function of X, i.e., $f(X)$, to disclose it. Here, $f(X)$ corresponds to the semantic of the information source which has less dimension compared to X. Since $f(X)$ can not be revealed directly (due to the privacy constraints) the encoder utilizes a privacy mechanism to produce disclosed data described by RV U. The goal is to design U based on the goal, semantic, and private data that reveals as much information as possible about $h(X)$ and satisfies a privacy criterion. We use mutual information to measure utility and privacy leakage. In this work, some bounded privacy leakage is allowed, i.e., $I(S;U) \leq \epsilon$.

The information theoretic privacy mechanism design and semantic communication problems are receiving increased attention recently. Related works can be found in [1]–[20]. Semantic communication involves transmitting a modified version of the original messages with reduced dimensionality to a receiver. The receiver's objective is to extract a specific goal or task, which is a lower-dimensional subset of the original message. Semantic communication takes into account not only the literal interpretation of the message but also the context, connotations, and nuances of language. It aims to avoid ambiguity and misunderstanding by considering how the message is perceived by the recipient and how it aligns with their knowledge, expectations, and cultural context [1]. Excluding the private data, relevant scenarios have been studied in [2] and [3].

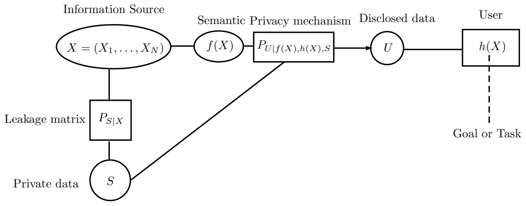


Fig. 1. Private semantic communication model. The goal is to design U such that it keeps as much information as possible about $h(X)$ while satisfying a privacy constraint.

In [4], a source coding problem with secrecy is studied. Privacy-utility trade-offs considering equivocation as measure of privacy and expected distortion as a measure of utility are studied in both [4] and [5]. In [6], the problem of privacy-utility trade-off considering mutual information both as measures of privacy and utility is studied. Under perfect privacy assumption, it has been shown that the privacy mechanism design problem can be reduced to a linear programming. In [7], we have designed privacy mechanisms with a per letter privacy criterion considering an invertible $P_{X|Y}$ where a small leakage is allowed. We generalized this result to a non-invertible leakage matrix in [8]. In [9], the problem of *secrecy by design* is studied where the results are derived under the perfect secrecy assumption. Bounds on secure decomposition have been derived using the Functional Representation Lemma and new bounds on privacy-utility trade-off are derived. In [10], the privacy problems considered in [9] are generalized by relaxing the perfect secrecy constraint and allowing some leakages. More specifically, we considered bounded mutual information, i.e., $I(U;X) \leq \epsilon$ for privacy leakage constraint. Moreover, the bounds obtained in [10] have been tightened in [20] by using *separation technique*.

In the present work, we utilize concepts from the privacy mechanism design outlined in [10] and [20] to introduce an innovative private semantic communication framework. The proposed scheme offers a mathematical approach to design a goal-oriented private utility function. This function not only facilitates the receiver in achieving the goal but also guarantees the privacy of the data from the recipient. For privacy mechanism design which corresponds to the acheivabilty we use different methods. To this end, extended versions of

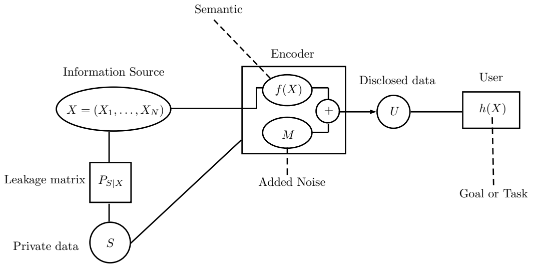


Fig. 2. Proposed approach for dealing with privacy concerns: adding artificial noise to the semantic $f(X)$.

the Functional Representation Lemma and the Strong Functional Representation Lemma and separation technique are used to address a *private semantic communication* problem. The Functional Representation Lemma (FRL) [9, Lemma 1] and the Strong Functional Representation Lemma (SFRL) [21, Theorem 1] are constructive lemmas that are valuable for the privacy design. Separation technique corresponds to representing a discrete random variable (RV) by two RVs which are correlated in general. We call this observation separation technique since it separates a RV into two RVs. To produce disclosed data U we propose an approach where artificial noise denoted by RV M is added to the semantic. We then use the privacy mechanism to design the artificial noise. In this work we assume that both semantic and goal are known by the encoder. We provide lower and upper bounds on the privacy-utility trade-off and study the bounds in special cases.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Let $P_{SX}=P_{SX_1\dots,X_N}$ denote the joint distribution of discrete random variables $X=(X_1,\dots,X_N)$ and S defined on finite alphabets $\mathcal{X}=\mathcal{X}_1\times\dots\mathcal{X}_N$ and $\mathcal{S}$. Here, X is vector with dimension N. We represent P_{SX} by a matrix defined on $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{X}|}$ and marginal distributions of S and X by vectors P_S and P_X defined on $\mathbb{R}^{|\mathcal{S}|}$ and $\mathbb{R}^{|\mathcal{X}|}$ given by the row and column sums of P_{SX}. We represent the leakage matrix $P_{S|X}$ by a matrix defined on $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{X}|}$. In this work, semantic of the source is a function of X denoted by $f(X)$ with dimension T where $T\leq N$. Furthermore, the goal or task of communication is represented by other function of X, i.e., $h(X)$, with dimension K where $K\leq N$. In most cases K is significantly smaller than N since semantic is designed based on the goal. In general $f\neq h$, since $f(\cdot)$ needs to be designed in an efficient way and efficiency can be defined based on different parameters.

As shown in Fig. 2, one approach of dealing with the privacy issue is to add artificial noise denoted by a discrete RV $M\in\mathcal{M}$ to the goal at the encoder side and communicate the resulting signal. However, it can annihilate the performance of the system by decreasing the utility achieved by the user. Thus, following this approach leads to a privacy-utility trade-off problem. In this model, the leakage is measured by the mutual information between S and U. Furthermore, the utility achieved by the user is measured by the mutual information between U and $h(X)$. As we mentioned earlier we assume that both semantic and goal are known to the encoder and the task is to design the artificial noise that achieves the optimal trade-off.

The private semantic communication design problem can be sated as follows

$$h_\epsilon(P_{S,f(X),h(X)})=\sup_{P_{U|S,f(X),h(X)}:I(U;S)\leq\epsilon,} I(h(X);U), \tag{1}$$

where $U=f(X)+M$, $P_{S,f(X),h(X)}$ is the joint distribution of $(S,f(X),h(X))$, and $P_{U|S,f(X),h(X)}$ describes the conditional distribution. In the following we study the case where $0\leq\epsilon<I(S;h(X))$, otherwise the optimal solution of $h_\epsilon(P_{S,f(X),h(X)})$ is $H(h(X))$ achieved by $U=h(X)$, i.e., $M=f(X)-h(X)$.

Remark 1. In (1), the privacy mechanism design is based on $f(X)$, $h(X)$, and S which are accessible by the encoder. Hence, the optimization is over $P_{U|S,f(X),h(X)}$ instead of $P_{U|S,X}$. In other words, the encoder does not require access to the information source X.

Remark 2. *A scenario that motivates our model can be stated as follows. Assume that the information source is not accessible directly to the encoder. Moreover, there exists a third party which designs $f(X)$ based on the task or goal $h(X)$ and shares it with the encoder. Since $f(X)$ is correlated with S, it can not be revealed directly. Thus, the encoder needs to design a message U based on $f(X)$, $h(X)$, and S and disclose it.*

Example 1. *Let X represent an image and the goal is to convey a specific feature of the image. For instance let X be the MNIST data set and the feature be the number which is illustrated inside the image. In this case, instead of encoding the whole image we only focus on the desired feature which is a number between 0 and 9 and encode it. The purpose of this paper is to develop a framework that guarantees the achievement of this goal while protecting the privacy of the private element.*

III. MAIN RESULTS

In this part, we provide lower and upper bounds for the privacy problems defined in (1) considering different scenarios. We study the tightness of the bounds in special cases and compare them in examples. To do so, in Appendix A, we present a simple observation which we call "separation technique". In the following results let $\mathcal{K}_S$ be all possible representations of S using separation technique where $X=(S_1,S_2)$. In other words we have $\mathcal{K}_S=\{(S_1,S_2):S=(S_1,S_2),\ |\mathcal{S}_1|\geq 2,\ |\mathcal{S}_2|\geq 2\}$.

Before stating the next theorem we derive an important expression for $I(X;U)$. For any correlated random variables S, X, and U, We have

$$\begin{aligned} I(X;U) &= I(S,X;U)-I(S;U|X),\\ &= I(S;U)+I(X;U|S)-I(S;U|X),\\ &= I(S;U)+H(X|S)-H(X|U,S)-I(S;U|X). \end{aligned} \tag{2}$$

Next, we derive lower and upper bounds on $h_\epsilon(P_{S,f(X),h(X)})$. For deriving lower bounds we use EFRL [10, Lemma 3], ESFRL [10, Lemma 4], and separation technique. For simplicity the remaining results are derived under the assumption $f(\cdot):\mathcal{X}_1\times\dots\mathcal{X}_N\rightarrow\mathbb{R}$, i.e., $T=1$.

Theorem 1. *For any* $0 \leq \epsilon < I(S; h(X))$ *and joint distribution* $P_{S,f(X),h(X)}$*, we have*

$$\max_{i\in\{1,..,4\}}\{L_h^i(\epsilon)\} \leq h_\epsilon(P_{S,f(X),h(X)}) \leq H(h(X)|S) + \epsilon, \tag{3}$$

where

$$\begin{aligned} L_h^1(\epsilon) &= H(h(X)|S) - H(S|h(X)) + \epsilon \\ &= H(h(X)) - H(S) + \epsilon, \\ L_h^2(\epsilon) &= H(h(X)|S) - \alpha H(S|h(X)) + \epsilon \\ &\quad - (1-\alpha)\left(\log(I(S;h(X)) + 1) + 4\right), \\ L_h^3(\epsilon) &= H(h(X)|S) + \epsilon - (\log(I(S;h(X)) + 1) + 4) \\ &\quad - \min_{(S_1,S_2)\in\mathcal{K}_S}\{\alpha_2 H(S_2|h(X))\}, \\ L_h^4(\epsilon) &= H(h(X)|S) + \epsilon - (\log(I(S;h(X))+1)+4) \end{aligned}$$
$$- \min_{(S_1,S_2)\in\mathcal{K}_S}\{\alpha_2\left(H(S|h(X) - \log(I(S;h(X))+1)+4)\right)\},$$

with $\alpha = \frac{\epsilon}{H(S)}$ *and* $\alpha_2 = \frac{\epsilon}{H(S_2)}$ *for any representation* $S = (S_1, S_2)$*. The lower bound in* (3) *is tight if* $H(S|h(X)) = 0$*, i.e.,* S *is a deterministic function of* $h(X)$*. Furthermore, if the lower bound* $L_h^1(\epsilon)$ *is tight then we have* $H(S|h(X)) = 0$*.*

Proof. Using (2) we have

$$\begin{aligned} I(f(X)+M; h(X)) &= I(f(X)+M; S) + H(h(X)|S) \\ &\quad - I(f(X)+M; S|h(X)) \\ &\quad - H(h(X)|S, f(X)+M), \end{aligned} \tag{4}$$

which results in

$$I(f(X)+M; h(X)) \leq \epsilon + H(h(X)|S). \tag{5}$$

For deriving the lower bounds $L_h^1(\epsilon)$ and $L_h^2(\epsilon)$ we use EFRL and ESFRL using $S \leftarrow X$ and $h(X) \leftarrow Y$. Let $\bar{U}$ and $\tilde{U}$ be the output of the EFRL and ESFRL. Using the same arguments in [10, Theorem 2] we have

$$I(\bar{U}; h(X)) \geq L_h^1(\epsilon), \tag{6}$$

$$I(\tilde{U}; h(X)) \geq L_h^2(\epsilon), \tag{7}$$

$$I(\bar{U}; S) = I(\tilde{U}; S) = \epsilon. \tag{8}$$

To achieve $L_h^1(\epsilon)$ let $M = \bar{U} - f(X)$ and to attain $L_h^2(\epsilon)$ let $M = \tilde{U} - f(X)$. The main idea for constructing a RV U that satisfies EFRL or ESFRL constraints is to add a randomized response to the output of FRL or SFRL. The randomization introduced in [22] is taken over S. To derive $L_h^3(\epsilon)$ and $L_h^4(\epsilon)$, let (S_1, S_2) be a possible representation of S, i.e., $S = (S_1, S_2)$. The main idea to achieve $L_h^3(\epsilon)$ and $L_h^4(\epsilon)$ is to take randomization over S_2 instead of S. In other words, we add a randomized response which is based on S_2 instead of S. Considering $L_h^2(\epsilon)$ and $L_h^4(\epsilon)$, α corresponds to the probability of randomizing over S, however, α_2 corresponds to the probability of randomizing over S_2 for any representation $S = (S_1, S_2)$. Let $\bar{U}$ be found by SFRL with $S = (X_1, X_2) \leftarrow X$ and $h(X) \leftarrow Y$. We have

$$I(\bar{U}; S_1, S_2) = H(h(X)|\bar{U}, S_1, S_2) = 0,$$
$$I(S_1, S_2; \bar{U}|h(X)) \leq \log(I(S_1, S_2; h(X)) + 1) + 4.$$

Moreover, let $U = (\bar{U}, W)$ with $W = \begin{cases} S_2, & \text{w.p. } \alpha_2 \\ c, & \text{w.p. } 1-\alpha_2 \end{cases}$, where c is a constant which does not belong to $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{X}$ and $\alpha_2 = \frac{\epsilon}{H(S_2)}$. First we show that $I(U; S_1, S_2) = \epsilon$. We have

$$\begin{aligned} I(U; S_1, S_2) &= I(\bar{U}, W; S_1, S_2) \overset{(a)}{=} I(W; S_1, S_2) \\ &= H(S_1,S_2) - \alpha_2 H(S_1|S_2) - (1-\alpha_2)H(S_1,S_2) = \epsilon, \end{aligned}$$

where (a) follows since $\bar{U}$ is independent of (S_1, S_2, W). Next, we expand $I(U; S_1, S_2|h(X))$.

$$I(U; S_1, S_2|h(X)) \tag{9}$$
$$= I(\bar{U}; S_1, S_2|h(X)) + I(W; S_1, S_2|h(X), \bar{U})$$
$$= I(\bar{U}; S_1, S_2|h(X)) + H(S_1, S_2|h(X), \bar{U}) \tag{10}$$
$$- H(S_1, S_2|h(X), \bar{U}, W)$$
$$= I(\bar{U}; S_1, S_2|h(X)) + \alpha_2 H(S_1, S_2|h(X), \bar{U}) \tag{11}$$
$$- \alpha_2 H(S_1|h(X), \bar{U}, S_2)$$
$$= I(\bar{U}; S_1, S_2|h(X)) - \alpha_2 H(S_1|h(X), \bar{U}, S_2) \tag{12}$$
$$+ \alpha_2\left(H(S_1, S_2|h(X)) - I(\bar{U}; S_1, S_2|h(X))\right)$$
$$= (1-\alpha)I(\bar{U}; S_1,S_2|h(X)) + \alpha_2 H(S_1, S_2|h(X))$$
$$- \alpha_2 H(S_1|h(X), \bar{U}, S_2). \tag{13}$$

In the following we bound (13) in two ways. We have

$$(13) = (1-\alpha_2)I(\bar{U}; S_1,S_2|h(X)) + \alpha_2 H(S_2|h(X)) \tag{14}$$
$$+ \alpha I(S_1; \bar{U}|h(X), S_2)$$
$$= I(\bar{U}; S_1,S_2|h(X)) + \alpha_2 H(S_2|h(X)) \tag{15}$$
$$- \alpha_2 I(\bar{U}; S_2|h(X))$$
$$\overset{(a)}{\leq} \log(I(S_1, S_2; h(X)) + 1) + 4 + \alpha_2 H(S_2|h(X)). \tag{16}$$

Furthermore,

$$(13) \leq (1-\alpha_2)I(\bar{U}; S_1,S_2|h(X)) + \alpha_2 H(S_1,S_2|h(X))$$
$$\overset{(b)}{\leqq} (1-\alpha_2)\left(\log(I(S_1, S_2; h(X)) + 1) + 4\right)$$
$$+ \alpha_2 H(S_1,S_2|h(X)). \tag{17}$$

Inequalities (a) and (b) follow since $\bar{U}$ is produced by SFRL, so that $I(\bar{U}; S_1, S_2|h(X)) \leq \log(I(S_1, S_2; h(X)) + 1) + 4$. Using (16), (17) and key equation in (2) we have

$$\begin{aligned} h_\epsilon(P_{XY}) &\geq I(U; h(X)) \\ &\overset{(c)}{\geq} \epsilon + H(h(X)|S_1, S_2) - \alpha_2 H(S_2|h(X)) \\ &\quad - (\log(I(S_1, S_2; h(X)) + 1) + 4) \\ &= \epsilon + H(h(X)|S) - \alpha_2 H(S_2|h(X)) \\ &\quad - \left(\log(I(S_1, S_2; h(X)) + 1) + 4\right), \end{aligned} \tag{18}$$

and

$$\begin{aligned} h_\epsilon(P_{XY}) &\geq I(U; h(X)) \\ &\overset{(d)}{\geq} \epsilon + H(h(X)|S_1, S_2) - \alpha_2 H(S_1, S_2|h(X)) \\ &\quad - (1-\alpha_2)(\log(I(S_1, S_2; h(X)) + 1) + 4) \\ &= \epsilon + H(h(X)|S) - \alpha_2 H(S|h(X)) \\ &\quad - (1-\alpha_2)(\log(I(S; h(X)) + 1) + 4). \end{aligned} \tag{19}$$

In steps (c) and (d) we used $H(h(X)|S_1, S_2, U) = 0$. The latter follows by definition of W and the fact that $\bar{U}$ is produced by SFRL. Note that since both (18) and (19) hold for any representation of X we can take maximum over all possible representations and we obtain

$$\begin{aligned}
&h_\epsilon(P_{S,f(X),h(X)})\\
&\geq H(h(X)|S)+\epsilon-(\log(I(S;h(X))+1)+4)\\
&-\min_{(S_1,S_2)\in\mathcal{K}_S}\{\alpha_2H(S_2|Y)\}=L_h^3(\epsilon),\\
&h_\epsilon(P_{S,f(X),h(X)})\\
&\geq H(h(X)|S)+\epsilon-(\log(I(S;h(X))+1)+4)\\
&-\min_{(S_1,S_2)\in\mathcal{K}_S}\{\alpha_2\left(H(S|h(X)-\log(I(S;h(X))+1)+4)\right)\},\\
&=L_h^4(\epsilon).
\end{aligned}$$

To design the artificial noise M, let RVs U_1 and U_2 achieve $L_h^3(\epsilon)$ and $L_h^4(\epsilon)$. Then, the privacy mechanism design that achieves $L_h^3(\epsilon)$ and $L_h^4(\epsilon)$ are obtained by $M_1=U_1-f(X)$ and $M_2=U_2-f(X)$. Finally, the results about tightness can be proved by using [10, Theorem 2]. □

Example 2. *Let $X=(\bar{X}_1,\bar{X}_2,\bar{X}_3)$, $h(X)=(\bar{X}_1,\bar{X}_2)$, and $S=\bar{X}_1+\bar{X}_2$, where $\bar{X}_1$, $\bar{X}_2$, and $\bar{X}_3$ are arbitrary correlated. In this case since S is a deterministic function of $h(X)$, by using Theorem 3 the lower bound $L_h^1(\epsilon)$ is tight and we have*

$$\begin{aligned}
h_\epsilon(P_{S,f(X),h(X)})&=H(h(X)|S)+\epsilon &(20)\\
&=H(\bar{X}_1,\bar{X}_2|\bar{X}_1+\bar{X}_2)+\epsilon. &(21)
\end{aligned}$$

Next, we show that tightness of the upper bound in Theorem 1 can be improved using the concept of *common information*. In other words, instead of having $H(S|h(X))=0$ we propose larger set of distributions that the upper bound is attained. To do so let us recall the definition of the common information between X and Y using [23]. For any pair of RVs (X,Y) defined on discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$, the common information between X and Y can be defined as follows

$$C(X;Y)=\inf_{P_{W|XY}:X-W-Y}I(X,Y;W). \qquad (22)$$

As shown in [23, Remark A] we have

$$I(X;Y)\leq C(X;Y)\leq\min\{H(X),H(Y)\}. \qquad (23)$$

One simple observation is that when $H(X|Y)=0$ or $H(Y|X)=0$ we have $I(X;Y)=C(X;Y)$. This follows since when $H(X|Y)=0$ we have

$$I(X,Y;W)=I(Y;W)$$

and W can be chosen as X, hence $X-W-Y$ holds. However, these are not the only cases where we have $I(X;Y)=C(X;Y)$ [24].

Corollary 1. *For any $0\leq\epsilon<I(S;h(X))$, if the common information and mutual information between the private data S and goal $h(X)$ are equal then we have*

$$h_\epsilon(P_{S,f(X),h(X)})=H(h(X)|S)+\epsilon.$$

Proof. The proof is based on [25, Theorem 3] and [25, Proposition 6]. □

Corollary 1 improves the condition $H(S|h(X))=0$ for achieving the upper bound in Theorem 1. Consequently, when the goal $h(X)$ is a deterministic function of S i.e., $H(h(X)|S)=0$, we have

$$h_\epsilon(P_{S,f(X),h(X)})=\epsilon.$$

Example 3. *Let $X=(\bar{X}_1,\bar{X}_2,\bar{X}_3)$, $h(X)=\bar{X}_1+\bar{X}_2$, and $S=(\bar{X}_1,\bar{X}_2,S_1)$, where $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, and S_1 are arbitrary correlated. In this case since $h(X)$ is a deterministic function of S, by using Corollary 1 we have*

$$h_\epsilon(P_{S,f(X),h(X)})=\epsilon. \qquad (24)$$

Moreover, considering perfect privacy constraint i.e., $\epsilon=0$, non-zero utilities can not be attained.

Special case: $H(h(X)|S)=0$

As we discussed earlier when $H(h(X)|S)=0$ we have $h_\epsilon(P_{S,f(X),h(X)})=\epsilon$. In this part, we propose a RV U that attains the upper bound ϵ.To achieve ϵ let $U=\begin{cases}h(X), & \text{w.p. } \alpha\\ c, & \text{w.p. } 1-\alpha\end{cases}$, where c is a constant which does not belong to the support of S and $\alpha=\frac{\epsilon}{I(S;h(X))}$. We emphasize that since we only consider the range $\epsilon<I(S;h(X))$, we have $\alpha<1$. To verify the privacy constraint we have

$$\begin{aligned}
I(U;S)&=H(S)-H(S|U)\\
&=H(X)-\alpha H(S|h(X))-(1-\alpha)H(S)\\
&=\alpha I(S;h(X))=\epsilon.
\end{aligned}$$

Using (2) we have

$$\begin{aligned}
I(U;h(X))&\overset{(a)}{=}\epsilon-H(S|h(X))+H(S|h(X),U)\\
&=\epsilon-H(S|h(X))+H(S|h(X))=\epsilon,
\end{aligned}$$

where in (a) we use $I(U;S)=\epsilon$ and $H(h(X)|S)=H(h(X)|S,U)=0$.

Comparison

In this part we study the bounds considering different cases. For simplicity let $X=(X_1,X_2)$ where X_1 and X_2 are arbitrary correlated. In this case we have

$$\begin{aligned}
&U_1^\epsilon=H(h(X)|S_1,S_2)-\epsilon,\\
&L_1^\epsilon=H(h(X)|S_1,S_2)-H(S_1,S_2|h(X))+\epsilon,\\
&L_2^\epsilon=H(h(X)|S_1,S_2)-\alpha H(S_1,S_2|Y)+\epsilon\\
&\quad-(1-\alpha)\left(\log(I(S_1,S_2;h(X))+1)+4\right),\\
&\bar{L}_3^\epsilon\triangleq H(h(X)|S_1,S_2)+\epsilon-(\log(I(S_1,S_2;h(X))+1)+4)\\
&\quad-\alpha_2H(S_2|Y)\leq L_3^\epsilon,\\
&\bar{L}_4^\epsilon\triangleq H(h(X)|S_1,S_2)+\epsilon+\alpha_2H(S_1,S_2|h(X))\\
&-(1-\alpha_2)(\log(I(S_1,S_2;h(X))+1)+4)\leq L_4^\epsilon,
\end{aligned}$$

where $\alpha=\frac{\epsilon}{H(S)}$ and $\alpha_2=\frac{\epsilon}{H(S_2)}$. Note that the lower bounds $\bar{L}_3^\epsilon$ and $\bar{L}_4^\epsilon$ are obtained based on the initial representation of $S=(S_1,S_2)$. Therefore, we have $\bar{L}_3^\epsilon\leq L_3^\epsilon$ and $\bar{L}_4^\epsilon\leq L_4^\epsilon$. Here, we compare the lower bounds L_2^ϵ, $\bar{L}_3^\epsilon$, and $\bar{L}_4^\epsilon$. To do so we consider the following scenarios.

Scenario 1: To compare $\bar{L}_4^\epsilon$ with L_2^ϵ, let us assume that $H(S_1,S_2|h(X))\leq\log(I(S_1,S_2;h(X))+1)+4$. A simple example can be considering S_1 and S_2 as binary RVs. In this case we have

$$\bar{L}_4^\epsilon - L_2^\epsilon = \epsilon(\frac{1}{H(S_2)} - \frac{1}{H(S_1,S_2)})\times$$
$$(\log(I(S_1,S_2;h(X))+1)+4-H(S_1,S_2|h(X))) \geq 0.$$

Scenario 2: To compare $\bar{L}_3^\epsilon$ with L_2^ϵ, let us assume that S_2 is a deterministic function of $h(X)$ and $H(S_1|h(X)) \geq \log(I(S_1,S_2;h(X))+1)+4$. A simple example is to let $4+H(h(X)) \leq H(S_1|h(X))$ which leads to $H(S_1|h(X)) \geq \log(I(S_1,S_2;h(X))+1)+4$. In this case we have

$$\bar{L}_3^\epsilon - L_2^\epsilon =$$
$$\frac{\epsilon}{H(S_1,S_2)}\left(H(S_1|h(X))-\log(I(S_1,S_2;h(X))+1)-4\right) \geq 0.$$

Moreover, we have

$$\begin{aligned}
&\bar{L}_3^\epsilon - \bar{L}_4^\epsilon\\
&= \alpha_2(H(S_1|h(X)) - \log(I(S_1;h(X)) + H(S_2|S_1)+1)-4)\\
&\overset{(a)}{\geq} \alpha_2\left(H(S_1|h(X)) - I(S_1;h(X)) - H(S_2|S_1) - 4\right)\\
&\overset{(b)}{\geq} \alpha_2\left(H(S_1|h(X)) - I(S_1;h(X)) - H(h(X)|S_1) - 4\right)\\
&= \alpha_2\left(H(S_1|h(X)) - H(h(X)) - 4\right)\\
&\geq 0, \qquad (25)
\end{aligned}$$

where (a) follows since $\log(1+x) \leq x$ and (b) holds since we have $H(S_2|S_1) \leq H(h(X)|S_1)$ and $H(S_2|h(X)) = 0$. Furthermore,

$$\begin{aligned}
&\bar{L}_3^\epsilon - L_1^\epsilon\\
&= H(S_1|h(X)) - \log(I(S_1;h(X)) + H(S_2|S_1)+1)-4\\
&\geq H(S_1|h(X)) - I(S_1;h(X)) - H(S_2|S_1) - 4\\
&\geq H(S_1|h(X)) - I(S_1;h(X)) - H(h(X)|S_1) - 4\\
&= H(S_1|h(X)) - H(h(X)) - 4 \geq 0. \qquad (26)
\end{aligned}$$

Finally, by using (25) and (26) we have

$$\bar{L}_3^\epsilon \geq \max\{L_2^\epsilon, \bar{L}_4^\epsilon, L_1^\epsilon\}.$$

IV. Conclusion

We have introduced a semantic communication with privacy constraint where it has been shown that using extended versions of the FRL, SFRL, and separation technique lower bounds on $h_\epsilon(P_{S,f(X),h(X)})$ are obtained. When the private data S is a deterministic function of the goal $h(X)$, the upper bound is achieved. Also, this statement is generalized by using the concept of common information. Finally, we have studied the bounds considering different scenarios.

References

[1] J. Feist, *Significance in language: A theory of semantics*. Taylor & Francis, 2022.

[2] E. C. Strinati and S. Barbarossa, "6g networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[3] Y. Shao, Q. Cao, and D. Gündüz, "A theory of semantic communication," *arXiv preprint arXiv:2212.01485*, 2022.

[4] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.)," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 918–923, 1983.

[5] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[6] B. Rassouli and D. Gündüz, "On perfect privacy," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 177–191, 2021.

[7] A. Zamani, T. J. Oechtering, and M. Skoglund, "A design framework for strongly χ^2-private data disclosure," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2312–2325, 2021.

[8] A. Zamani, T. J. Oechtering, and M. Skoglund, "Data disclosure with non-zero leakage and non-invertible leakage matrix," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 165–179, 2022.

[9] Y. Y. Shkel, R. S. Blum, and H. V. Poor, "Secrecy by design with applications to privacy and compression," *IEEE Transactions on Information Theory*, vol. 67, no. 2, pp. 824–843, 2021.

[10] A. Zamani, T. J. Oechtering, and M. Skoglund, "Bounds for privacy-utility trade-off with non-zero leakage," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 620–625.

[11] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Information Theory Workshop*, 2014, pp. 501–505.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[13] F. P. Calmon, A. Makhdoumi, M. Medard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5011–5038, Aug 2017.

[14] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, 2020.

[15] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1512–1534, 2019.

[16] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 594–603, 2020.

[17] B. Rassouli, F. E. Rosas, and D. Gündüz, "Data disclosure under perfect sample privacy," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2019.

[18] I. Issa, S. Kamath, and A. B. Wagner, "Maximal leakage minimization for the shannon cipher system," in *2016 IEEE International Symposium on Information Theory*, 2016, pp. 520–524.

[19] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, 2016. [Online]. Available: https://www.mdpi.com/2078-2489/7/1/15

[20] A. Zamani, T. J. Oechtering, and M. Skoglund, "New privacy mechanism design with direct access to the private data," *arXiv preprint arXiv:2309.09033*, 2023.

[21] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, 2018.

[22] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[23] A. D. Wyner, "The wire-tap channel," *The Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, 1975.

[24] R. Ahlswede and J. Körner, *On common information and related characteristics of correlated information sources*. Springer, 2006.

[25] A. Zamani, T. J. Oechtering, and M. Skoglund, "On the privacy-utility trade-off with and without direct access to the private data," *IEEE Transactions on Information Theory*, pp. 1–1, 2023.

Appendix A

Observation. *(Separation technique) Any discrete RV S supported on $\mathcal{S} = \{1,\ldots,|\mathcal{S}|\}$ can be represented by two RVs (S_1,S_2).*

Proof. First, let $|\mathcal{S}|$ be not a prime number. Thus, there exist $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ such that $|\mathcal{S}| = |\mathcal{S}_1| \times |\mathcal{S}_2|$ where $|\mathcal{S}_1| \geq |\mathcal{S}_2| \geq 2$. We can uniquely map each $x \in \mathcal{S}$ into a pair (s_1,s_2) where $s_1 \in \mathcal{S}_1$ and $s_2 \in \mathcal{S}_2$. As a result, we can represent S by the pair (S_1,S_2) where $\mathcal{S}_1 = \{1,\ldots,|\mathcal{S}_1|\}$, $\mathcal{S}_2 = \{1,\ldots,|\mathcal{S}_2|\}$, and $P_S(s) = P_{S_1S_2}(s_1,s_2)$. Next, let $|\mathcal{S}|$ be a prime number. Hence, there exist $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ such that $|\mathcal{S}|+1 = |\mathcal{S}_1| \times |\mathcal{S}_2|$ and we can represent S by the pair (S_1,S_2) where $P_{S_1S_2}(s_1 = |\mathcal{S}_1|, s_2 = |\mathcal{S}_2|) = 0$. In other words, the last pair $(|\mathcal{S}_1|,|\mathcal{S}_2|)$ is not mapped to any $s \in \mathcal{S}$. □

Multi-terminal Strong Coordination over Noiseless Networks with Secrecy Constraints

Viswanathan Ramachandran, Tobias J. Oechtering and Mikael Skoglund
Department of Information Science and Engineering
KTH Royal Institute of Technology, Stockholm, Sweden
{visra,oech,skoglund}@kth.se

***Abstract*—We investigate the problem of securely emulating a two-user multiple-access channel (MAC) aided by a multiple-access network of noiseless links as a resource. In this configuration, two encoders observe independent and identically distributed (i.i.d.) samples of a source random variable each and send rate-limited messages over their respective pairwise (noiseless) communication links to the decoder. The decoder also receives i.i.d. samples of a side-information random variable. At limited rates, each encoder and the decoder additionally possess independent pairwise shared randomness. The objective is for the decoder to generate approximately i.i.d. samples of another random variable, which is jointly distributed with the two sources and the side information. Furthermore, we require that an external eavesdropper who intercepts the communication links and has correlated observations but no access to the shared randomness variables, learns virtually nothing about the input sources and the simulated output sequence. We are interested in the rate tuples which permit such simulation with strong secrecy. We establish a complete characterization for this secure multi-terminal coordination problem when the sources are independent and one of the pairwise shared randomness rates is unconstrained. Moreover, we derive an achievable region and an outer bound for the general case of correlated sources and limited shared randomness rates.**

I. Introduction

Cuff *et al.* [1] introduced the framework of *coordination capacity*, which shifts the focus of network communication from traditional data transmission between nodes to establishing a desired joint distribution of actions among all nodes. In this framework, the nodes communicate with each other to coordinate their actions, making it particularly relevant in scenarios where distributed agents must achieve decentralized cooperation (wireless sensor networks, self-driving cars, etc. being a few examples). In [1], two criteria were introduced: *empirical coordination* where the joint type of the actions must approach a target distribution, and *strong coordination*, where the distribution of the sequence of actions must be close in total variation to a target distribution. Strong coordination is especially useful in adversarial settings, where the node actions must appear random to an external adversary overhearing the communication. This work falls within the realm of strong coordination, and examines it in a two-sender noiseless multiple-access network setting with an eavesdropper.

Strong coordination, also known as *channel simulation*, aims to describe the minimal communication required to achieve remote correlation. An encoder observing an independently and identically distributed (i.i.d.) source X^n with distribution q_X transmits a message to a decoder through a noiseless link in the point-to-point formulation. The decoder's task is to produce a sequence Y^n such that the total variation distance between the induced joint distribution of (X^n, Y^n) and the i.i.d. joint distribution obtained by transmitting the source X^n through a discrete memoryless channel $q_{Y|X}$ vanishes asymptotically with blocklength. Both the encoder and the decoder may benefit from common randomness to accomplish such channel simulation. The complete optimal trade-off region between communication and shared randomness rates was independently discovered by [2] and [3], where the available shared randomness reduces the communication rate requirements. Expanding on this characterization, [4] obtained a complete characterization for a point-to-point network involving interactive communications between the two nodes.

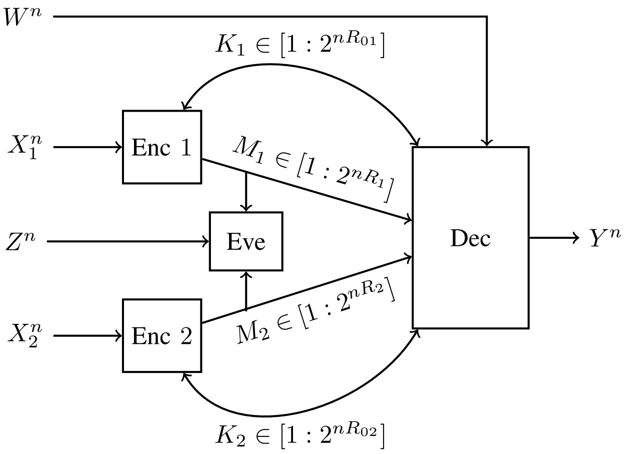


Fig. 1. Strong coordination over a two-sender multiple-access network of noiseless links subject to secrecy constraints.

Conclusive results on channel simulation in multi-terminal networks are comparatively rare – we outline a few of them. A cascade network as an extension of the point-to-point network with secrecy constraints was examined in [5], and the optimal trade-off between communication and common randomness rates was determined. A generalization of [3] to include side information at the receiver was addressed in [6], where some tight characterizations were obtained for specific cases. In [7], strong coordination over a multiple-access network consisting of noiseless links was explored, where a complete characterization was established for the case of independent sources. The role of shared randomness amongst the encoders in reducing the communication rate requirements for channel simulation was also investigated therein.

Channel simulation has also been investigated subject to secrecy constraints. For instance, an achievable region was obtained in [8] for a two-terminal noiseless setting with an external eavesdropper who taps into the communication between legitimate nodes and has access to correlated observations. Secure randomized function computation between two mutually distrusting users (with no external adversaries) was addressed in [9] and the class of randomized functions which can be computed with perfect security was characterized.

In this paper, we address secure multiple access channel simulation over a multiple-access network of noiseless links. This is a three-terminal version of [3], where we require that correlated sources be encoded in a distributed manner so as to achieve strong coordination with a decoder output, while also ensuring strong secrecy against an external eavesdropper. The current setting thus constitutes an extension of the multi-terminal noiseless network coordination problem addressed in [7] to account for insecure communication links and provide secrecy guarantees against the external eavesdropper (please refer to Remarks 1 and 2 in Section III that highlight the differences compared to this work). Compared to [8], our setting explores secure channel simulation in a multi-terminal network rather than a point-to-point network with an eavesdropper.

Main Contributions.

- We derive an achievable region (Theorem 1) and an outer bound (Theorem 2) for the general case. Our achievable scheme combines coordination coding in the spirit of [3] with encryption of the communication between the nodes using the shared randomness as secret keys (as a one-time pad [10]).
- For the case when the input sources and side information are mutually independent, and one of the shared randomness rates is unconstrained, we establish a complete characterization for this secure multi-terminal coordination problem (Theorem 3).

II. System Model

We investigate strong coordination of signals in a two-sender multiple-access network of noiseless links with secrecy constraints. The setup comprises two encoders, with encoder $j \in \{1,2\}$ observing an input given by X_j^n, and a decoder which observes a side information sequence W^n. For $j \in \{1,2\}$, encoder j and the decoder can harness pairwise shared randomness K_j, assumed to be uniformly distributed on $[1:2^{nR_{0j}}]$. Encoder $j \in \{1,2\}$ (which observes X_j^n and has access to K_j) transmits a rate-limited message $M_j \in [1:2^{nR_j}]$ over its respective noiseless communication link to the decoder. It is assumed that the message communication between the encoder-decoder pairs occurs over a public channel, where an eavesdropper (Eve) can tap into the messages (M_1, M_2) sent over the channel (Eve does not have access to the shared randomness (K_1, K_2)). In addition, Eve has access to correlated observations Z^n which is jointly distributed with the input sources and decoder side information. In particular, $(X_{1i}, X_{2i}, W_i, Z_i)$, $i = 1, 2, \ldots, n$, are assumed to be independent and identically distributed (i.i.d.) with joint distribution specified by nature as $q_{X_1X_2WZ}$. The random variables X_1, X_2, W, Z assume values in finite alphabets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{W}, \mathcal{Z}$, respectively. The shared randomness indices K_1 and K_2 are assumed to be independent of each other and of (X_1^n, X_2^n, W^n, Z^n). The decoder obtains (M_1, M_2) along with (K_1, K_2, W^n) and simulates an output sequence Y^n (where Y_i, $i = 1, \ldots, n$, assumes values in a finite alphabet $\mathcal{Y}$) which along with the input sources, side information and eavesdropper observations must be approximately i.i.d. according to the joint distribution $q^{(n)}_{X_1X_2WZY}(x_1^n, x_2^n, w^n, z^n, y^n) := \prod_{i=1}^n q_{X_1X_2WZY}(x_{1i}, x_{2i}, w_i, z_i, y_i)$ (refer Figure 1). Moreover, we require strong secrecy against the eavesdropper in the sense that the messages (M_1, M_2) must appear to be independent of $(X_1^n, X_2^n, W^n, Z^n, Y^n)$.

Definition 1. *A* $(2^{nR_1}, 2^{nR_2}, 2^{nR_{01}}, 2^{nR_{02}}, n)$ code *comprises two randomized encoder maps* $p^{\text{Enc}_j}(m_j|x_j^n, k_j)$ *for* $j \in \{1,2\}$ *and a randomized decoder map* $p^{\text{Dec}}(y^n|m_1, m_2, k_1, k_2, w^n)$, *where the shared randomness and communication indices assume values* $k_j \in [1:2^{nR_{0j}}]$ *and* $m_j \in [1:2^{nR_j}]$ *respectively for* $j \in \{1,2\}$.

The induced joint distribution of all the random variables $(X_1^n, X_2^n, W^n, Z^n, M_1, M_2, K_1, K_2, Y^n)$, the resulting induced marginal distribution on $(X_1^n, X_2^n, W^n, Z^n, Y^n, M_1, M_2)$ and the induced marginal distribution on (M_1, M_2) are respectively given by

$$\begin{aligned} &p(x_1^n, x_2^n, w^n, z^n, m_1, m_2, k_1, k_2, y^n) \\ &\quad = \frac{1}{2^{n(R_{01}+R_{02})}} q(x_1^n, x_2^n, w^n, z^n) \prod_{j=1}^{2} p^{\text{Enc}_j}(m_j|x_j^n, k_j) \\ &\qquad \times p^{\text{Dec}}(y^n|m_1, m_2, k_1, k_2, w^n), \end{aligned}$$

and

$$\begin{aligned} &p^{\text{ind}}(x_1^n, x_2^n, w^n, z^n, y^n, m_1, m_2) \\ &\quad = \sum_{k_1,k_2} p(x_1^n, x_2^n, w^n, z^n, m_1, m_2, k_1, k_2, y^n), \\ &p^{\text{ind}}(m_1, m_2) \\ &\quad = \sum_{k_1,k_2,x_1^n,x_2^n,w^n,z^n,y^n} p(x_1^n, x_2^n, w^n, z^n, m_1, m_2, k_1, k_2, y^n). \end{aligned}$$

Definition 2. *A rate quadruple* $(R_1, R_2, R_{01}, R_{02})$ *is said to be* achievable for a target joint distribution $q_{X_1X_2WZY}$ with secrecy *provided there exists a sequence of* $(2^{nR_1}, 2^{nR_2}, 2^{nR_{01}}, 2^{nR_{02}}, n)$ *codes such that*

$$\lim_{n\to\infty} ||p^{\text{ind}}_{X_1^n,X_2^n,W^n,Z^n,Y^n,M_1,M_2} - p^{\text{ind}}_{M_1,M_2} q^{(n)}_{X_1X_2WZY}||_1 = 0, \tag{1}$$

where $q^{(n)}_{X_1X_2WZY}$ is the target i.i.d. product distribution defined as

$$\begin{aligned}&q^{(n)}_{X_1X_2WZY}(x_1^n, x_2^n, w^n, z^n, y^n)\\&\quad := \prod_{i=1}^{n} q_{X_1X_2WZY}(x_{1i}, x_{2i}, w_i, z_i, y_i).\end{aligned}$$

We notice that (1) implies the following strong secrecy criterion against the eavesdropper

$$\lim_{n\to\infty} I(M_1, M_2; X_1^n, X_2^n, W^n, Z^n, Y^n) = 0. \tag{2}$$

Moreover, (2) along with the approximate i.i.d. nature of $(X_1^n, X_2^n, W^n, Z^n, Y^n)$ from (1) can be alternatively expressed as

$$\begin{aligned}\lim_{n\to\infty} |I(M_1, M_2, Z^n; &X_1^n, X_2^n, W^n, Y^n)\\&- nI(Z; X_1, X_2, W, Y)| = 0. \end{aligned}\tag{3}$$

Definition 3. *The* rate region $\mathcal{R}^{\text{secrecy}}_{\text{coordination}}$ *is the closure of the set of all achievable rate quadruples* $(R_1, R_2, R_{01}, R_{02})$.

Let $\mathcal{R}^{\text{secrecy}}_{\text{coordination, } R_{02}\to\infty}$ be the rate region when the available pairwise shared randomness K_2 is unconstrained, i.e.,

$$\begin{aligned}&\mathcal{R}^{\text{secrecy}}_{\text{coordination, } R_{02}\to\infty} = \{(R_1, R_2, R_{01}) : \exists\ R_{02}\\&\text{s.t. } (R_1, R_2, R_{01}, R_{02}) \in \mathcal{R}^{\text{secrecy}}_{\text{coordination}}\}.\end{aligned}\tag{4}$$

III. Main Results

Firstly, let us present an inner bound to the rate region $\mathcal{R}^{\text{secrecy}}_{\text{coordination}}$.

Theorem 1 (Achievable Rate Region)**.** *Given a target p.m.f.* $q_{X_1X_2WZY}$*, the rate quadruple* $(R_1, R_2, R_{01}, R_{02})$ *is in* $\mathcal{R}^{\text{secrecy}}_{\text{coordination}}$ *provided*

$$\begin{aligned}R_1 &\ge I(U_1; X_1|U_2, W, T)\\R_2 &\ge I(U_2; X_2|U_1, W, T)\\R_1 + R_2 &\ge I(U_1, U_2; X_1, X_2|W, T)\\R_{01} &\ge I(U_1; X_1, X_2, Z, Y|W, T) - I(U_1; U_2|W, T)\\R_{02} &\ge I(U_2; X_1, X_2, Z, Y|W, T) - I(U_1; U_2|W, T)\\R_2 + R_{01} &\ge I(U_1; X_1, X_2, Z, Y|W, T)\\&\quad + I(U_2; X_2|U_1, W, T)\\R_1 + R_{02} &\ge I(U_2; X_1, X_2, Z, Y|W, T)\\&\quad + I(U_1; X_1|U_2, W, T)\\R_{01} + R_{02} &\ge I(U_1, U_2; X_1, X_2, Z, Y|W, T),\end{aligned}$$

for some p.m.f.

$$\begin{aligned}p(x_1,&x_2, w, z, t, u_1, u_2, y) =\\&p(x_1, x_2, w, z)p(t)\prod_{j=1}^{2} p(u_j|x_j, t)p(y|u_1, u_2, w, t)\end{aligned}\tag{5}$$

such that

$$\sum_{u_1,u_2} p(x_1, x_2, w, z, u_1, u_2, y|t) = q(x_1, x_2, w, z, y), \textit{ for all } t.$$

A detailed proof of Theorem 1 can be found in the long version [11]. We may think of (U_1, U_2) as quantization codebooks for the respective encoder source observations (X_1, X_2). Furthermore, to ensure secrecy, the compressed source descriptions are encrypted using the shared randomness variables as secret keys. In particular, the shared randomness K_j is used as a one-time pad on the message M_j for $j = 1, 2$.

Remark 1. Compared to the setting of [7] without secrecy constraints, the proof of Theorem 1 in the full version [11] differs considerably due to the stronger constraint (1) imposed on the joint distribution for ensuring strong secrecy. In particular, we show that there exists a sequence of $(2^{nR_1}, 2^{nR_2}, 2^{nR_{01}}, 2^{nR_{02}}, n)$ codes with encoder and decoder mappings along with the particular realization of random binning resulting in vanishing total variation distance as well as the required strong secrecy constraint.

We next derive an outer bound to the rate region $\mathcal{R}^{\text{secrecy}}_{\text{coordination}}$.

Theorem 2 (Outer Bound)**.** *Given a target p.m.f.* $q_{X_1X_2WZY}$*, any rate quadruple* $(R_1, R_2, R_{01}, R_{02})$ *in* $\mathcal{R}^{\text{secrecy}}_{\text{coordination}}$ *obeys, for every* $\epsilon \in (0, \frac{1}{4}]$,

$$\begin{aligned}R_1 &\ge I(U_1; X_1|W, T)\\R_2 &\ge I(U_2; X_2|W, T)\\R_1 + R_2 &\ge I(U_1, U_2; X_1, X_2|W, T)\\R_{01} &\ge I(U_1; X_1, X_2, Z, Y|W, T) - 2g(\epsilon)\\R_{02} &\ge I(U_2; X_1, X_2, Z, Y|W, T) - 2g(\epsilon)\\R_{01} + R_{02} &\ge I(U_1, U_2; X_1, X_2, Z, Y|W, T) - 2g(\epsilon),\end{aligned}$$

with

$$\begin{aligned}g(\epsilon) = 2\sqrt{\epsilon}\bigg(&H_q(X_1, X_2, W, Z, Y) + R_1 + R_2\\&+ \log\frac{(|\mathcal{X}_1||\mathcal{X}_2||\mathcal{W}||\mathcal{Z}||\mathcal{Y}|)}{\epsilon}\bigg)\end{aligned}$$

(where $g(\epsilon) \to 0$ *as* $\epsilon \to 0$*), for some p.m.f.*

$$\begin{aligned}p(x_1,&x_2, w, z, t, u_1, u_2, y) =\\&p(x_1, x_2, w, z)p(t)p(u_1, u_2|x_1, x_2, t)p(y|u_1, u_2, w, t)\end{aligned}\tag{6}$$

such that

$$p(u_1|x_1, x_2, w, z, t) = p(u_1|x_1, t), \tag{7}$$

$$p(u_2|x_1, x_2, w, z, t) = p(u_2|x_2, t), \tag{8}$$

$$||p(x_1, x_2, w, z, y|t) - q(x_1, x_2, w, z, y)||_1 \le \epsilon \text{ for all } t.$$

The details of the proof can be found in the full version [11]. It should be noted that the outer bound in Theorem 2 represents an epsilon rate region, as discussed in [3, Section VI-C].

Remark 2. Compared to the setting of [7] without secrecy constraints, the proof of Theorem 2 in the long version [11] makes use of the stronger constraint (10) implied by a successful code in the current setting. This allows us to prove the strong secrecy constraint as in steps (12)–(14) in Appendix A. Another crucial difference compared to [7] is that the stricter lower bounds on the shared randomness rates in the current setting (due to the dual purpose of the shared randomness variables for achieving coordination as well as encrypting the public communication), R_{01} in (16) of Appendix A, are facilitated by the same strong secrecy constraint.

When the random variables (X_1, X_2, W) are mutually independent, and one of the shared randomness rates is unlimited, we can demonstrate the tightness of the inner bound presented in Theorem 1. This is accomplished by obtaining a stronger outer bound than the one in Theorem 2 and establishing cardinality bounds on the auxiliary variables, that allows us to prove the continuity of the outer bound at $\epsilon = 0$, which in turn allows us to fully characterize the region $\mathcal{R}^{\text{secrecy}}_{\text{coordination, } R_{02} \to \infty}$.

Theorem 3 (Tight Characterization - Independent Sources)**.** *Consider a target p.m.f. $q_{X_1X_2WZY}$ such that the random variables (X_1, X_2, W) are mutually independent, i.e., $q_{X_1X_2W} = q_W q_{X_1} q_{X_2}$. Then the rate region $\mathcal{R}^{\text{secrecy}}_{\text{coordination, } R_{02} \to \infty}$ is specified by the set of all rate triples (R_1, R_2, R_{01}) such that*

$$\begin{aligned} R_1 &\geq I(U_1; X_1|T) \\ R_2 &\geq I(U_2; X_2|T) \\ R_{01} &\geq I(U_1; X_1, Z, Y|X_2, W, T), \end{aligned}$$

for some p.m.f.

$$\begin{aligned} p(x_1,&x_2, w, z, t, u_1, u_2, y) = \\ &p(w)p(x_1)p(x_2)p(z|x_1, x_2, w)p(t) \\ &\times \prod_{j=1}^{2} p(u_j|x_j, t)p(y|u_1, u_2, w, t) \end{aligned} \tag{9}$$

such that

$$\sum_{u_1,u_2} p(x_1, x_2, w, z, u_1, u_2, y|t) = q(x_1, x_2, w, z, y), \textit{ for all } t,$$

with $|\mathcal{U}_1| \leq |\mathcal{X}_1||\mathcal{X}_2||\mathcal{W}||\mathcal{Z}||\mathcal{Y}|$, $|\mathcal{U}_2| \leq |\mathcal{U}_1||\mathcal{X}_1||\mathcal{X}_2||\mathcal{W}||\mathcal{Z}||\mathcal{Y}|$ and $|\mathcal{T}| \leq 3$.

The converse of Theorem 3 involves deriving a single-letter characterization that exhibits a p.m.f. structure matching that of the inner bound in Theorem 1. This is accomplished by leveraging the independence condition on the sources. The achievability follows from Theorem 1 by invoking the independence condition $q(x_1, x_2, w) = q(w)q(x_1)q(x_2)$ along with the fact that R_{02} is unconstrained. For a detailed proof of converse, please refer to Appendix A.

IV. Conclusion

We investigated secure strong coordination in a multiple-access network of noiseless links in the presence of an external eavesdropper. General inner and outer bounds were derived on the rate region of communication and shared randomness rates, along with a tight characterization for the special case of independent sources. It would be interesting to explore a more general scenario where the noiseless links are replaced by a (noisy) multiple access channel as a resource, which is part of our ongoing work.

Appendix A
Converse Proof of Theorem 3

Consider a coding scheme that induces a joint distribution on $(X_1^n, X_2^n, W^n, Z^n, Y^n, M_1, M_2)$ which satisfies

$$||p_{X_1^n,X_2^n,W^n,Z^n,Y^n,M_1,M_2} - p_{M_1,M_2} q^{(n)}_{X_1X_2WZY}||_1 \leq \epsilon, \tag{10}$$

for $\epsilon \in (0, \frac{1}{4}]$. To simplify notation, we define $\Theta_{\sim i} \triangleq (\Theta^{i-1}, \Theta^n_{i+1})$ for any vector Θ^n. The following lemma will be useful in establishing the outer bound.

Lemma 1. *[12, Lemma 6] Let p_{S^n} be such that $||p_{S^n} - q_S^{(n)}||_1 \leq \epsilon$, where $q_S^{(n)}(s^n) = \prod_{i=1}^n q_S(s_i)$, then*

$$\sum_{i=1}^{n} I_p(S_i; S_{\sim i}) \leq n g_1(\epsilon), \tag{11}$$

with $g_1(\epsilon) = 2\sqrt{\epsilon}\left(H(S) + \log|\mathcal{S}| + \log\frac{1}{\sqrt{\epsilon}}\right) \to 0$ as $\epsilon \to 0$.

We take $S = (X_1, X_2, W, Z, Y)$ in Lemma 1 for our purposes. We note that for $\epsilon \in (0, \frac{1}{4}]$, the following holds

$$\begin{aligned} \max&\left\{g_1(\epsilon), 4\epsilon\left(\log|\mathcal{S}| + \log\frac{1}{\epsilon}\right)\right\} \leq g(\epsilon) \\ &:= 2\sqrt{\epsilon}\left(H(S) + R_1 + R_2 + \log|\mathcal{S}| + 2\log\frac{1}{\sqrt{\epsilon}}\right) \\ &= 2\sqrt{\epsilon}\Bigg(H(X_1, X_2, W, Z, Y) + R_1 + R_2 \\ &\qquad + \log|\mathcal{X}_1||\mathcal{X}_2||\mathcal{W}||\mathcal{Z}||\mathcal{Y}| + 2\log\frac{1}{\sqrt{\epsilon}}\Bigg), \end{aligned} \tag{12}$$

which is $g(\epsilon)$ in Theorem 2. Thus, one can replace $g_1(\epsilon)$ in Lemma 1 by $g(\epsilon)$, which satisfies $\lim_{\epsilon\to 0} g(\epsilon) = 0$ as well.

We first prove the strong secrecy constraint. We can bound the total variation distance between the induced joint distribution on $(X_1^n, X_2^n, W^n, Z^n, Y^n, M_1, M_2)$ and the product of the induced distributions on $(X_1^n, X_2^n, W^n, Z^n, Y^n)$ and (M_1, M_2) as follows:

$$\begin{aligned} ||&p_{X_1^n,X_2^n,W^n,Z^n,Y^n,M_1,M_2} - p_{X_1^n,X_2^n,W^n,Z^n,Y^n} p_{M_1,M_2}||_1 \\ &\overset{(a)}{\leq} ||p_{X_1^n,X_2^n,W^n,Z^n,Y^n,M_1,M_2} - p_{M_1,M_2} q^{(n)}_{X_1X_2WZY}||_1 \\ &\quad + ||p_{X_1^n,X_2^n,W^n,Z^n,Y^n} p_{M_1,M_2} - p_{M_1,M_2} q^{(n)}_{X_1X_2WZY}||_1 \\ &= ||p_{X_1^n,X_2^n,W^n,Z^n,Y^n,M_1,M_2} - p_{M_1,M_2} q^{(n)}_{X_1X_2WZY}||_1 \\ &\quad + ||p_{X_1^n,X_2^n,W^n,Z^n,Y^n} - q^{(n)}_{X_1X_2WZY}||_1 \\ &\overset{(b)}{\leq} 2||p_{X_1^n,X_2^n,W^n,Z^n,Y^n,M_1,M_2} - p_{M_1,M_2} q^{(n)}_{X_1X_2WZY}||_1 \\ &\overset{(c)}{\leq} 2\epsilon, \end{aligned} \tag{13}$$

where (a) follows from triangle inequality, (b) follows from [3, Lemma V.1] while (c) follows by the definition of achievability, i.e., from (10). Consequently, it follows from [13, Theorem 17.3.3] that the mutual information can be bounded as

$$\begin{aligned} I(&X_1^n, X_2^n, W^n, Z^n, Y^n; M_1, M_2) \\ &= H(X_1^n, X_2^n, W^n, Z^n, Y^n) + H(M_1, M_2) \end{aligned}$$

$$
\begin{aligned}
&- H(X_1^n, X_2^n, W^n, Z^n, Y^n, M_1, M_2)\\
&\le 4n\epsilon\bigg(\log|\mathcal{X}_1| + \log|\mathcal{X}_2| + \log|\mathcal{W}| + \log|\mathcal{Z}| + \log|\mathcal{Y}|\\
&\qquad + R_1 + R_2 + \log\frac{1}{\epsilon}\bigg) \le ng(\epsilon). \qquad (14)
\end{aligned}
$$

Let us next prove the lower bound on R_j for $j \in \{1, 2\}$.

$$
\begin{aligned}
nR_j &\ge H(M_j) \ge H(M_j|K_j)\\
&\ge I(M_j; X_j^n|K_j) \overset{(a)}{=} I(M_j, K_j; X_j^n)\\
&= \sum_{i=1}^{n} I(M_j, K_j; X_{ji}|X_{j,i+1}^n)\\
&\overset{(b)}{=} \sum_{i=1}^{n} I(M_j, K_j, X_{j,i+1}^n; X_{ji})\\
&\overset{(c)}{=} \sum_{i=1}^{n} I(U_{ji}; X_{ji}) \overset{(d)}{=} nI(U_{jT}; X_{jT}|T)\\
&\overset{(e)}{=} nI(U_j; X_j|T), \qquad (15)
\end{aligned}
$$

where (a) follows from the independence between K_j and X_j^n, (b) follows from the i.i.d. nature of X_{ji} for $i = 1, \dots, n$, (c) follows from an identification of $U_{1i} = (M_1, K_1, X_{1,i+1}^n)$ and $U_{2i} = (M_2, K_2, X_{2,i+1}^n)$, (d) follows by the introduction of a uniform time-sharing random variable $T \in [1 : n]$ that is independent of all other variables, while (e) follows by defining $U_1 := U_{1T}$, $U_2 := U_{2T}$, $X_1 := X_{1T}$, $X_2 := X_{2T}$, $Y := Y_T$, $W := W_T$ and $Z := Z_T$.

Let us next derive the lower bound on R_{01}.

$$
\begin{aligned}
&nR_{01} = H(K_1)\\
&\ge H(K_1|M_1, X_2^n, W^n)\\
&\ge I(K_1; X_1^n, Z^n, Y^n|M_1, X_2^n, W^n)\\
&= I(M_1, K_1; X_1^n, Z^n, Y^n|X_2^n, W^n)\\
&\qquad - I(M_1; X_1^n, Z^n, Y^n|X_2^n, W^n)\\
&\overset{(a)}{\ge} I(M_1, K_1; X_1^n, Z^n, Y^n|X_2^n, W^n) - ng(\epsilon)\\
&= \sum_{i=1}^{n} I(M_1, K_1; X_{1i}, Z_i, Y_i|X_{1,i+1}^n, Z_{i+1}^n\\
&\qquad\qquad Y_{i+1}^n, X_2^n, W^n) - ng(\epsilon)\\
&= \sum_{i=1}^{n} I(M_1, K_1, X_{1,i+1}^n, Z_{i+1}^n, Y_{i+1}^n, X_{2\sim i},\\
&\qquad\qquad W_{\sim i}; X_{1i}, Z_i, Y_i|X_{2i}, W_i)\\
&\quad - \sum_{i=1}^{n} I(X_{1,i+1}^n, Z_{i+1}^n, Y_{i+1}^n, X_{2\sim i}, W_{\sim i}; X_{1i},\\
&\qquad\qquad Z_i, Y_i|X_{2i}, W_i) - ng(\epsilon)\\
&\overset{(b)}{\ge} \sum_{i=1}^{n} I(M_1, K_1, X_{1,i+1}^n; X_{1i}, Z_i, Y_i|X_{2i}, W_i) - 2ng(\epsilon)\\
&= \sum_{i=1}^{n} I(U_{1i}; X_{1i}, Z_i, Y_i|X_{2i}, W_i) - 2ng(\epsilon)\\
&= nI(U_{1T}; X_{1T}, Z_T, Y_T|X_{2T}, W_T, T) - 2ng(\epsilon)\\
&= nI(U_1; X_1, Z, Y|X_2, W, T) - 2ng(\epsilon), \qquad (16)
\end{aligned}
$$

where (a) follows from (14) and (b) follows since

$$
\begin{aligned}
&\sum_{i=1}^{n} I(X_{1,i+1}^n, Z_{i+1}^n, Y_{i+1}^n, X_{2\sim i}, W_{\sim i}; X_{1i}, Z_i, Y_i|X_{2i}, W_i)\\
&\le \sum_{i=1}^{n} I(X_{1\sim i}, X_{2\sim i}, Y_{\sim i}, Z_{\sim i}, W_{\sim i}; X_{1i}, X_{2i}, Y_i, Z_i, W_i)\\
&\le ng(\epsilon) \qquad (17)
\end{aligned}
$$

by (10) and Lemma 1. When R_{02} is unlimited, the auxiliary random variable alphabet cardinalities can be limited to:

$$
\begin{aligned}
&|\mathcal{U}_1| \le |\mathcal{X}_1||\mathcal{X}_2||\mathcal{W}||\mathcal{Z}||\mathcal{Y}|,\\
&|\mathcal{U}_2| \le |\mathcal{U}_1||\mathcal{X}_1||\mathcal{X}_2||\mathcal{W}||\mathcal{Z}||\mathcal{Y}|, \text{ and } |\mathcal{T}| \le 3,
\end{aligned}
$$

where $|\mathcal{T}|$ follows using the support lemma [14, Appendix C]. On the other hand, the cardinalities of U_1 and U_2 can be restricted as above following the perturbation method of [15]. Finally, by invoking the continuity properties of total variation distance and mutual information in the probability simplex, similar to the approach in [3, Lemma VI.5] and [4, Lemma 6], the converse for Theorem 3 is complete.

References

[1] P. Cuff, H. Permuter, and T. Cover, "Coordination capacity," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4181–4206, 2010.

[2] C. Bennett, I. Devetak, A. Harrow, P. Shor, and A. Winter, "The quantum reverse Shannon theorem and resource tradeoffs for simulating quantum channels," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2926–2959, 2014.

[3] P. Cuff, "Distributed channel synthesis," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7071–7096, 2013.

[4] M. Yassaee, A. Gohari, and M. Aref, "Channel simulation via interactive communications," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 2964–2982, 2015.

[5] S. Satpathy and P. Cuff, "Secure cascade channel synthesis," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6081–6094, 2016.

[6] V. Ramachandran, S. R. B. Pillai, and V. M. Prabhakaran, "Strong coordination with side information," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 1564–1569.

[7] G. R. Kurri, V. Ramachandran, S. R. B. Pillai, and V. M. Prabhakaran, "Multiple access channel simulation," *IEEE Transactions on Information Theory*, vol. 68, no. 11, pp. 7575–7603, 2022.

[8] A. Gohari, M. H. Yassaee, and M. R. Aref, "Secure channel simulation," in *2012 IEEE Information Theory Workshop*. IEEE, 2012, pp. 406–410.

[9] D. Data, G. R. Kurri, J. Ravi, and V. M. Prabhakaran, "Interactive secure function computation," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5492–5521, 2020.

[10] C. E. Shannon, "Communication theory of secrecy systems," *The Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[11] V. Ramachandran, T. Oechtering, and M. Skoglund, "Multi-terminal strong coordination over noiseless networks with secrecy constraints." [Online]. Available: https://www.tinyurl.com/zurichl1

[12] G. Cervia, L. Luzzi, M. Le Treust, and M. Bloch, "Strong coordination of signals and actions over noisy channels with two-sided state information," *IEEE Transactions on Information Theory*, 2020.

[13] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[14] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.

[15] A. A. Gohari and V. Anantharam, "Evaluation of Marton's inner bound for the general broadcast channel," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 608–619, 2012.

Author Index