International Zurich Seminar on Information and Communication (IZS 2020)

Proceedings

International Zurich Seminar on Information and Communication

February 26 – 28, 2020

Sorell Hotel Zürichberg, Zurich, Switzerland

Proceedings

Acknowledgment of Support

The International Zurich Seminar on Information and Communication (IZS)
is organized by 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

Yair Be'ery
Stephan ten Brink
Shraga Bross
Yuval Cassuto
Terence H. Chan
Giuseppe Durisi
Robert Fischer
Bernard Fleury
Albert Guillén i Fàbregas
Martin Hänggi
Franz Hlawatsch
Ashish Khisti
Tobias Koch
Gerhard Kramer
Frank Kschischang
Hans-Andrea Loeliger
Ron Roth
Igal Sason
Robert Schober
Yossef Steinberg
Ido Tal
Giorgio Taricco
Emre Telatar
Pascal Vontobel
Ligong Wang
Armin Wittneben
Ram Zamir

Organizers of Invited Sessions

Alex Alvarado
Thomas Mittelholzer
Haim H. Permuter
Yury Polyanskiy
Frans M. J. Willems

Local Organization

Silvia Tempel (Secretary)
Michael Lerjen (Web and Publications)
Patrick Strebel (Registration)

Table of Contents

Keynote Lectures

Session 1 — Wed 10:00 – 11:20

Signal-Shaping Methods

Invited session organizers: Frans M. J. Willems and Alex Alvarado (Eindhoven University of Technology)

---

*Invited papers are marked by an asterisk.

Session 2 Wed 13:30 – 14:50

Coding Theory and Applications

Chaired by Yuval Cassuto (Technion – Israel Institute of Technology)

Session 3 Wed 15:20 – 17:00

Coding for Data Storage and for Low Error-Rate Applications

Invited session organizer: Thomas Mittelholzer (HSR University of Applied Sciences)

Session 4 — Thu 10:00 – 11:40

Information Inequalities and Statistics

Invited session organizer: Yury Polyanskiy (MIT)

Session 5 — Thu 13:30 – 14:50

Shannon Theory

Chaired by Emmanuel Abbé (EPFL)

Session 6 Thu 15:20 – 17:00

Machine Learning and Information Theory

Invited session organizer: Haim H. Permuter (Ben Gurion University)

Session 7 Fri 10:00 – 11:40

Classical and Nonclassical Information Measures

Chaired by Tobias Koch (Universidad Carlos III de Madrid)

Session 8 Fri 13:30 – 14:50

Information Theoretic Aspects of Communication

Chaired by Robert Fischer (Ulm University)

Session 9 Fri 15:20 – 16:40

Modern Detection Theory

Chaired by Sergey Loyka (University of Ottawa)

Recent-Results Posters

Wednesday, February 26

Adaptive Coded Modulation Scheme for Free-Space Optical Communication
Ahmed Elzanaty (King Abdullah University of Science and Technology, Saudi Arabia)
Mohamed-Slim Alouini (King Abdullah University of Science and Technology, Saudi Arabia)

Achievable Information Rates of Probabilistic Amplitude Shaping: An Alternative Approach via Random Sign-Coding Arguments
Yunus Can Gültekin (Eindhoven University of Technology, Eindhoven, The Netherlands)
Alex Alvarado (Eindhoven University of Technology, Eindhoven, The Netherlands)
Frans M. J. Willems (Eindhoven University of Technology, Eindhoven, The Netherlands)

On Achieving Low Bit Error Rates with QAM Signaling and LDPC Coding over the AWGN
Gada Rezgui (CY Paris University, Cergy-Pontoise, France)
Iryna Andriyanova (CY Paris University, Cergy-Pontoise, France)
Charly Poulliat (University of Toulouse, Toulouse, France)
Asma Maaloui (University of Toulouse, Toulouse, France)

Thursday, February 27

Secure Distributed Multiple Matrix Multiplication
Nitish Mital (Imperial College London, UK)
Cong Ling (Imperial College London, UK)
Deniz Gündüz (Imperial College London, UK)

Finite Blocklength Rate for Soft Covering
Lanqing Yu (Princeton University, Princeton, USA)
Semih Yagli (Princeton University, Princeton, USA)
Paul Cuff (Renaissance Technologies LLC, Long Island, USA)

From Feedback Capacity to Tight Achievable Rates without Feedback for AGN Channels with Stable and Unstable Autoregressive Noise
Christos Kourtellaris (University of Cyprus, Nicosia, Cyprus)
Charalambos D. Charalambous (University of Cyprus, Nicosia, Cyprus)
Sergey Loyka (University of Ottawa, Ontario, Canada)

Friday, February 28

Topology Optimization for 6G Networks: A Network Information-Theoretic Approach
Abdulkadir Çelik (King Abdullah University of Science and Technology, Saudi Arabia)
Anas Chaaban (University of British Columbia, Vancouver, Canada)
Basem Shihada (King Abdullah University of Science and Technology, Saudi Arabia)
Mohamed-Slim Alouini (King Abdullah University of Science and Technology, Saudi Arabia)

Generalized Gaussian Model for Data-Driven Learning in Communications
Khac-Hoang Ngo (LSS, CentraleSupélec, France; Paris Research Center, Huawei Technologies, France)
Sheng Yang (LSS, CentraleSupélec, France)
Maxime Guillaud (Paris Research Center, Huawei Technologies, France)

Weight Enumeration, RM-Polar Codes, List Decoding
Kumud S. Altmayer (University of Maryland, Adelphi, USA; University of Virginia, Charlottesville, USA)

Indoor Location Estimation based on Images and Object Identification
Dimitris Milioris (Nokia Bell Labs, France)
Philippe Jacquet (Inria Paris-Saclay, France)

Signal Shaping: Fundamentals, Potentials, and Techniques

Robert F.H. Fischer

Institut für Nachrichtentechnik, Universität Ulm, Ulm, Germany
Email: robert.fischer@uni-ulm.de

***Abstract*—*Source coding* and *channel coding* are well-established fields providing various and flexible techniques for eliminating redundancy and protecting data against errors, respectively. Nowadays transmission systems extensively utilize source and channel coding techniques adapted as closely as possible to the specific situations. Less common is the application of *signal shaping*—in principle, the task of signal shaping is to generate (transmit) signals which meet specific demands. The most popular aim of signal shaping is to generate signals with least average power. Without sacrificing performance, this is possible by replacing uniformly distributed transmit symbols by Gaussian ones.**

Other shaping aims as, e.g., controlling the *power spectral density* [3], limiting the dynamic range [6], or enhancing the performance of schemes for physical-layer security [7] are possible, too.

In some sense, source and channel coding are dual to each other. Signal shaping can be seen as dual to both source and channel coding—these three operations complement each other and schemes from one field can be transferred to the other ones, cf., e.g., the utilization of linear channel codes for source coding [1].

In the talk, first the fundamentals and potentials of signal shaping are explained. The possible gains and principle operations are derived from basic geometry. Based on the dualities, specific signal-shaping techniques are classified and explained. This includes the use of a *source decoder as shaping encoder* [4] which currently is experiencing a renaissance via a so-called distribution matcher [2], *trellis shaping* [5] as dual operation to Ungerböcks *trellis coding*, and *shell mapping* which has a vector-quantization counterpart.

Details on the fundamentals and schemes can be found in the monography [3].

References

[1] T.C. Ancheta. Syndrome-Source-Coding and its Universal Generalization. *IEEE Transactions on Information Theory*, vol. 22, no. 4, pp. 432–436, July 1976.

[2] G. Böcherer, F. Steiner, P. Schulte. Bandwidth Efficient and Rate-Matched Low-Density Parity-Check Coded Modulation. *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4651–4665, Dec. 2015.

[3] R.F.H. Fischer. *Precoding and Signal Shaping for Digital Transmission*, John Wiley & Sons, New York, 2002.

[4] G.D. Forney, R.G. Gallager, G.R. Lang, F.M. Longstaff, and S.U.H. Qureshi. Efficient Modulation for Band-Limited Channels. *IEEE Journal on Selected Areas in Communications*, vol. 2, no. 5, pp. 632–647, Sep. 1984.

[5] G.D. Forney. Trellis Shaping. *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 281–300, March 1992.

[6] W. Henkel, B. Wagner. Another Application for Trellis Shaping: PAR Reduction for DMT (OFDM). *IEEE Transactions on Communications*, vol. 48, no. 9, pp. 1471–1476, Sep. 2000.

[7] J. Pfeiffer, R.F.H. Fischer. Multilevel Coding for Physical-Layer Security in Optical Networks. In *19th ITG-Symposium on Photonic Networks*, Leipzig, Germany, May 2018.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Probabilistic Shaping: A Random Coding Experiment

Georg Böcherer
Optical Communications Technology Lab
Huawei France Technologies S.A.S.U.
Boulogne-Billancourt 92100, France
Email: georg.boecherer@ieee.org

Patrick Schulte and Fabian Steiner
Technical University of Munich
Institute for Communications Engineering
80333 München, Germany
Email: {patrick.schulte, fabian.steiner}@tum.de

***Abstract*—A layered probabilistic shaping (PS) ensemble is considered, which contains probabilistic amplitude shaping (PAS) as a practical instance. Layered PS consists of an inner layer for forward error correction (FEC) and an outer layer for PS. In the PS layer, message bits are mapped to FEC encoder inputs that map to channel input sequences in a shaping set. The shaping set specifies desired properties, for instance, it may consist of all sequences that have a capacity-achieving distribution for the considered channel. By random coding arguments, the probability of encoding failure and decoding failure is analyzed and it is shown that the layered PS architecture is capacity-achieving for a discrete input memoryless channel. Practical achievable spectral efficiencies of the layered PS architecture are discussed.**

I. Introduction

Probabilistic amplitude shaping (PAS) was proposed in [1] to integrate non-uniform channel input distributions with off-the-shelf linear forward error correction (FEC) codes. PAS quickly found industrial application in transceivers for fiber-optic transmission, e.g., [2]–[4]. Since PAS is not a sample of the classical random code ensemble (see Remarks 1, 2, and 3), the calculation of appropriate achievable rates for PAS is intricate, and several attempts were taken [2, Sec. III.C], [5], [6]. In [7] and [8, Chap. 10], achievable rates for PAS are derived using random sign coding and partially systematic FEC encoding. In this work, we discuss layered probabilistic shaping (PS), a random code ensemble that was developed in the line of work [8]–[11]. Layered PS contains PAS as a practical instance, but is more general, e.g., it also covers the probabilistic parity bit shaping proposed in [12].

In Sec. II, we define layered PS and derive a general channel coding theorem. In Sec. III, we show that layered PS achieves the capacity of discrete input memoryless channels and discuss practical matched and mismatched decoding metrics.

II. Layered Probabilistic Shaping

Consider a channel with finite input alphabet $\mathcal{X}$ and define

$$m = \log_2 |\mathcal{X}|. \tag{1}$$

The channel output alphabet can be continuous or discrete.

A. Classical Random Code Ensemble

The classical random code ensemble [13, Ch. 5] for a channel with input alphabet $\mathcal{X}$ and codeword length n symbols in $\mathcal{X}$ is

$$\mathcal{C} = \left\{C^n(w), w = 1, 2, \ldots, 2^{nmR_{\text{fec}}}\right\} \tag{2}$$

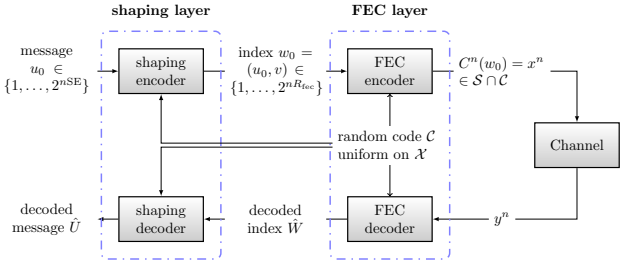


Fig. 1. The layered PS architecture discussed in Sec. II. In PAS [1], the FEC encoder is systematic and the shaping encoder is realized by a DM [14]–[19] that shapes the systematic symbols. The shaping encoder of PAS is zero error.

where the entries of the $|\mathcal{C}| = 2^{nmR_{\text{fec}}}$ codewords are independently and identically distributed according to P_X on the constellation $\mathcal{X}$. We require $0 \leq R_{\text{fec}} \leq 1$ so that $mR_{\text{fec}} \leq \log_2 |\mathcal{X}|$. By [13, Eq. (5.2.5)], the decoding rule for a memoryless channel with transition density $p_{Y|X}$ is

$$\hat{w} = \underset{w \in \{1,\ldots,|\mathcal{C}|\}}{\operatorname{argmax}} \prod_{i=1}^{n} p_{Y|X}(y_i|c_i(w)) \tag{3}$$

where y^n is the sequence observed at the channel output. The spectral efficiency (SE) in bits per channel use is $\text{SE} = mR_{\text{fec}}$ and the classical random code ensemble achieves

$$\text{SE}^* = \mathbb{I}(X;Y). \tag{4}$$

In particular, it achieves the capacity $\max_{P_X} \mathbb{I}(X;Y)$ when the optimal P_X is used.

B. Layered Random Code Ensemble

The layered PS architecture is displayed in Fig. 1. We consider the random code ensemble

$$\mathcal{C} = \left\{C^n(w), w = 1, 2, \ldots, 2^{nmR_{\text{fec}}}\right\} \tag{5}$$

where the entries of the $|\mathcal{C}| = 2^{nmR_{\text{fec}}}$ codewords are chosen independently and *uniformly* distributed on the constellation $\mathcal{X}$. As above, we require $0 \leq R_{\text{fec}} \leq 1$.

Remark 1. Note that the classical random code ensemble of Sec. II-A samples the codeword entries according to the desired channel input distribution P_X. In contrast, layered PS always uses the uniform distribution.

TABLE I
PS AND FEC OVERHEADS

	FEC	Shaping Set
Rate	R_{fec}	$R_{\text{ss}} = \frac{\log_2 \lvert\mathcal{S}\rvert}{nm}$
Redundancy	$1 - R_{\text{fec}}$	$1 - R_{\text{ss}}$
Overhead in %	$100 \cdot \left(\frac{1}{R_{\text{fec}}} - 1\right)$	$100 \cdot \left(\frac{1}{R_{\text{ss}}} - 1\right)$
Total overhead in %	$100 \cdot \left(\frac{1}{R_{\text{ss}}+R_{\text{fec}}-1} - 1\right)$	

C. Encoding

We consider a general shaping set $\mathcal{S} \subseteq \mathcal{X}^n$. Define the shaping set rate by

$$R_{\text{ss}} = \frac{\log_2 |\mathcal{S}|}{nm}. \tag{6}$$

Note that by the definition of m in (1), $0 \leq R_{\text{ss}} \leq 1$. We divide the FEC code into $2^{n\text{SE}}$ partitions, so that the number of codewords in each partition is

$$\frac{2^{nmR_{\text{fec}}}}{2^{n\text{SE}}} = 2^{nm(R_{\text{fec}} - \frac{\text{SE}}{m})}. \tag{7}$$

The PS encoder maps message $u \in \{1, 2, \ldots, 2^{n\text{SE}}\}$ to a codeword in the uth partition that is in $\mathcal{S}$. By double indexing $\mathcal{C}$, the chosen codeword has index $w = (u, v)$ for some $v \in \{1, 2, \ldots, 2^{nm(R_{\text{fec}} - \frac{\text{SE}}{m})}\}$. An encoding error occurs if the PS encoder cannot find such a codeword.

Theorem 1 ([11, Theorem 1])**.** *The probability that the PS encoder cannot map its input to a codeword in $\mathcal{S} \cap \mathcal{C}$ is upper bounded by*

$$\begin{aligned}&\Pr(\text{PS encoding failure})\\ &\quad\leq \exp\left(-2^{nm\left[1-(1-R_{\text{ss}})-(1-R_{\text{fec}})-\frac{\text{SE}}{m}\right]}\right).\end{aligned} \tag{8}$$

Remark 2. By the theorem, the SE is determined by two overheads (see Table I), namely the PS overhead and the FEC overhead. For a desired SE, the overhead allocation is a degree of freedom that can be exploited in the transceiver design, for example, a low FEC overhead may be desirable for complexity reasons. Note that in the classical random coding experiment, the SE is always equal to mR_{fec}.

D. Decoding

We consider a generic FEC decoder with a decoding metric q. For an observation y^n, the metric assigns to each sequence $x^n \in \mathcal{X}^n$ a non-negative score $q(x^n, y^n)$ (see [11, Sec. V.A] for the definition and detailed discussion of non-negative scores). The FEC encoder maps a message w to a codeword $c^n(w)$. For an observed output y^n, the decoder outputs as its decision the message that maps to the codeword with the maximum score, i.e,

$$\hat{w} = \underset{w \in \{1,\ldots,|\mathcal{C}|\}}{\operatorname{argmax}} \; q\left(c^n(w), y^n\right). \tag{9}$$

Theorem 2 ([11, Theorem 2])**.** *Suppose the codeword $C^n(w_0) = x^n$ is transmitted, let y^n be a channel output sequence, and let q be a non-negative decoding metric. Define the empirical cross-entropy*

$$\text{ж}(q, x^n, y^n) = -\frac{1}{n} \log_2 \frac{q(x^n, y^n)}{\sum_{a^n \in \mathcal{X}^n} q(a^n, y^n)}. \tag{10}$$

The probability that the decoder (9) *does not recover the index w_0 from the sequence y^n is bounded from above by*

$$\begin{aligned}&\Pr(\hat{W} \neq w_0 | C^n(w_0) = x^n, Y^n = y^n)\\ &\quad\leq 2^{-nm\left(1 - R_{\text{fec}} - \frac{\text{ж}(q, x^n, y^n)}{m}\right)}.\end{aligned} \tag{11}$$

Note that in Fig. 1, if the index decision $\hat{W}$ is correct, then the shaping decoder can error-free recover the message u_0 from $\hat{W}$. That is, $\Pr(\hat{W} \neq w_0)$ upper bounds $\Pr(\hat{U} \neq u_0)$.

E. Channel Coding Theorem

We now consider a memoryless channel

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i) \tag{12}$$

and memoryless decoding metrics

$$q(x^n, y^n) = \prod_{i=1}^{n} q(x_i, y_i). \tag{13}$$

Furthermore, we require that most sequences in the shaping set $\mathcal{S}$ have the distribution P_X, so that with high probability

$$\text{ж}(q, X^n, Y^n) \approx \mathbb{E}\left[\text{ж}(q, X, Y)\right] =: \mathbb{X}(q, X, Y) \tag{14}$$

where $\mathbb{X}(q, X, Y)$ is a cross-entropy. By Theorems 1 and 2, following the line of arguments in [20] (leaving out the ϵs and δs) we arrive at the following channel coding theorem.

Corollary 1. *For a shaping set with distribution P_X, an achievable spectral efficiency allowing for successful encoding and decoding with high probability is*

$$\text{SE}^* = \left[mR_{\text{ss}} - \mathbb{X}(q, X, Y)\right]^+ \tag{15}$$

where $[\cdot]^+ = \max\{0, \cdot\}$ ensures non-negativity.

Note that (15) is the same as [11, Eq. (1)] with slightly different notation.

III. DECODING METRICS

We now instantiate the achievable SE in (15) for various shaping sets and decoding metrics. See Table II for an overview.

A. Capacity-Achieving Symbol-Metric

We use as shaping set $\mathcal{S}$ all sequences with distribution P_X. For sufficiently large n, we have $R_{\text{ss}}m \approx \mathbb{H}(X)$. With the decoding metric $P_{X|Y}$, the achievable SE becomes equal to the mutual information $\mathbb{I}(X;Y)$, which shows that the layered PS architecture is capacity-achieving.

Remark 3. Note that the classical random code ensemble achieves capacity with a maximum likelihood (ML) rule on a codebook of size $2^{n\text{SE}}$ while layered PS achieves capacity with a maximum a posteriori (MAP) rule on a codebook of size $2^{n(\text{SE}+m(1-R_{\text{ss}}))}$, which is larger.

TABLE II
IMPORTANT DECODING METRICS

	mR_{ss}	q	$\mathbb{X}^*(q,X,Y)$	SE*
symbol-metric	mR_{ss}	$P_{X\|Y}$	$\mathbb{H}(X\|Y)$	$[mR_{\text{ss}}-\mathbb{H}(X\|Y)]^+$
+ capacity-achieving	$\mathbb{H}(X)$	$P_{X\|Y}$	$\mathbb{H}(X\|Y)$	$\mathbb{I}(X;Y)$
bit-metric	mR_{ss}	$\prod_{i=1}^m P_{B_i\|Y}$	$\sum_{i=1}^m \mathbb{H}(B_i\|Y)$	$\left[mR_{\text{ss}}-\sum_{i=1}^m \mathbb{H}(B_i\|Y)\right]^+$
	$\mathbb{H}(X)$	$\prod_{i=1}^m P_{B_i\|Y}$	$\sum_{i=1}^m \mathbb{H}(B_i\|Y)$	$\left[\mathbb{H}(X)-\sum_{i=1}^m \mathbb{H}(B_i\|Y)\right]^+$
mismatched metric	mR_{ss}	q	$\min_{s>0}\mathbb{X}(q^s,X,Y)$	$\max_{s>0}[mR_{\text{ss}}-\mathbb{X}(q^s,X,Y)]^+$

B. Bit-Metric

Bit metric decoding uses an m-bit label $\boldsymbol{B}=B_1B_2\dots B_m$ of the channel input alphabet and a bit-metric

$$q(\boldsymbol{b},y)=\prod_{i=1}^{m} q_i(b_i,y). \tag{16}$$

Table II shows achievable SEs when $q_i=P_{B_i|Y}$. By defining the L-value $L_i=\log P_{B_i|Y}(0|Y)/P_{B_i|Y}(1|Y)$, the conditional entropy sum can also be written as

$$\sum_{i=1}^{m}\mathbb{H}(B_i|Y)=\sum_{i=1}^{m}\mathbb{E}\left[\log_2\{1+\exp\left[-(1-2B_i)L_i\right]\}\right]. \tag{17}$$

C. Mismatched Metrics

For $s>0$, the non-negative metric q and the metric q^s implement exactly the same decision rule. Consequently, their error probability is the same. This allows us to tighten the error bound in Theorem 2 and thereby the achievable SE in Corollary 1. The tightened cross-entropy is

$$\mathbb{X}^*(q,X,Y)=\min_{s>0}\mathbb{X}(q^s,X,Y). \tag{18}$$

For uniform distributions P_X, the mismatched achievable SE recovers the generalized mutual information (GMI) in [21]. For non-uniform P_X, it is different from the GMI, because in [21], the classical random code ensemble of Sec. II-A is considered.

IV. CONCLUSIONS

We defined layered probabilistic shaping (PS) and derived achievable rates. In particular, we showed that layered PS is capacity-achieving for a particular shaping sets and decoding metrics. Several differences between layered PS and the classical random code ensemble were pointed out. The achievable rates of layered PS are directly applicable for probabilistic amplitude shaping (PAS). An interesting future work is the study of finite length error exponents for layered PS, accounting for the distribution spectrum of the sequences in the shaping set.

REFERENCES

[1] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4651–4665, Dec. 2015.

[2] F. Buchali, F. Steiner, G. Böcherer, L. Schmalen, P. Schulte, and W. Idler, "Rate adaptation and reach increase by probabilistically shaped 64-QAM: An experimental demonstration," *J. Lightw. Technol.*, vol. 34, no. 8, Apr. 2016.

[3] A. Ghazisaeidi, I. F. de Jauregui Ruiz, R. Rios-Müller, L. Schmalen, P. Tran, P. Brindel, A. C. Meseguer, Q. Hu, F. Buchali, G. Charlet, *et al.*, "Advanced C+L-band transoceanic transmission systems based on probabilistically shaped PDM-64QAM," *J. Lightw. Technol.*, vol. 35, no. 7, pp. 1291–1299, Apr. 2017.

[4] J. Li, A. Zhang, C. Zhang, X. Huo, Q. Yang, J. Wang, J. Wang, W. Qu, Y. Wang, J. Zhang, *et al.*, "Field trial of probabilistic-shaping-programmable real-time 200-Gb/s coherent transceivers in an intelligent core optical network," in *Asia Commun. Photonics Conf. (ACP)*, 2018.

[5] J. Cho, L. Schmalen, and P. J. Winzer, "Normalized generalized mutual information as a forward error correction threshold for probabilistically shaped QAM," in *Proc. Eur. Conf. Optical Commun. (ECOC)*, Paper M.2.D.2, Gothenburg, Sweden, Sep. 2017.

[6] T. Yoshida, M. Karlsson, and E. Agrell, "Performance Metrics for Systems With Soft-Decision FEC and Probabilistic Shaping," *IEEE Photon. Technol. Lett.*, vol. 29, no. 23, pp. 2111–2114, Dec. 2017.

[7] R. A. Amjad, "Information rates and error exponents for probabilistic amplitude shaping," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, Nov. 2018.

[8] G. Böcherer, *Principles of coded modulation*, Habilitation thesis, Technical University of Munich, 2018. [Online]. Available: http://www.georg-boecherer.de/bocherer2018principles.pdf.

[9] ——, "Achievable rates for shaped bit-metric decoding," *arXiv preprint*, 2016. [Online]. Available: http://arxiv.org/abs/1410.8075.

[10] G. Böcherer, "Achievable rates for probabilistic shaping," *arXiv preprint*, [Online]. Available: https://arxiv.org/abs/1707.01134v5.

[11] G. Böcherer, P. Schulte, and F. Steiner, "Probabilistic shaping and forward error correction for fiber-optic communication systems," *J. Lightw. Technol.*, vol. 37, no. 2, pp. 230–244, Jan. 2019.

[12] G. Böcherer, D. Lentner, A. Cirino, and F. Steiner, "Probabilistic parity shaping for linear codes," *arXiv preprint*, 2019. [Online]. Available: https://arxiv.org/abs/1902.10648.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[14] G. Böcherer and R. Mathar, "Matching dyadic distributions to channels," in *Proc. Data Compression Conf. (DCC)*, 2011, pp. 23–32.

[15] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 430–434, Jan. 2016.

[16] Y. C. Gültekin, F. M. Willems, W. van Houtum, and S. Serbetli, "Approximate enumerative sphere shaping," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, Colorado, USA, 2018, pp. 676–680.

[17] P. Schulte and F. Steiner, "Divergence-optimal fixed-to-fixed length distribution matching with shell mapping," *IEEE Wireless Commun. Letters*, vol. 8, no. 2, pp. 620–623, Apr. 2019.

[18] T. Yoshida, M. Karlsson, and E. Agrell, "Hierarchical distribution matching for probabilistically shaped coded modulation," *J. Lightw. Technol.*, vol. 37, no. 6, pp. 1579–1589, 2019.

[19] J. Cho, "Prefix-free code distribution matching for probabilistic constellation shaping," *IEEE Trans. Commun.*, 2019, accepted for publication, DOI: 10.1109/TCOMM.2019.2924896.

[20] Y. Lomnitz and M. Feder, "A simpler derivation of the coding theorem," *arXiv preprint*, 2012. [Online]. Available: https://arxiv.org/abs/1205.1389.

[21] G. Kaplan and S. Shamai (Shitz), "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEÜ*, vol. 47, no. 4, pp. 228–239, 1993.

Short-length Probabilistic Shaping: Improved Methods and Mitigation of Fiber Nonlinearities

Tobias Fehenberger
ADVA
Fraunhoferstr. 9a
82152 Martinsried/Munich, Germany
Email: tfehenberger@adva.com

***Abstract*—Advanced amplitude shapers that improve upon the conventional constant-composition distribution matching in terms of rate loss and computational complexity are reviewed. In a comprehensive comparison, we focus on energy considerations, rate loss, and decoding performance. We further study the mitigating effects of short-length probabilistic shaping on the Kerr nonlinearities occurring during optical fiber transmission and discuss the impact of interleavers on this effect.**

In this invited contribution, we discuss probabilistic shaping (PS) in the short-length regime where the block sizes are at most a few hundred symbols. Focusing on probabilistic amplitude shaping (PAS) as underlying coded modulation framework to realize PS [1], we study in detail the amplitude shaping block that maps a block of uniformly distributed data bits into a shaped amplitude sequence. The first amplitude shaper proposed for PAS is constant-composition distribution matching (CCDM) [2], which, as its name suggests, outputs sequences with identical compositions, i.e., they are permutations of each other. While asymptotically lossless, CCDM has suboptimal finite-length performance. Furthermore, the conventional arithmetic coding method used for implementing CCDM is inherently sequential, which introduces latency and limits high-throughput application [3]. It is mainly this combination of requiring long blocks and having a sequential implementation that lead to a great deal of investigation into advanced amplitude shapers. In this contribution, we review such advanced shapers and present a comprehensive comparison. The investigated schemes include multiset-partition distribution matching [4], enumerative sphere shaping [5], [6], and Huffman coded sphere shaping [7]. A numerical analysis of rate loss and performance after forward error correction (FEC) decoding is supported by a study of the signal space occupied by the respective amplitude shaping schemes and the corresponding energy considerations.

In the second part of this contribution, we study the impact that probabilistically shaped signaling has on the Kerr nonlinearities which are present in the optical fiber channel. It has been shown theoretically and demonstrated in simulations that for asymptotically long CCDM block lengths, the effective signal-to-noise ratio (SNR) after fiber transmission and digital signal processing is smaller for shaped signaling than for uniform distributions, which is due to fiber nonlinearities being exacerbated by PS [8]. Surprisingly, the inverse behavior is observed for short CCDM sequences where fiber nonlinearities are mitigated by shaping [9]. This inverse proportionality of SNR with block length can for CCDM be attributed to the fact that certain overall transmit sequences, such as those with long runs of identical amplitudes, cannot occur when several short CCDM blocks are concatenated and combined into a FEC codeword [10]. For long CCDM sequences, on the other hand, this restriction does not apply, and an SNR penalty due to the fiber nonlinearities is observed. We investigate this behavior numerically and show how the utilization of interleavers affects the capability of mitigating fiber nonlinearities by PS.

References

[1] G. Böcherer, P. Schulte, and F. Steiner, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4651–4665, Dec. 2015.

[2] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 430–434, Jan. 2016.

[3] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Parallel-amplitude architecture and subset ranking for fast distribution matching," *IEEE Transactions on Communications*, Jan. 2020. [Online]. Available: https://arxiv.org/abs/1902.08556

[4] ——, "Multiset-partition distribution matching," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 1885–1893, Mar. 2019.

[5] Y. C. Gültekin, W. van Houtum, A. Koppelaar, and F. M. Willems, "Enumerative sphere shaping for wireless communications with short packets," *arXiv*, Mar. 2019. [Online]. Available: https://arxiv.org/abs/1903.10244

[6] Y. C. Gültekin, T. Fehenberger, A. Alvarado, and F. M. J. Willems, "Probabilistic shaping for finite blocklengths: distribution matching and sphere shaping," *arXiv*, Sep. 2019. [Online]. Available: https://arxiv.org/abs/1909.08886

[7] D. S. Millar, T. Fehenberger, T. Yoshida, T. Koike-Akino, K. Kojima, N. Suzuki, and K. Parsons, "Huffman coded sphere shaping with short length and reduced complexity," in *Proc. European Conference on Optical Communications (ECOC)*, Dublin, Ireland, Sep. 2019.

[8] T. Fehenberger, A. Alvarado, G. Böcherer, and N. Hanik, "On probabilistic shaping of quadrature amplitude modulation for the nonlinear fiber channel," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 22, pp. 5063–5073, Nov. 2016.

[9] A. Amari, S. Goossens, Y. C. Gultekin, O. Vassilieva, I. Kim, T. Ikeuchi, C. Okonkwo, F. M. J. Willems, and A. Alvarado, "Introducing enumerative sphere shaping for optical communication systems with short blocklengths," *arXiv:1904.06601 [cs, math]*, Apr. 2019.

[10] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, K. Parsons, and H. Griesser, "Analysis of nonlinear fiber interactions for finite-length constant-composition sequences," *IEEE/OSA Journal of Lightwave Technology*, Sep. 2019.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Hierarchical Distribution Matching with Massively Parallel Interfaces for Fiber-Optic Communications

Tsuyoshi Yoshida, Erik Agrell, and Magnus Karlsson

(Invited Paper)

***Abstract*—The design of the distribution matching (DM) encoder and decoder is essential in the implementation of probabilistic shaping. Recently, techniques for low-complexity implementation of DM have been studied. This work consists of three contributions on this topic. Firstly, the mismatch between required throughput and clock frequency in the electric circuitry of fiber-optic transceivers is explained. The throughput of one DM module determines the number of parallel DM modules needed, which in turn determines the circuit size and power consumption. Our previously proposed hierarchical DM (HiDM) has massively parallel input/output interfaces and thus around 100 times fewer instances are required compared with run-length-coding-based DM. Secondly, the HiDM construction is exemplified and described in detail for a DM word length of more than 100 symbols. Thirdly, the capability of HiDM to shape probability mass functions suitable for nonlinear fiber-optic channels is demonstrated, considering higher-order moments.**

***Index Terms*—Coding, hierarchical distribution matching, implementation, modulation, optical fiber communication, probabilistic shaping, reverse concatenation, throughput.**

I. Introduction

Constellation shaping has been deeply investigated over several decades to approach the Shannon capacity over the additive white Gaussian noise (AWGN) channel. The two main types of shaping schemes are *geometric shaping* [1] and *probabilistic shaping* (PS) [2], [3]. Fiber-optic communication channels with optical amplifiers are suitable target applications for PS. The first reason is the existence of the linear optical amplifier. When the optical signal is shaped, the average optical power inside an optical modulator is reduced, but the power will soon be recovered by optical amplifiers, which gives an almost linear gain without waveform degradation. The second reason is the channel stability because of the confined waveguide (fiber) transmission.

Probabilistic amplitude shaping (PAS) [4] provides an attractive method to implement PS by using reverse concatenation, which means forward error correction (FEC) inside the shaping. The PAS scheme was early examined in optical fiber communications [5] and had a significant impact on the community. The shaping encoding and decoding functions for PAS are called *distribution matching* (DM) and *distribution dematching* (invDM), resp. DMs can be classified in terms of symbolwise or bitwise operation, and further into computation-based or LUT-based. The LUT-based DMs can be further classified into fixed- or variable-length LUTs. A symbolwise DM directly controls the probability mass function (PMF) of the output symbols [6]–[16], whereas a bitwise DM [17]–[19] controls the probability of the output bits in a bit tributary, and the PMF of the symbol is controlled by multiple binary bit tributaries, according to an architecture called bit-level DM or product DM [20], [21]. LUT-based DMs, on the other hand, provide lower computational complexity at the expense of memory. LUT-based DMs with fixed-length interfaces include [7], [10], [12], [17], while others use (virtually) variable-length interfaces [8], [14], [16], [18].

In the original PAS scheme, *constant-composition DM* (CCDM) [6] was employed, which is a symbolwise, computation-based DM, similar to the arithmetic coding scheme proposed in [22]. Our previous works include DM based on *run-length coding* (RLC) [18], which is a bitwise, LUT-based, variable-length DM, and *hierarchical DM (HiDM)* [12], [23], which is a symbolwise, LUT-based, fixed-length DM. HiDM, having a unique tree structure of LUTs, shows good performance, reasonable implementation complexity leading to low power consumption, high throughput, and small error rate increase in the invDM processing.

A main issue in the design of logic circuitry for optical fiber communications is the mismatch between the required throughput (several 100 Gb/s to a few Tb/s) and the clock frequency of the electrical circuitry (several 100 MHz). Most PS coding schemes operate in a highly sequential manner, so that their numbers of input/output bits per clock cycle (throughput) and their numbers of physical wirings (bus widths) would be one or a few bits. To realize transmission at 1 Tb/s using a 500 MHz clock, 2000 parallel *instances* are required if the number of input/output bits is only one bit per clock cycle. Even if one DM module operating at 500 MHz for 500 Mb/s is very small, the total required circuit area would be 2000 times larger. On the other hand, HiDM can input and output several 100 bits or even 1000 bits per clock cycle because of its massively parallel interfaces. To realize 1 Tb/s from a 500 MHz clock, we need just a few instances. This is an important advantage in high-speed fiber-optic communications.

In this work, we firstly raise important issues of throughput and bus width by exemplifying two of our previously proposed DMs [12], [18]. Next, we explain recommended design principles of HiDM in detail to make our previous work

T. Yoshida is with Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, 247-8501, Japan. He also belongs to Graduate School of Engineering, Osaka University, Suita, 505-0871, Japan (e-mail: Yoshida.Tsuyoshi@ah.MitsubishiElectric.co.jp).

M. Karlsson is with the Dept. of Microtechnology and Nanoscience and E. Agrell is with the Dept. of Electrical Engineering, both at Chalmers University of Technology, SE-41296 Gothenburg, Sweden.

This work was partly supported by "Massively Parallel and Sliced Optical Network (MAPLE)," the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan (project no. 20401) and by the Swedish Research Council (project no. 2017-03702).

TABLE I
PARAMETERS IN REVERSE CONCATENATION PS SYSTEMS.

Notation	Description
m	number of bits per QAM symbol
m_{sb}	number of shaped bit tributaries per QAM symbol
N_{s}	number of PAM symbols per DM word
N_{in}	number of information bits per DM word

[12] more reproducible. Finally, we show that the DM word choice in HiDM can be flexibly adapted to different PMFs of the transmitted symbols by considering higher-order moments. This was partly addressed in [24] to improve the performance over nonlinear fiber links.

II. THROUGHPUT AND BUS WIDTH AT THE INTERFACES

In this section, we compare two previously proposed LUT-based DMs in terms of throughput and bus width. Some key parameters of reverse concatenation PS are defined in Tab. I. For simplicity, nonshaped bits are excluded from the explanation.

A. Run-length-coding-based bitwise DM

As a submodule for bit-level DM, we proposed a binary-output DM with RLC (variable-length coding) [18]. The encoder schematic is shown in Fig. 1. The incoming N_{in} bits are first demultiplexed into K lanes. In each lane j, the input bits are mapped into a binary word s_j in which exactly half of the bits are 1 using a uniformalizer (UFL), which employs bit flipping and adjustment sequence insertion. Blocks of K bits are converted into RLC words, having a length from 1 to 2^K bits, using a variable-length LUT as in Tab. II. Finally, the RLC words are stored in a first-input/first-output (FIFO) buffer, where they are concatenated into a DM word and output with some latency to account for the DM conversion speed. As the RLC codebook is prefix-free, the codewords are uniquely invertible at the receiver by reading the bits from the beginning. The RLC word corresponding to input 111 in Tab. II is chosen to be 00000000 instead of the more natural 0000000, since this makes the DM word length fixed at $(2^K+1)N_{\text{in}}/(2K)$ for all inputs, at the expense of a slightly higher rate loss. More details can be found in [18].

The key element in a hardware implementation of this bitwise DM with RLC is the variable-length LUT. The number of entries (addresses) is significantly smaller than that with a fixed-length LUT to realize the same rate loss. On the other hand, it is known that a variable-length LUT is not straightforward [25], [26]. An available LUT element usually has a fixed bus width at input/output interfaces. Thus how to realize a virtual variable-length LUT with a fixed-length LUT or other available elements is a critical issue for the implementation. According to the exemplified RLC in Tab. II, the bus width at the output interface can be 8 bits. For example, in case that the output length is shorter than 8, arbitrary bits should be padded. The address for writing into the FIFO buffer is updated after writing the current RLC word based on its effective (unpadded) length.

The throughputs for the bitwise DM with RLC in Fig. 1 are K bits at the input interface and $(2^K+1)/2$ bits at the output.

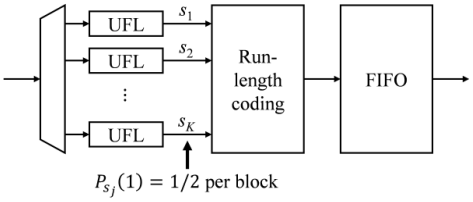


Fig. 1. Schematic for bitwise DM encoding with RLC and periodical uniformalization.

TABLE II
AN EXAMPLE OF RLC ($K = 3$).

Input bits $s_1s_2s_3$	Input symbol	RLC word	Effective RLC word length
000	0	1	1
001	1	01	2
010	2	001	3
011	3	0001	4
100	4	00001	5
101	5	000001	6
110	6	0000001	7
111	7	00000000	8

Relevant values of K are from 3 to 6, so the throughputs range from 3 to 6 bits at the input and from 4.5 to 32.5 bits at the output. In these cases, the bus widths should be 6 and 33 bits at input and output interfaces, resp. Since this is a bit-level DM, m_{sb} instances are needed control the PMF of one quadrature amplitude modulation (QAM) symbol. At a clock frequency of 500 MHz, the minimum throughput at the encoding output is $4.5 \cdot 500 = 2.25$ Gsymbol/s. To achieve a symbol rate of 100 Gsymbol/s, $\lceil 2 \cdot 100/2.25 \rceil m_{\text{sb}} = 89 m_{\text{sb}}$ instances are required for polarization-multiplexed PS-QAM signal generation. At the same condition, the throughput of CCDM [6] is one pulse amplitude modulation (PAM) symbol per clock cycle, i.e., 500 Msymbol/s, and $\lceil 2 \cdot 100/0.5 \rceil = 400$ instances are required. This would be the same for other computation-based symbolwise DMs. If m_{sb} is 4 (i.e., 64-QAM), the required number of instances is comparable between CCDM and bit-level DM with RLC, but CCDM needs complex arithmetic coding [6], [12]. A drawback with bitwise DMs such as this RLC-based scheme is that the obtained PMFs are usually constrained to products of bit probabilities[1], which is not the case for symbolwise DMs.

B. Hierarchical DM

Fig. 2 shows the schematic of HiDM, which is a LUT-based fixed-length-to-fixed-length conversion scheme [12]. The parameters in HiDM are defined in Tab. III. The N_{in} input bits (excluding sign bits) are partitioned and input to LUTs, hierarchically placed in L layers. In the top layer, an LUT receives s_L bits and outputs $u_L = t_{L-1}r_{L-1}$ bits. In layer $\ell = L-1, L-2, \ldots, 2$, each LUT receives r_ℓ bits from layer $\ell+1$ and s_ℓ bits from the input of the DM as information bits. Totally $v_\ell = r_\ell + s_\ell$ bits are converted into $u_\ell = t_{\ell-1}r_{\ell-1}$ bits. These bits are fed into $t_{\ell-1}$ LUTs in layer $\ell-1$, which each receive $r_{\ell-1}$ bits. In layer 1, each LUT receives r_1 bits from layer 2 and s_1 bits from the input of the DM. Totally

[1]If a bitwise DM generates binary DM words with a fixed number of ones, a parallel amplitude architecture [27] can approximate an arbitrary PMF.

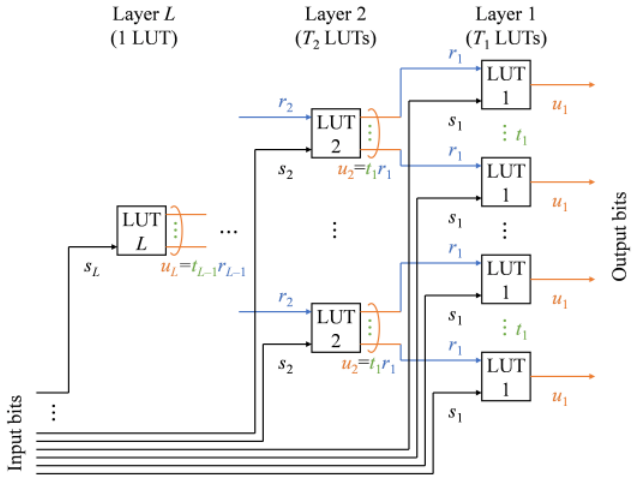


Fig. 2. Schematic of HiDM encoding.

TABLE III
KEY PARAMETERS IN HIDM.

Notation	Description
ℓ	layer index
L	number of layers
t_ℓ	number of LUTs in layer ℓ connected to a single LUT in layer $\ell+1$
T_ℓ	number of LUTs in layer ℓ
r_ℓ	number of input bits in an LUT in layer ℓ from layer $\ell+1$
s_ℓ	number of input information bits in an LUT in layer ℓ
v_ℓ	number of input bits in an LUT in layer ℓ ($r_\ell+s_\ell$)
u_ℓ	number of output bits in an LUT in layer ℓ ($mN_\mathrm{s}/2$ if $\ell=1$ or $t_{\ell-1}r_{\ell-1}$ else)

$v_1 = r_1 + s_1$ bits are converted into u_1 bits, which corresponds to u_1/m_sb QAM symbols. The number of DM output bits and QAM symbols are $m_\mathrm{sb}N_\mathrm{s}/2 = T_1u_1 = (\Pi_{\ell=1}^{L-1}t_\ell)u_1$ and $N_\mathrm{s}/2$, resp.

HiDM has massively parallel input/output interfaces, which are well suited for hardware implementation. The bus widths are $\sum_{\ell=1}^{L} T_\ell s_\ell$ and T_1u_1 at the input and the output of the DM encoding, resp. If for example $L=7$, $T_\ell = 2^{7-\ell}$, $s_1 = \cdots = s_6 = 5$, and $s_7 = 10$, then the bus widths are 640 bits at both the input and the output interfaces. At a clock frequency of 500 MHz, the output throughput is 320 Gb/s or $320/m_\mathrm{sb}$ Gsymbol/s. Under a symbol rate of 100 Gsymbol/s, $\lceil 2\cdot 100/(320/m_\mathrm{sb})\rceil = \lceil (5/8)m_\mathrm{sb}\rceil$ instances are required. Thus, the larger bus width supports around 100 times larger throughput, which requires 100 times fewer instances compared with the RLC-based DM in Sec. II-A.

III. DESIGN AND EVALUATION OF HIDM

A fiber-optic communication channel can be approximated by the AWGN channel with an average power constraint if the dominant impairment is amplified spontaneous emission noise from optical amplifiers. The target PMF for such channel is the discrete Gaussian, or Maxwell–Boltzmann (MB) distribution. For simplicity, the shaped QAM symbols are obtained by combining two shaped PAM symbols. In an example for PS-256-QAM generation [12], the total number of bits per PAM symbol $m/2$ is 4, and both the sign bit (the most significant bit) and the least significant bit are not shaped. Only the second and third significant bits are shaped in each dimension, so that $m_\mathrm{sb}=4$. Tab. IV exemplifies the parameters used. The number of DM input bits per DM word $\sum_{\ell=1}^{L} T_\ell s_\ell$ is 507, and the number of DM output bits per DM word $m_\mathrm{sb}N_\mathrm{s}/2 = T_1u_1$ is 640. Thus the maximum spectral efficiency per 2D symbol is $\beta = 2(1+mN_\mathrm{in}/N_\mathrm{s}) = 2(2+507/320) = 7.169$ bit per channel use (bpcu). The entropy of a 2D symbol $2H(X)$ will be larger than β, where X denotes a PAM symbol.

TABLE IV
CHOSEN PARAMETERS USED IN [12, TAB. IV, FIG. 4].

ℓ	t_ℓ	T_ℓ	r_ℓ	s_ℓ	v_ℓ	u_ℓ
7		1		5	5	12
6	2	2	6	5	11	12
5	2	4	6	5	11	12
4	2	8	6	5	11	12
3	2	16	6	5	11	12
2	2	32	6	5	11	12
1	2	64	6	3	9	10

The values of T_ℓ, v_ℓ, and u_ℓ determine the accumulated size of the LUTs, i.e., $\sum_{\ell=1}^{L} T_\ell 2^{v_\ell} u_\ell$ for DM. If a simple mirror structure is employed for the invDM, its size will be[2] $\sum_{\ell=1}^{L} T_\ell 2^{u_\ell} v_\ell$. Thus, there would be practical constraints on the values of v_ℓ and u_ℓ, which depend on the acceptable hardware resource usage. Under such constraints, a binary tree structure ($t_\ell = 2$, $\forall \ell$) gives the best shaping performance.

The LUT contents are determined from layer 1 and up. There are 2^{u_1} output-word candidates for LUT1, of which 2^{v_1} should be selected based on some criterion, e.g., minimum average symbol energy $E = \mathbb{E}[X^2]$. Thus, the output-word candidates are sorted by increasing E, assuming a Gray-mapped PAM constellation. The top 2^{v_1} candidates are selected, and assigned input symbols in natural order (i.e., $0\cdots00$, $0\cdots01$, …, $1\cdots11$), with $0\cdots00$ assigned to the word with the smallest E. The process then continues with layers $\ell = 2,\ldots,L$. There are 2^{u_ℓ} output-word candidates for the level-ℓ LUT. For each candidate, E is computed based on the selected contents for LUTs $1,\ldots,\ell-1$. The output words are again sorted by increasing E, and the top 2^{v_2} are selected.

We generated PS-256-QAM signals having a DM word length of 320 16-PAM symbols by employing CCDM [6] and HiDM [12]. The target PMF for CCDM was set to the MB distribution with $\beta = 7.169$ bpcu. For HiDM, the scheme exemplified above and in Tab. IV was used, which also has $\beta = 7.169$ bpcu. In Tab. V [12], some key statistics of the shaped PAM symbols X and QAM symbols X_c are summarized, viz. the PMF P_X, average QAM symbol energy $E_\mathrm{c} = \mathbb{E}[|X_\mathrm{c}|^2]$, QAM symbol entropy $H(X_\mathrm{c}) = 2H(X)$, maximum spectral efficiency β, rate loss $R_\mathrm{loss} = H(X_\mathrm{c}) - \beta$, and constellation gain $G = (2^\beta - 1)d_\mathrm{min}^2/(6E_\mathrm{c})$, where d_min denotes the minimum Euclidean distance. The rate loss of a QAM symbol was 0.07 and 0.08 bpcu for CCDM and HiDM, resp. In each case, the constellation gain G was more than 1 dB, while G is 0 dB for uniform square QAM. The gap in G from the ideal Maxwell–Boltzmann (MB) distribution was within 0.4 dB even though we did not shape the least

[2]There may exist techniques to reduce the LUT size without sacrificing performance.

TABLE V
STATISTICS OF THE SHAPED SYMBOLS [12].

	CCDM	HiDM	MB
N_s (PAM symbols)	320	320	–
$P_{\lvert X\rvert}(1)$	0.2453	0.2376	0.2628
$P_{\lvert X\rvert}(3)$	0.2453	0.2376	0.2355
$P_{\lvert X\rvert}(5)$	0.1625	0.1684	0.1891
$P_{\lvert X\rvert}(7)$	0.1625	0.1684	0.1360
$P_{\lvert X\rvert}(9)$	0.0719	0.0757	0.0877
$P_{\lvert X\rvert}(11)$	0.0719	0.0757	0.0506
$P_{\lvert X\rvert}(13)$	0.0203	0.0183	0.0262
$P_{\lvert X\rvert}(15)$	0.0203	0.0183	0.0121
E_c	74.00	74.70	68.31
$H(X_c)$ (bpcu)	7.242	7.252	7.169
β (bpcu)	7.169	7.169	7.169
R_{loss} (bpcu)	0.073	0.083	0
G (dB)	1.097	1.056	1.444

significant bit.[3]

IV. FLEXIBLE TUNING OF TWO-DIMENSIONAL PMFS

In fiber-optic links where AWGN is not the dominant impairment, different PMFs than MB can give better performance. The received signal-to-noise ratio (SNR) after propagation through a nonlinear fiber-optic channel depends on the transmitted PMF, especially for short links with negligible chromatic and polarization-mode dispersion, where the transmitted waveform shape is maintained. The nonlinear self-channel interference increases with the excess kurtosis $\Phi = \mathbb{E}[|X_c|^4]/\mathbb{E}^2[|X_c|^2] - 2$ [28], [29] of the QAM symbols X_c, which equals 0 for a Gaussian distribution.

HiDM can shape PMFs in an arbitrary number of dimensions as long as the complexity is acceptable. Here, we improve the tolerance to fiber nonlinearity by two-dimensional shaping using HiDM. The LUTs are designed as in Sec. III, except that the output-word candidates are sorted by increasing $\mathbb{E}\left[|X_c|^{F/2}\right]$, for some $F = 1, 2, \ldots, 8$, to reduce Φ.

As in [24], the base constellation is 64-QAM, 32-QAM, or a 1:1 hybrid of 16-QAM and 32-QAM. The target number of coded bits per QAM symbol is $m = 4.25$ bpcu. The assumed FEC code rate is $5/6$, so the target information rate is 3.542 bpcu. When we employ 64-QAM, 32-QAM, or hybrid 32/16-QAM, the FEC throughput increases by $6/4.25 - 1 = 41.1\%$, $5/4.25 - 1 = 17.6\%$, or $44.5/4.25 - 1 = 5.9\%$, resp., compared with uniform signaling. Fig. 3 shows the PMFs generated by HiDM for different F values. The PMFs for the linear AWGN channel have a relatively high peak at small amplitudes, wheras the the PMFs for nonlinear channels are more uniform.

Fig. 4 illustrates the tradeoff between linear and nonlinear performance. The horizontal axis shows the normalized generalized mutual information (NGMI) [30], [31] or asymmetric information (ASI) [32], [33] with matched bit-metric decoding [4] over the AWGN channel with an SNR of 12 dB. The vertical axis shows the excess kurtosis Φ, which approximately quantifies the nonlinear interference. PS-64-QAM and PS-32-QAM show comparable linear performance. They are almost

[3]If the least significant bit is shaped, the energy gap will be reduced to less than 0.3 dB.

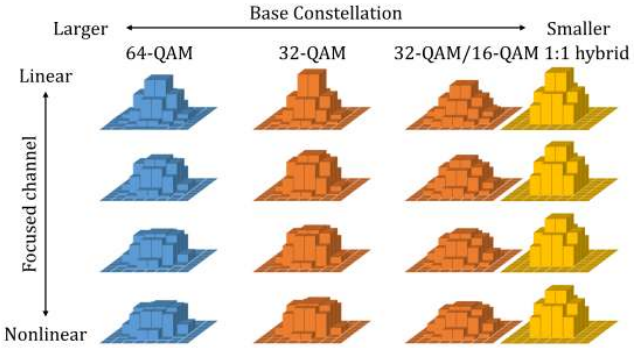


Fig. 3. Two-dimensional PMFs obtained by HiDM for (from top to bottom) $F = 2, 4, 6, 8$.

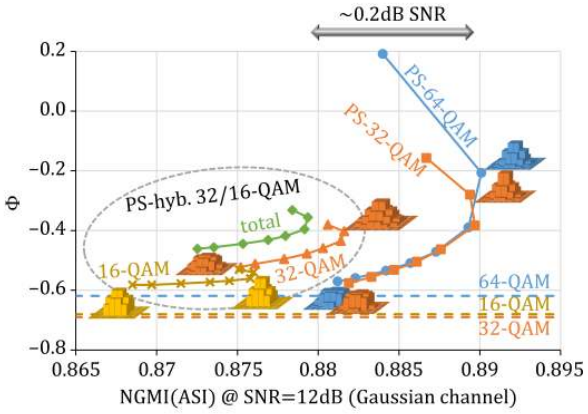


Fig. 4. Tradeoff between linear and nonlinear performance. The top marker on each curve corresponds to $F = 1$ and the bottom one to $F = 8$.

0.2 dB better in terms of required SNR than hybrid PS-32/16-QAM. When the DM word lists in the LUTs are adapted to nonlinear fiber-optic channels by increasing F, the linear performance degrades, but the nonlinear performance improves. For PS-64-QAM and PS-32-QAM with high F, Φ becomes comparable with uniform 64-QAM, at the expense of a linear performance reduction of about 0.2 dB.

The best linear performance was observed at $F = 2$, with minimum E_c, for PS-64-QAM and PS-hybrid-32/16-QAM, and $F = 3$ for PS-32-QAM. To investigate why the best F is 3 for PS-32-QAM, we computed the rate loss as shown in Fig. 5. The rate loss decreases as F increases, i.e., as Φ decreases, for PS-64-QAM and PS-32-QAM. The balance between E_c and rate loss causes the peculiar performance of PS-32-QAM in Fig. 4.

V. SUMMARY

Some aspects of low-complexity implementations of DM PS in fiber-optic communications were studied, in terms of throughput, bus width, and circuit area. HiDM has around 100 times larger throughput than a DM with RLC. A large-scale HiDM example was given in detail, realizing a DM word length of 160 256-QAM symbols using a 7-layer LUT hierarchy. The resulting energy gap from the ideal MB distribution is less than 0.4 dB, while keeping four bits per QAM symbol uniformly distributed (nonshaped). A simple method

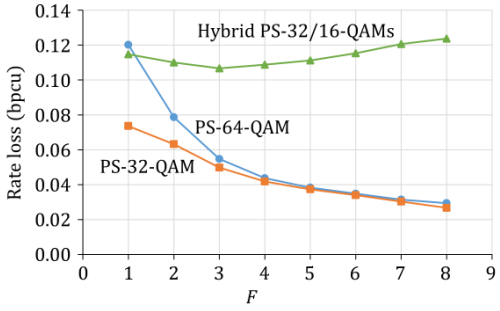


Fig. 5. Rate loss as a function of the sorting parameter F.

to flexibly adapt HiDM to nonlinear channels with granular base constellations was also described.

As shown in [34], this HiDM is useful also for a rudimentary form of joint source–channel coding. This realizes simultaneous data compression and probabilistic shaping, which can further reduce the required SNR or system power consumption in future optical networks. Thanks to the high throughput by massively parallel interfaces, both the encoding and decoding of HiDM, including joint source–channel coding, was implemented in a single field-programmable gate array chip [35].

Acknowledgment

We thank Koji Igarashi of Osaka University for fruitful discussions.

References

[1] G. D. Forney, Jr. and L.-F. Wei, "Multidimensional constellations—Part I: introduction, figure of merit, and generalized cross constellation," *IEEE J. Selected Areas Commun.*, vol. 7, no. 6, pp. 877–892, Aug. 1989.

[2] A. R. Calderbank and L. H. Ozarow, "Nonequiprobable signaling on the Gaussian channel," *IEEE Trans. Inf. Theory*, vol. 36, no. 4, pp. 726–740, July 1990.

[3] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 913–929, May 1993.

[4] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4651–4665, Dec. 2015.

[5] F. Buchali, F. Steiner, G. Böcherer, L. Schmalen, P. Schulte, and W. Idler, "Rate adaptation and reach increase by probabilistically shaped 64-QAM: an experimental demonstration," *J. Lightw. Technol.*, vol. 34, no. 7, pp. 1599–1609, Apr. 2016.

[6] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 430–434, Jan. 2016.

[7] J. Cho, S. Chandrasekhar, R. Dar, and P. J. Winzer, "Low-complexity shaping for enhanced nonlinearity tolerance," in *Proc. Eur. Conf. on Opt. Comm. (ECOC)*, Düsseldorf, Germany, Sep. 2016, p. W1C.2.

[8] G. Böcherer, F. Steiner, and P. Schulte, "Fast probabilistic shaping implementation for long-haul fiber-optic communication systems," in *Proc. Eur. Conf. on Opt. Comm. (ECOC)*, Göteborg, Sweden, Sep. 2017, p. Tu.2.D.3.

[9] F. Steiner, P. Schulte, and G. Böcherer. "Approaching waterfilling capacity of parallel channels by higher order modulation and probabilistic amplitude shaping," in *Proc. 52nd Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, Mar. 2018.

[10] T. Yoshida, M. Karlsson, and E. Agrell, "Technologies toward implementation of probabilistic constellation shaping," in *Proc. Eur. Conf. on Opt. Comm. (ECOC)*, Roma, Italy, Sep. 2018, p. Th.1.H.1.

[11] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Multiset-partition distribution matching," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 1885–1893, Mar. 2019.

[12] T. Yoshida, M. Karlsson, and E. Agrell, "Hierarchical distribution matching for probabilistically shaped coded modulation," *J. Lightw. Technol.*, vol. 37, no. 6, pp. 1579–1589, Mar. 2019.

[13] P. Schulte and F. Steiner, "Divergence-optimal fixed-to-fixed length distribution matching with shell mapping," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 620–623, Apr. 2019.

[14] J. Cho and P. J. Winzer, "Multi-rate prefix-free code distribution matching ," in *Proc. Opt. Fib. Commun. Conf. (OFC)*, San Diego, CA, USA, Mar. 2019, Paper M4B.7.

[15] Y. C. Gültekin, W. J. van Houtum, A. Koppelaar, and F. M. J. Willems, "Enumerative sphere shaping for wireless communications with short packets," *IEEE Trans. Wireless Commun.*, to appear, 2020.

[16] J. Cho, "Prefix-free code distribution matching for probabilistic constellation shaping," *IEEE Trans. Commun.*, to appear, 2020.

[17] T. Yoshida, M. Karlsson, and E. Agrell, "Short-block-length shaping by simple mark ratio controllers for granular and wide-range spectral efficiencies," in *Proc. Eur. Conf. on Opt. Comm. (ECOC)*, Göteborg, Sweden, Sep. 2017, p. Tu.2.D.2.

[18] T. Yoshida, M. Karlsson, and E. Agrell, "Low-complexity variable-length output distribution matching with periodical distribution uniformalization," in *Proc. Opt. Fib. Commun. Conf. (OFC)*, San Diego, CA, USA, Mar. 2018, p. M.4.E.2.

[19] Y. Koganei, K. Sugitani, H. Nakashima, and T. Hoshida, "Optimum bit-level distribution matching with at most $O(N^3)$ implementation complexity," in *Proc. Opt. Fib. Commun. Conf. (OFC)*, San Diego, CA, USA, Mar. 2019, Paper M4B.4.

[20] M. Pikus and W. Xu, "Bit-level probabilistically shaped coded modulation,", *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 1929–1932, Sep. 2017.

[21] G. Böcherer, F. Steiner, and P. Schulte, "High throughput probabilistic shaping with product distribution matching," [Online]. Available: www.arxiv.org/abs/1702.07510

[22] T. V. Ramabadran, "A coding scheme for *m*-out-of-*n* codes," *IEEE Trans. Commun.*, vol. 38, no. 8, pp. 1156–1163, Aug. 1990.

[23] S. Civelli and M. Secondini, "Hierarchical distribution matching: a versatile tool for probabilistic shaping," [Online]. Available: www.arxiv.org/abs/1911.08243

[24] T. Yoshida and N. Suzuki, "Flexible and low-power probabilistic shaping for fiber-optic communications," in *Proc. Signal Processing in Photonic Communications (SPPCom)*, Burlingame, CA, USA, July 2019, Paper SpT3E.2.

[25] Z. Aspar, Z. M. Yusof, and I. Suleiman, "Parallel Huffman decoder with an optimized look up table option on FPGA," in *2000 TENCON Proc. Intelligent Systems and Technologies for the New Millennium (Cat. No.00CH37119)*, vol. 1, pp. 73–76, 2000.

[26] H. J. Huang, C.-H. Fang, and C.-P. Fan, "Very large scale integration design of a low power and cost effective context based adaptive variable length coding decoder for H. 264/AVC portable applications," IET Image Processing, vol. 6, no. 2, pp. 104–114, 2012.

[27] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Parallel-amplitude architecture and subset ranking for fast distribution matching," [Online]. Available: arxiv.org/abs/1902.08556

[28] A. Carena, G. Bosco, V. Curri, Y. Jiang, P. Poggiolini, and F. Forghieri, "EGN model of non-linear fiber propagation," *Opt. Express*, vol. 22, no. 13, pp. 16335–16362, June 2014.

[29] E. Sillekens, D. Semrau, G. Liga, N. A. Shevchenko, Z. Li, A. Alvarado, P. Bayvel, R. I. Killey, and D. Lavery, "A simple nonlinearity-tailored probabilistic shaping distribution for square QAM," in *Proc. Opt. Fib. Commun. Conf. (OFC)*, San Diego, CA, USA, Mar. 2018, Paper M3C.4.

[30] J. Cho, L. Schmalen, and P. Winzer, "Normalized generalized mutual information as a forward error correction threshold for probabilistically shaped QAM," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Göteborg, Sweden, Sep. 2017, Paper M.2.D.2.

[31] J. Cho and P. J. Winzer, "Probabilistic constellation shaping for optical fiber communications," *IEEE/OSA J. Lightw. Technol.*, vol. 37, no. 6, pp. 1590–1607, Mar. 2019.

[32] T. Yoshida, M. Karlsson, and E. Agrell, "Performance metrics for systems with soft-decision FEC and probabilistic shaping," *IEEE Photon. Technol. Lett.*, vol. 29, no. 23, pp. 2111–2114, Dec. 2017.

[33] T. Yoshida, A. Alvarado, M. Karlsson, and E. Agrell, "Post-FEC BER prediction for bit-interleaved coded modulation with probabilistic shaping," [Online]. Available: arxiv.org/abs/1911.01585

[34] T. Yoshida, M. Karlsson, and E. Agrell, "Joint source-channel coding via compressed distribution matching in fiber-optic communications," in *Proc. Opt. Fib. Commun. Conf. (OFC)*, San Diego, CA, USA, Mar. 2019, Paper M4B.6.

[35] T. Yoshida, M. Binkai, S. Koshikawa, S. Chikamori, K. Matsuda, N. Suzuki, M. Karlsson, and E. Agrell, "FPGA implementation of distribution matching and dematching," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Dublin, Ireland, Sep. 2019, Paper M.2.D.2.

Prefix-Free Code Distribution Matching for 5G New Radio

Junho Cho and Ori Shental
Nokia Bell Labs
Holmdel, NJ 07733, USA
Email: {junho.cho, ori.shental}@nokia-bell-labs.com

Abstract—**We use prefix-free code distribution matching (PCDM) for rate matching (RM) in some 5G New Radio (NR) deployment scenarios, realizing a wide range of information rates from 1.4 to 6.0 bit/symbol in fine granularity of 0.2 bit/symbol. We study the performance and implementation of the PCDM-based RM, in comparison with the low-density parity-check (LDPC)-based RM, as defined in the 5G NR standard. Simulations in the additive white Gaussian noise channel show that up to 2.16 dB gain in the signal-to-noise ratio can be obtained with the PCDM-based RM at a block error rate of 10^{-2} when compared to LDPC-based RM in the tested scenarios, potentially at a smaller hardware cost.**

I. Introduction

In the 5th Generation (5G) New Radio (NR) mobile broadband standard, low-density parity-check (LDPC) codes have been adopted as the channel coding scheme for user data, as recently specified in the 3rd Generation Partnership Project (3GPP) Release 15 [1]. A notable feature of the 5G NR LDPC codes is the great flexibility to support a wide range of information block lengths K_C, ranging from 40 to 8448 bits, and various code rates, ranging from 1/5 to 8/9 [2]–[4]. This ensures reliable transmission of user data in dynamically varying cellular channel conditions, and in various deployment scenarios where different amount of radio and hardware resources is available.

Among the many available 5G NR LDPC code parameters, finding a set of parameters to maximize the information throughput under given channel conditions and resources is a task of *rate matching* (RM). The 5G NR standard performs RM in two steps: first, coarse-grained RM chooses one of the two base graphs (BGs) and a submatrix size to lift the BG, then fine-grained RM shortens and punctures parts of the derived code in single-bit granularity. There are 51 different submatrix sizes Z_C defined in the standard, in the form of $Z_C = A \times 2^j$ for $A \in \{2, 3, 5, 7, 9, 11, 13, 15\}$ and $j = 0,1, \ldots$, within the range $2 \leq Z_C \leq 384$. Transmission begins with a high-rate LDPC code first, and in case the decoding fails at the receiver, incremental-redundancy hybrid automatic repeat request (HARQ) is operated such that more parity bits are transmitted for the same data until the decoding succeeds. The BGs of the 5G NR LDPC codes are made to have a special structure such that a high-rate code is always a submatrix of a lower-rate code, in order to facilitate the incremental-redundancy HARQ. Overall, the coarse- and fine-grained RM with incremental-redundancy HARQ make the number of all possible codes in an order of thousands.

Although essential to support the broad 5G NR deployment scenarios, the sheer number of LDPC codes poses a significant challenge in hardware implementation. In [5], for example, it was shown that a flexible decoder for only 12 LDPC codes (defined in the WiFi standard IEEE 802.11n/ac, with 3 different submatrix sizes and 4 code rates) consumes about 2.2× larger area than an inflexible decoder for a single code for the same throughput, when implemented on a field-programmable gate array (FPGA). In particular, multiple submatrix sizes add a greater implementation complexity than multiple code lengths, due to the intricacy associated with the design of a routing network [5]. It is therefore a daunting task to implement the whole set of 5G NR LDPC codes with as many as 51 different submatrix sizes. Moreover, this flexible coding scheme should attain up to 20 Gb/s of the downlink throughput, as required by the standard.

While RM for user data is almost solely performed by LDPC in the 5G NR standard, recent optical communication systems use *probabilistic constellation shaping* (PCS) for RM, in conjunction with a single or a few forward error correction (FEC) codes [6]. PCS shapes the probability distribution of modulation symbols such that symbols with a low energy are sent more frequently than those with a high energy, thereby reducing the average symbol energy. This implies an increased Euclidean distance between modulation symbols for the same transmit power, hence probabilistically-shaped symbols better resist the channel impairments than symbols with uniform probability distribution. Since a non-uniform distribution has a lower entropy than the uniform distribution over the same support, PCS can intrinsically adjust the information rate (IR), i.e., it can realize RM. In optical communications, PCS-based RM served as a key technology to obtain record-high spectral-efficiency transmission results in recent experiments and field trials, which led to rapid adoption in the commercial sector [6].

Motivated by the remarkable success of PCS in optical communications, we study in this work the application of PCS to mobile broadband services. We realize PCS in the probabilistic amplitude shaping (PAS) architecture [7] using *prefix-free code distribution matching* (PCDM) [8]. By transferring the role of RM to PCDM, while only a small subset of the 5G NR LDPC codes is used for FEC, we demonstrate up to 2.16 dB gain in the signal-to-noise ratio (SNR) for the same IR, at a block error rate (BLER) of 10^{-2}. Importantly, this SNR gain may be achieved at a smaller hardware cost than the conventional LDPC-based RM, as recently shown by an FPGA implementation in optical communications scenarios [9].

TABLE I
RATE MATCHING WITH 5G LDPC CODES [1] OF LENGTH $N_C = 600$

QAM	BG	Z_C	K_C	IR	QAM	BG	Z_C	K_C	IR	QAM	BG	Z_C	K_C	IR
16	2	28	210	1.4	64	2	36	280	2.8	256	2	44	330	4.4
16	2	30	240	1.6	64	2	40	300	3.0	256	2	44	345	4.6
16	2	36	270	1.8	64	2	40	320	3.2	256	2	48	360	4.8
16	2	40	300	2.0	64	2	44	340	3.4	256	2	48	375	5.0
16	2	44	330	2.2	64	2	48	360	3.6	256	2	52	390	5.2
16	2	48	360	2.4	64	2	48	380	3.8	256	1	20	405	5.4
16	2	52	390	2.6	64	2	52	400	4.0	256	1	20	420	5.6
16	1	20	420	2.8	64	1	20	420	4.2	256	1	20	435	5.8
16	1	22	450	3.0	64	1	20	440	4.4	256	1	22	450	6.0
16	1	22	480	3.2	64	1	22	460	4.6					
16	1	24	510	3.4	64	1	22	480	4.8					
					64	1	24	500	5.0					

II. RATE MATCHING WITH 5G NR LDPC

When a rate-R_C LDPC code is used with M^2-ary quadrature amplitude modulation (QAM) for $M^2 \in \{4, 16, 64, 256\}$, as specified in the 5G NR standard, the achievable IR of the system is given by

$$R_{Info} = 2mR_c \tag{1}$$

in bit/symbol, where $m \coloneqq \log_2 M$. This IR is said to be achieved *if* the decoding is error-free. For the 5G NR LDPC codes with incremental-redundancy HARQ, error-free decoding needs not be ensured in every transmission block, but rather a marginally low BLER (typically within the range of 10^{-3} to 10^{-1}) is set as the target error performance to avoid too frequent retransmission. In this case, RM involves finding a code-modulation pair that produces the greatest R_{Info} among all pairs defined in the standard such that the target BLER is fulfilled under the given channel condition. Also engaged in RM are the available radio and hardware resources in hand, and the practical requirements such as latency.

In this work, three sets of codes are selected from the 5G NR LDPC codes to produce IRs ranging from 1.4 to 6.0 bit/symbol in 0.2 bit/symbol increments to cover a wide range of channel conditions. Each set of codes has a fixed code length $N_C \in \{600, 1200, 4800\}$, which deals with a scenario with few to many resources, incurring short to long latency. For example, the extensive set of codes defined in the current 5G NR standard for the case of $N_C = 600$ is shown in Table I, where 32 different codes with 10 different submatrix sizes Z_C are needed to realize the target IRs, with three different QAM orders. To support all three N_C for the target IRs, 96 different LDPC codes are needed in total, with 27 different submatrix sizes.

III. RATE MATCHING WITH PCDM

A. PCDM

An essential component of PCS realized using the PAS architecture is the distribution matching (DM), which receives binary information bits of equal probabilities and produces modulation symbols of a target probability distribution. The transmitter of a PCS system, in the PAS architecture [7], first synthesizes a target distribution of positive real symbols using a DM, as shown in Fig. 1, then the binary representation of the positive real symbols is encoded by a binary *systematic* FEC code. The parity bits are then used as sign bits to produce real

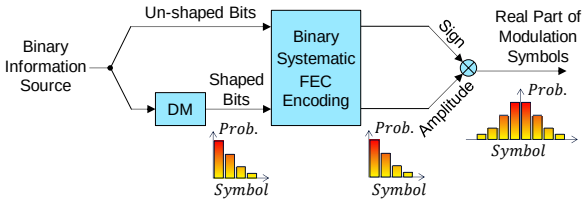


Fig. 1. PCS based on the PAS architecture [7].

TABLE II
PCDM CODE $\mathcal{C}_2$

Input Bits	Output Symbols
0	111111
100	113
1010	111113
1011	11113
1100	1113
1101	1311
1110	3111
111100	133
111101	3113
1111100	1313
1111101	3131
1111110	3311
11111110	3133
111111110	3313
111111111	3331

symbols that are symmetrically distributed around zero, while the systematic information bits preserve the symbol-domain probability distribution made by the DM. At the receiver side, as long as the FEC decoding recovers error-free systematic bits, the DM operation can be undone without error.

PCDM is a method to implement DM by using *prefix-free codes* (often called *Huffman codes* [10, Ch. 5.6]). As shown in Table II, a PCDM code is constructed by concatenating two prefix-free codes, namely, binary prefix-free codewords in the left entries and non-binary (including binary) codewords in the right entries of a look-up table (LUT) in an order. A PCDM encoder reads information *bits* in a bit-by-bit manner until the first (hence shortest) matching bit sequence is found from the left entries of the LUT, then instantaneously produces a *symbol* sequence in the corresponding right entry. This variable-length bit-to-symbol encoding is repeated in an iterative manner, where each iteration starts from the first bit in the bit stream that has not been encoded yet. For example, the code in Table II (denoted by $\mathcal{C}_2$ throughout the paper) encodes a bit stream "0 1100…" into the symbol stream "111111 1113…" Note that the right entries of $\mathcal{C}_2$ contain only the positive real part of complex-valued 16-QAM symbols $X + iY$ for $X, Y \in \{\pm 1, \pm 3\}$, which simplifies the description and implementation. The negative real part of the symbols can be produced by using the symmetry of a probabilistic distribution around zero, as typically done in PCS systems, allowing one more information bit to be encoded as a *sign* bit in a symmetrically distributed real symbol. Generating the imaginary component is trivial; we can, for instance, use the real symbols alternately for real and imaginary components of a complex-valued QAM symbol (this approach is taken in this work). PCDM *decoding* can be described in the same manner as PCDM encoding, by changing only the role of bits and symbols, thus the details of the decoding process are omitted.

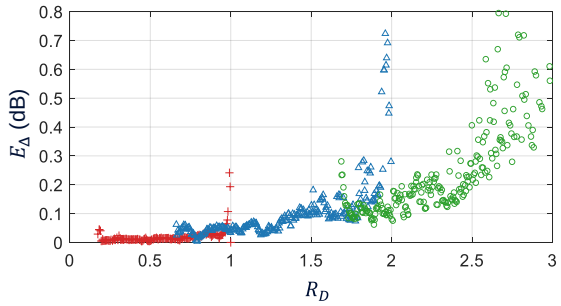


Fig. 2. Performance of PCDM codes of cardinality $|\mathcal{C}| = 24$, with real positive symbols for 16-ary (pluses), 64-ary (triangles), and 256-ary (circles) QAM.

The amount of information bits carried by each DM output symbol, called the *DM rate*, denoted by R_D (in bits per positive real symbol), and the average energy E of output symbols of PCDM can be easily calculated from the LUT in the limit of encoding iterations [8], assuming independent and identically distributed (IID) information bits with equal probabilities. For example, the code $\mathcal{C}_2$ realizes $R_D \approx 0.504$ with $E \approx 1.904$ asymptotically. The performance of a DM can be quantified by the *energy gap* defined as $E_\Delta := E/E^*$, which evaluates the additional energy consumed by the DM relative to the theoretical lower limit of energy E^* to achieve the same rate R_D. The limit E^* to achieve R_D is given by the average output energy of a stationary ergodic random process that generates letter X from the same alphabet as the PCDM code, where X follows the IID Maxwell-Boltzmann distribution [11] and produces entropy $H(X) = R_D$. The problem of constructing a good PCDM code is then to find a code that produces the smallest average energy E (hence smallest E_Δ) among all possible codes subject to the rate constraint $R_D \geq R_D^*$, with R_D^* being the target DM rate. If we restrict the cardinality of PCDM codes (i.e., the number of rows in the LUTs) to a small number, good PCDM codes can be found by exhaustive or algorithmic search [8]. $\mathcal{C}_2$ has indeed been found in such a way, and its E_Δ is only ~0.03 dB. Note that, as per the aforementioned method in constructing a complex-valued symbol from positive real symbols, rate-R_D PCDM yields $1 + R_D$ information bits per real symbol, and $2(1 + R_D)$ information bits per complex-valued symbol.

The PCDM procedure is, however, not compliant with the 5G NR standard in its current form, since it produces variable-length output at each iteration; i.e., it cannot realize fixed-rate transmission in a block-by-block manner as required by the standard. This compliance issue can be circumvented in the following manner. Namely, we use the *framing* method presented in [8], which switches the encoding method from PCDM to typical bit-to-symbol mapping for uniform QAM during the successive process. The switching position is dynamically determined from the input bit values, such that the given fixed-length bit block can be contained in a fixed-length symbol block. Framing slightly increases E_Δ in general; and the shorter the block length, the more E_Δ increases (see [8] for details). For example, the code $\mathcal{C}_2$ with $R_D \approx 0.504$ and $E \approx 1.904$ can be framed to encode an input block of length $K_D = 150$ bits in an output block of length $N_D = 300$ positive real symbols, to realize a fixed $R_D = 0.5$ in each block with a little greater average symbol energy than 1.904.

There are other known DM methods such as the constant composition DM (CCDM) [12], shell mapping (SM) [13], and enumerative sphere shaping (ESS) [14]. The CCDM needs multiplications and divisions at each iteration, making its hardware implementation very costly. The complexity of SM and ESS is much lower than CCDM, but increases with the block length. Furthermore, due to the inherently limited parallelism [14, Table 3], it is unclear if the CCDM, SM, or ESS can support 20 Gb/s of downlink throughput. There are no published papers on hardware implementation of these methods to date. On the other hand, PCDM has a low complexity, independent of the block length, and was proven through an FPGA implementation to achieve a high throughput with a massive parallelism [9], as will be discussed in Sec. III-C in more detail.

B. *Rate Matching with PCDM*

We first note that the PCDM is characterized by the input and output block lengths K_D and N_D, respectively, realizing the DM rate $R_D = K_D/N_D$ in each block, as if an LDPC code of input and output block lengths K_C and N_C, respectively, realizes the code rate $R_C = K_C/N_C$ in each block. This already illustrates that PCDM can be used for RM, instead of the LDPC. With reference to the PAS architecture in Fig. 1, it can easily be seen that the IR of a PCS system with rate-R_D DM and rate-R_C coding is given by

$$R_{Info} = 2[1 + R_D - m(1 - R_C)] \tag{2}$$

(see [6], [7] for details). As a matter of fact, this shows exactly how the IR can be varied by adjusting either R_D or R_C.

In order to perform RM with PCDM, we construct PCDM codes $\mathcal{C}$ for various R_D ranging from 0.2 to 3.0 bits per positive real symbol, under the cardinality constraint $|\mathcal{C}| = 24$. There exist an enormous number of PCDM codes even with this small cardinality of 24, since the number of possible codes grows exponentially with the cardinality; e.g., for positive real symbols of 16-QAM, more than 3.4×10^{11} different cardinality-24 codes can be constructed. Among all possible codes, the performance of the PCDM codes that have the smallest E_Δ in each R_D bin of width 0.005 is shown in Fig. 2, where small E_Δ below 0.4 dB are observed across a wide range of R_D.

To compare PCDM- and LDPC-based RM in the considered 5G deployment scenarios, we realize the same IRs as in Sec. II using PCDM codes, in conjunction with much fewer LDPC codes than in Table I. Fixed-length framing is applied to PCDM such that each PCDM output block is mapped to exactly one LDPC code of length $N_C \in \{600, 1200, 4800\}$. This is achieved by making the PCDM output block length N_D equal to N_C/m for a given N_C. The PCDM input block length K_D is then determined to meet the target IR according to (2). Shown in Table III are such determined PCDM parameters for the case of $N_C = 600$, made to be compatible with the LDPC-based RM scenario of Table I. We use 28 PCDM codes and 3 LDPC codes of 3 different submatrix sizes in Table III, one LDPC code for each QAM order (cf. top of Table III). Note, however, that it is

TABLE III
RATE MATCHING WITH PCDM CODES AND 5G NR LDPC CODES OF LENGTH $N_C = 600$

BG = 1, $Z_C = 20$, $N_C = 600$, $K_C = 420$

QAM	N_D	K_D	IR
16	300	90	1.4
16	300	120	1.6
16	300	150	1.8
16	300	180	2.0
16	300	210	2.2
16	300	240	2.4
16	300	270	2.6

BG = 1, $Z_C = 22$, $N_C = 600$, $K_C = 480$

QAM	N_D	K_D	IR
64	200	180	2.6
64	200	200	2.8
64	200	220	3.0
64	200	240	3.2
64	200	260	3.4
64	200	280	3.6
64	200	300	3.8
64	200	320	4.0
64	200	340	4.2
64	200	360	4.4
64	200	380	4.6

BG = 1, $Z_C = 24$, $N_C = 600$, $K_C = 510$

QAM	N_D	K_D	IR
256	150	255	4.2
256	150	270	4.4
256	150	285	4.6
256	150	300	4.8
256	150	315	5.0
256	150	330	5.2
256	150	345	5.4
256	150	360	5.6
256	150	375	5.8
256	150	390	6.0

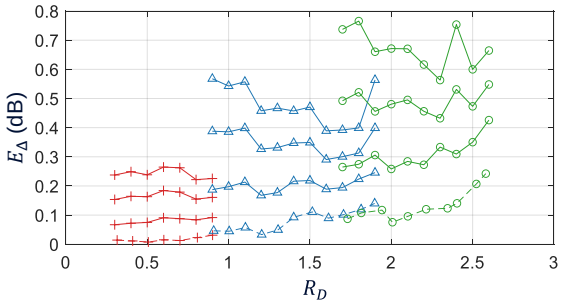


Fig. 3. Performance of PCDM codes $\mathcal{C}$ that realize $R_{Info} = 1.4, 1.6, \ldots, 6.0$ with cardinality $|\mathcal{C}| = 24$ using real positive symbols of 16-ary (pluses), 64-ary (triangles), and 256-ary (circles) QAM. The PCDM is compatible with 5G NR LDPC codes of lengths $N_C = 600$ (upper solid lines), 1200 (middle sold lines), and 4800 (lower solid lines). Also shown is the performance without a fixed-length constraint (dashed lines).

possible to use only one submatrix size Z_C for all QAM orders, if we design new LDPC codes by taking PCDM into account. More importantly, to further support the other code lengths $N_C \in \{1200, 4800\}$, we need more LDPC codes but *no* more PCDM codes, since the set of PCDM codes for $N_C = 600$ can be used for an *arbitrary* integer N_C with mere change in the framing constraint, causing virtually no additional hardware cost.

Fig. 3 shows the performance of 28 PCDM codes chosen from Fig. 2, which produce the target IRs under the framing constraints to comply with the LDPC codes of $N_c \in \{600, 1200, 4800\}$. With large PCDM block lengths N_D compatible with $N_c = 4800$ (markers connected by lower solid lines), energy gap of approximately 0.1 to 0.4 dB is achieved. The energy gap increases as N_D decreases, reaching almost 0.8 dB for the case of $N_c = 600$ and 256-QAM. This large gap is attributed partly to the fixed-length constraint, and partly to the cardinality constraint that becomes more prominent as the QAM order grows. However, as will be shown in Sec. IV, PCDM-based RM provides significant SNR gain even with 0.8 dB of the energy gap.

C. Implementation Aspects of PCDM-based RM

Table IV summarizes the implementations required for LDPC- and PCDM-based RM in the 5G NR deployment scenarios with $N_C \in \{600, 1200, 4800\}$, where the numbers in

TABLE IV
IMPLEMENTATIONS REQUIRED FOR LDPC- AND PCDM-BASED RM WITH CODE LENGTH $N_C = 600, 1200, 4800$

		$N_C = 600$	$N_C = 1200$	$N_C = 4800$	Total
LDPC-Based RM	# LDPC submatrix sizes	10	7	10	27
PCDM-Based RM	# LDPC submatrix sizes	3 (1)	3 (1)	3 (1)	9 (3)
	# PCDM codes	28	28	28	28

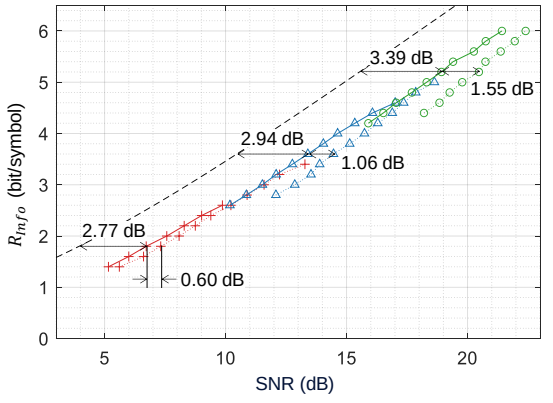
the parentheses show the possible numbers if a new LDPC design criterion is applied. The PCDM-based RM uses 28 PCDM codes and 9 (3) LDPC codes of 9 (3) different submatrix sizes in total, whereas the LDPC-based RM uses 96 LDPC codes of 27 different submatrix sizes. A universal PCDM architecture is presented in [9] that can support all the 28 PCDM codes of Table IV. In this universal architecture [9], PCDM encoding is performed in a massively parallel manner, achieving 16.7 Gb/s of throughput on an FPGA platform. Moreover, to achieve the same throughput, PCDM uses substantially smaller hardware than LDPC, even with finer rate granularity [9, Sec. 4]. This shows that PCDM is a viable option to realize the fine-grained RM with the maximum throughput of 20 Gb/s, as per the 5G NR requirement.

Another important aspect is that, when PCDM performs RM, the rate of LDPC codes can be made much higher than LDPC-based RM; for example, PCDM-based RM needs $R_C \geq 0.7$ to realize all the target IRs (cf. Table III), whereas LDPC-based RM needs R_C as low as 0.35 for the same IRs (cf. Table I). A higher code rate translates into a smaller number of rows in the parity-check matrix (PCM) for a fixed code length (i.e., for the same number of columns in the PCM). In case of $R_{Info} = 1.4$ bit/symbol and $N_C = 600$, the PCM for the PCDM-based RM has 44% fewer number of rows than the PCM for the LDPC-based RM, which greatly reduces the hardware cost required to implement an LDPC decoder.

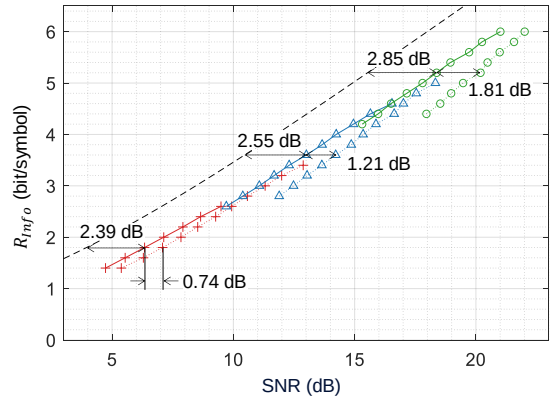
IV. PERFORMANCE EVALUATION

We evaluate the performance of the PCDM-based RM in the additive white Gaussian noise (AWGN) channel for the 5G NR deployment scenarios with $N_C = 600$, 1200, 4800, in comparison with the LDPC-based RM. For each pair of PCDM and LDPC codes, we generate 10^4 blocks of K_D IID random bits of equal probabilities, and perform PCDM encoding. Each PCDM output block is encoded into an LDPC codeword, then mapped to QAM symbols in the PAS architecture (cf. Fig. 1). After going through the AWGN channel, the received data is decoded by the belief propagation algorithm with 12 iterations. Due to the configuration of PCDM and LDPC chosen in this paper, a PCDM block error occurs if and only if an LDPC block error occurs, making the BLERs the same for the LDPC and the PCDM.

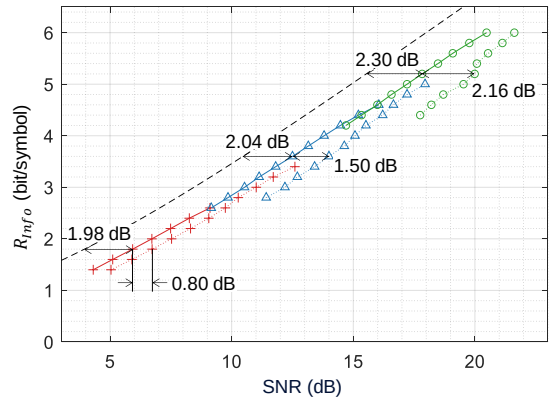
Figs. 4(a)-(c) show the IR and the SNR that is required to achieve a BLER of 10^{-2} with $N_C = 600$, 1200, 4800, respectively. In case of the LDPC-based RM (markers connected by dotted lines), when an IR can be achieved by multiple code-modulation pairs, a high-rate code with a low-



(a)



(b)



(c)

Fig. 4. IR and SNR to achieve BLER = 10^{-2} using LDPC- (dotted lines) and PCDM-based (solid lines) RM schemes, with (a) $N_C = 600$, (b) $N_C = 1200$, (c) $N_C = 4800$, and with 16-ary (pluses), 64-ary (triangles), and 256-ary (circles) QAM. Also shown is the AWGN channel capacity (dashed lines).

order QAM offers a greater IR than a low-rate code with a high-order QAM. The gap to the AWGN capacity (dashed lines) generally increases as the QAM order grows, as an anticipated consequence of bit-interleaved coded-modulation (BICM) with equally probable modulation symbols [15]. By contrast, when PCDM performs RM (markers connected by solid lines), the IR smoothly increases as the QAM order increases, yielding more consistent gap to the capacity than LDPC-based RM. The actual SNR gain obtained from PCDM varies with the IR and N_C, but significant gains are observed in a wide range of the IR, reaching up to 2.16 dB for a large N_C and a large QAM order.

V. Concluding Remarks

We studied the performance and implementation aspects of the PCDM-based RM in some 5G NR deployment scenarios. We realize a wide range of IRs from 1.4 to 6.0 bit/symbol with fine granularity of 0.2 bit/symbol, using 28 PCDM codes and only a few 5G NR LDPC codes. AWGN simulations show that up to 2.16 dB of SNR gain can be obtained with PCDM at a working point of BLER=10^{-2}. Furthermore, this SNR gain can potentially be achieved with a reduced hardware cost than the LDPC-based RM as currently defined in the 5G NR standard.

Although not included in the reported simulation and results, incremental-redundancy HARQ can be incorporated with PCDM. We can, for instance, use the PCDM only for the initial transmission, and transmit additional parity bits via uniform QAM if the initial transmission fails. IRs lower than 1.4 bit/symbol are not studied in this work, as it is difficult to realize them using the proposed method, but the lower IRs can be realized by using the incremental-redundancy HARQ. Full-pledged 5G NR simulations of PCDM-based RM are left for future work, which include the evaluation of the throughput with incremental-redundancy HARQ in fading channels.

References

[1] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; Multiplexing and Channel Coding (Release 15)*, 3GPP TS 38.212 v15.7.0, Sep. 2019.

[2] T. Richardson and S. Kudekar, "Design of low-density parity check codes for 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 28–34, Mar. 2018.

[3] D. Hui et al., "Channel coding in 5G New Radio: A tutorial overview and performance comparison with 4G LTE," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 60–69, Dec. 2018.

[4] J. H. Bae et al., "An overview of channel coding for 5G NR cellular communications," *APSIPA Trans. Signal Inf. Process.*, vol. 8, e17, pp. 1–14, May 2019.

[5] P. Hailes et al., "A flexible FPGA-based quasi-cyclic LDPC decoder," *IEEE Access*, vol. 5, pp. 20965–20984, Mar. 2017.

[6] J. Cho and P. J. Winzer, "Probabilistic constellation shaping for optical fiber communications," *J. Lightw. Technol.*, vol. 37, no. 6, pp. 1590–1607, Mar. 2019.

[7] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4651–4665, Dec. 2015.

[8] J. Cho, "Prefix-free code distribution matching for probabilistic constellation shaping," *IEEE Trans. Commun.*, accepted for publication.

[9] Q. Yu, S. Corteselli, and J. Cho, "FPGA implementation of prefix-free code distribution matching for probabilistic constellation shaping," in *Proc. Opt. Fiber Commun. Conf.*, Mar. 2019, accepted for publication.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[11] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 913–929, May 1993.

[12] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 430–434, Jan. 2016.

[13] P. Schulte and F. Steiner, "Divergence-optimal fixed-to-fixed length distribution matching with shell mapping," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 620–623, Apr. 2019.

[14] Y. C. Gültekin et al., "Probabilistic shaping for finite blocklengths: distribution matching and sphere shaping," 2019, arXiv:1909.08886.

[15] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 927–946, May 1998.

Improved Lower Bounds for Pliable Index Coding using Absent Receivers

Lawrence Ong
University of Newcastle
lawrence.ong@newcastle.edu.au

Badri N. Vellambi
University of Cincinnati
badri.vellambi@uc.edu

Jörg Kliewer
New Jersey Institute of Technology
jkliewer@njit.edu

Parastoo Sadeghi
Australian National University
parastoo.sadeghi@anu.edu.au

***Abstract*—This paper studies pliable index coding, in which a sender broadcasts information to multiple receivers through a shared broadcast medium, and the receivers each have some message a priori and want any message they do not have. An approach, based on receivers that are absent from the problem, was previously proposed to find lower bounds on the optimal broadcast rate. In this paper, we introduce new techniques to obtained better lower bounds, and derive the optimal broadcast rates for new classes of the problems, including all problems with up to four absent receivers.**

I. Introduction

This papers studies pliable index coding, where one transmitter sends information to multiple receivers in a noiseless broadcast setting. In the original index-coding setup [1, 2], each receiver is described by the set of messages that it has, referred to as side information, and the message that it wants from the transmitter. In the *pliable* variant of the problem [3], each receiver is described by only its side information, and its decoding requirement is relaxed to any message not in the side-information set.

The aim for both the original and the pliable problems is to determine the minimum codelength normalised to the message length, referred to as the optimal broadcast rate, that the transmitter must broadcast to satisfy all receivers. As with original index-coding problems, the optimal broadcast rate is not known for pliable-index-coding problems in general.

Even though the two index-coding versions share many similarities, their decoding requirements set them apart in non-trivial ways. As a result, different techniques have been attempted to solve each of them. To date, only a small number of classes of pliable-index-coding problems have been solved. In particular, two classes of *symmetric* problems have been solved [4, 5]. These problems are symmetric in the sense that if a receiver is present in the problem, every receiver with the same cardinality of messages as side information as that of the present receiver is also present. For asymmetric problems, we derived the optimal broadcast rate for some classes of problems based on the absent receivers [6]. We label a receiver by its side-information set, for instance, receiver H has side-information set H. With this notation, we lower bounded the optimal broadcast rate by the longest chain of nested absent receivers, that is, there exist absent receivers $H_1, H_2, \ldots, H_{L_{\max}}$ such that $H_1 \subsetneq H_2 \subsetneq \cdots \subsetneq H_{L_{\max}}$. We characterised the optimal broadcast rate when (i) there exists a message not in the side-information set of any absent receiver, (ii) there is no nested absent receiver pair, (iii) there is only one nested absent receiver pair, and (iv) the absent receivers are formed by taking unions of some message partitions.

However, with the existing results, even a simple problem with three absent receivers remained unsolved (see problem $\mathcal{P}_1$ in Section III). In this paper, we strengthen our previous results to obtain new lower bounds. As a result of the improved lower bounds, we can solve all pliable-index-coding problems with four or fewer absent receivers (which includes $\mathcal{P}_1$).

Our previous results [6] were derived based on our proposed algorithm to construct a decoding chain. Fix a *decoding choice* for each present receiver. The algorithm then iteratively adds messages to the chain. When the current decoding chain corresponds to a present receiver H, the message that receiver H chooses to decode is added to the chain. If the current chain does not correspond to any present receiver, we will arbitrarily "skip" a message not in the chain and also add the same message to the decoding chain. This continues till the chain equals to the whole message set. The fewer the skipped messages, the tighter the lower bound. In this paper, we propose two improvements. First, we modify the algorithm such that even if receiver H is absent, we may not need to skip a message, by looking at receivers $H^- \subsetneq H$, and the messages to be decoded by them. Second, instead of arbitrarily skipping a message, we consider the next absent receiver H' that the algorithm will encounter, and skip a message in such a way that we will be able to avoid skipping a message when the algorithm reaches H'.

We will formally define pliable-index-coding problems in Section II, after which we will use an example to illustrate the above-mentioned two new ideas in Section III. These two ideas will be formally presented in Sections IV and VI. In Section V, we will also present a simpler lower bound. The results will be combined to characterise the optimal broadcast rate for new classes of pliable-index-coding problems in Section VII.

II. Problem Formulation

We use the following notation: $\mathbb{Z}^+$ denotes the set of natural numbers, $[a:b] := \{a, a+1, \ldots, b\}$ for $a, b \in \mathbb{Z}^+$ such that $a < b$, and $X_S = (X_i : i \in S)$ for some ordered set S.

Consider a sender having $m \in \mathbb{Z}^+$ messages, denoted by $X_{[1:m]} = (X_1, \ldots, X_m)$. Each message $X_i \in \mathbb{F}_q$ is independently and uniformly distributed over a finite field of size q. There are n receivers having distinct subsets of messages, which we refer to as side information. Each receiver is labelled by its

side information, i.e., the receiver that has messages X_H, for some $H \subsetneq [1:m]$, will be referred to as receiver H. The aim of the pliable-index-coding problem is to devise an encoding scheme for the sender and a decoding scheme for each receiver satisfying pliable recovery of a message at each receiver.

Without loss of generality, the side-information sets of the receivers are distinct; all receivers having the same side information can be satisfied if and only if (iff) any one of them can be satisfied. Also, no receiver has side information $H = [1:m]$ because this receiver cannot be satisfied. So, there can be at most $2^m - 1$ receivers present in the problem. A pliable index coding problem is thus defined uniquely by m and the set $\mathbb{U} \subseteq 2^{[1:m]} \setminus \{[1:m]\}$ of all present receivers. Any receiver that is not present, i.e., receiver $H \in 2^{[1:m]} \setminus (\{[1:m]\} \cup \mathbb{U}) := \mathbb{U}^{\text{abs}}$, is said to be *absent*.

Given a pliable-index-coding problem with m messages and present receivers $\mathbb{U}$, a pliable index code of length $\ell \in \mathbb{Z}^+$ consists of

- an encoding function of the sender, $\mathsf{E} : \mathbb{F}_q^m \to \mathbb{F}_q^\ell$; and
- for each receiver $H \in \mathbb{U}$, a decoding function $\mathsf{G}_H : \mathbb{F}_q^\ell \times \mathbb{F}_q^{|H|} \to \mathbb{F}_q$, such that $\mathsf{G}_H(\mathsf{E}(X_{[1:m]}), X_H) = X_i$, for some $i \in [1:m] \setminus H$.

Define *decoding choice* D as follows:

$$D : \mathbb{U} \to [1:m], \text{ such that } D(H) \in [1:m] \setminus H. \tag{1}$$

Here, $D(H)$ is the message decoded by receiver H.

The above formulation requires the decoding of only one message at each receiver. Lastly, the aim is to find the optimal broadcast rate for a particular message size q, denoted by $\beta_q := \min_{\mathsf{E},\{\mathsf{G}\}} \ell$ and the optimal broadcast rate over all q, denoted by $\beta := \inf_q \beta_q$.

III. A Motivating Example

We will now use an example to illustrate two ideas proposed in this paper. Consider a pliable-index-coding problem $\mathcal{P}_1$ with six messages and each receiver requires one new message. All receivers are present except receivers $H_1 = \{3\}$, $H_2 = \{1,2,3,4\}$, and $H_3 = \{3,4,5,6\}$. $\mathcal{P}_1$ does not fall into any category for which the optimal rate $\beta_q(\mathcal{P}_1)$ is known.

A. Existing lower bounds

We have previously established a lower bound [6]

$$\beta_q \geq m - L_{\max}, \tag{2}$$

where $L_{\max}$ is the maximum length of any nested chain of absent receivers, that is, $H_1 \subsetneq H_2 \subsetneq \cdots \subsetneq H_{L_{\max}}$, with each $H_i \in \mathbb{U}^{\text{abs}}$. In $\mathcal{P}_1$, $L_{\max} = 2$, which can be obtained from $H_1 \subsetneq H_2$ or $H_1 \subsetneq H_3$. So, $\beta_q(\mathcal{P}_1) \geq 6 - 2 = 4$.

This lower bound can also be obtained by considering another pliable-index-coding problem $\mathcal{P}_1^-$ formed by removing all receivers each having at least one and up to four messages. It has been shown [4] that $\beta_q(\mathcal{P}_1^-) = 4$. Combined with the result $\beta_q(\mathcal{P}_1) \geq \beta_q(\mathcal{P}_1^-)$ [6], we get $\beta_q(\mathcal{P}_1) \geq 4$.

Another lower bound can be obtained by using our previously proposed algorithm [6] to construct a decoding chain of messages. Our previous algorithm is a restriction of our improved Algorithm 1 devised in this paper, in which we have defined $\mathcal{P}_{m,\mathbb{U},D}$ as a pliable-index-coding problem with m messages, receivers $\mathbb{U}$, and decoding choice D. If in lines 4–16 of Algorithm 1, we always choose Option 1 instead of Option 2, we will retrieve our previous algorithm, which for brevity we will refer to as Algorithm 2 in this paper. Using Algorithm 2 on problem $\mathcal{P}_1$, the following lower bound has been shown [6]:

$$\beta_q(\mathcal{P}_1) \geq m - \max_D |S|, \tag{3}$$

where the maximisation is taken over all possible decoding choices D of the receivers, and S is the set of skipped messages obtained from any instance of Algorithm 2 for a specific D.

For Algorithms 1 and 2, we say that the algorithm "hits" a (present or absent) receiver C if it constructs C upon the execution of lines 8, 14, or 19.

For $\mathcal{P}_1$, there exists D for which Algorithm 2 will always hit two absent receivers (either H_1 and H_2, or H_1 and H_3) regardless of which messages we skip. This gives a lower bound $\beta_q(\mathcal{P}_1) \geq m - 2 = 4$. To see this, note that receiver $\emptyset$ is present. Let $D(\emptyset) = 3$. Executing line 19 of the algorithm, we hit $C = \{D(\emptyset)\} = H_1$. Since receiver H_1 is absent, we execute lines 6–10. Supposing that we skip message 1, we will hit $C = \{3,1\}$. Let $D(\{3,1\}) = 2$ and $D(\{3,1,2\}) = 4$. Repeating lines 19–21, we will hit the second absent receiver $H_2 = \{3,1,2,4\}$. So, by defining D in such a way that no matter which message we choose to skip after hitting H_1, the messages to be subsequently added to C stay within H_2 or within H_3 (until we hit H_2 or H_3 respectively), we will always hit H_2 or H_3.

Algorithm 1: A new and generalised algorithm to construct a decoding chain with skipped messages

input : $\mathcal{P}_{m,\mathbb{U},D}$
output : A *decoding chain* C (a totally ordered set with a total order $\preceq$) and a set of *skipped messages* S

1 $C \leftarrow \emptyset$; (initialise C)
2 $S \leftarrow \emptyset$; (initialise S)
3 **while** $C \neq [1:m]$ **do**
4 **if** $C \notin \mathbb{U}$ **then** (receiver C is absent)
5 Choose any of the following options:
6 **Option 1:** (skip a message)
7 Choose any $a \in [1:m] \setminus C$; (skip a)
8 $C \leftarrow C \cup \{a\}$; (expand C)
9 Define $i \preceq a$, for all $i \in C$; (define order in C)
10 $S \leftarrow S \cup \{a\}$; (expand S)
11 **Option 2:** (avoid skipping)
12 Choose any present receiver $B \subsetneq C$, such that $D(B) \notin C$;
13 (look for a subset B, a present receiver)
14 $C \leftarrow C \cup \{D(B)\}$;
15 (add the message that receiver B decodes)
16 Define $i \preceq D(B)$, for all $i \in C$;
17 (define order in C)
18 **else** (receiver C is present)
19 $C \leftarrow C \cup \{D(C) = x\}$;
20 (add the message that receiver C decodes)
21 Define $i \preceq x$, for all $i \in C$; (define order in C)

B. Two new ideas

We will explain the new ideas in this paper by juxtaposing them with Algorithm 2. Since skipping fewer messages gives a tighter lower bound, we introduce the following new ideas to skip fewer messages compared to Algorithm 2:

(a) **Avoid skipping messages:** This is done by using the subsets of C. Using Algorithm 2, when the algorithm hits C, and if receiver C is absent, we skip a message. In Algorithm 1, even if receiver C is absent, if there exists a receiver $B \subsetneq C$ such that $D(B) \notin C$, then the decoding chain can continue by adding $D(B)$ into the chain C without skipping a message (via Option 2).

(b) **Look ahead then skip messages:** Instead of arbitrarily selecting a message $a \in [1:m] \setminus C$ in Option 1, we will base our choice of skipped messages on D. More specifically, we skip a specially chosen message such that the next absent receiver C to be hit will contain a receiver $B \subsetneq C$ whose decoding choice $D(B) \notin C$, and using idea (a), we need not skip a message. This choice will be detailed in Section VI.

C. A new lower bound

Using the above-mentioned ideas, we now construct a new lower bound for $\mathcal{P}_1$. Note that for any D, if fewer than two absent receivers are hit, then $|S| \leq 1$, and this can only lead to the right-hand side of (3) evaluated to 5 or more. So, we only need to consider scenarios where two absent receivers are hit, and in this case the first one must be H_1.

To work out the appropriate choice of skipped message upon hitting H_1, we *look ahead* and consider $D(H_2 \cap H_3) = D(\{3,4\}) = x$. It is necessary that $x \in H_i \setminus H_j$ for some $i, j \in \{2,3\}$ and $i \neq j$. We then skip any message $y \in H_j \setminus H_i$, and update the decoding chain as $C \leftarrow (C \cup \{y\})$. As y is in C now, the only remaining absent receiver that can be hit is H_j. If H_j is not hit, then the algorithm terminates with $|S| = 1$; otherwise, it hits H_j.

When H_j is hit, we can *avoid* skipping a message by noting that (i) there is a present receiver $H_2 \cap H_3 \subsetneq H_j$, and (ii) it decodes $D(H_2 \cap H_3) = x \notin H_j$. The decoding chain continues and terminates without hitting another absent receiver.

This means for any D, we can always choose S such that $|S| \leq 1$. This gives a lower bound of $\beta_q(\mathcal{P}_1) \geq 6 - 1 = 5$. This bounds can be shown to be tight by using a cyclic code for achievability.

More generally, we have the following proposition (which will be proven rigorously later):

Proposition 1: Consider a pliable-index-coding problem $\mathcal{P}_{m,\mathbb{U}}$, where the set of absent receivers is $\mathbb{U}^{\text{abs}} = \{H_1, H_2, H_3\}$, such that $H_1 \subsetneq H_2 \cap H_3$, and $H_2 \cup H_3 = [1:m]$. We have $\beta_q(\mathcal{P}_{m,\mathbb{U}}) = \beta_q(\mathcal{P}_{m,\mathbb{U}}) = m - 1$.

IV. A New and Generalised Algorithm

Compared to Algorithm 2, the new Algorithm 1 has Option 2, which implements the two new ideas in Section III-B. It allows us to avoid skipping a message even when an absent receiver C is hit, as long as a suitable present receiver $B \subsetneq C$ can be found. If Option 1 is always selected, we revert back to Algorithm 2 as a special case. Although choosing Option 1 may seem counter-intuitive, we will see that later that choosing Option 1 simplifies the proof of our results as it avoids evaluating $D(B)$ required in Option 2.

The sketch of proof for the lower bound (3) for Algorithm 2 is as follows [6]: We started with a bipartite graph G_D that describes $\mathcal{P}_{m,\mathbb{U},D}$. We showed that for each instance of Algorithm 2, there is a series of pruning operations on G_D that yield an acyclic graph G'_D with $m - |S|$ remaining messages. The graph G_D is acyclic because, by construction, all directed edges flow from message nodes that are *larger* to message nodes that are *smaller* with respect to the order $\preceq$. As $m - |S|$ is a lower bound on $\mathcal{P}_{m,\mathbb{U},D}$ [7, Lem. 1], and that $\beta_q(\mathcal{P}_{m,\mathbb{U}}) = \min_D \beta_q(\mathcal{P}_{m,\mathbb{U},D})$, we have (3).

We now show that the lower bound (3) is still valid using Algorithm 1. Algorithm 1 differs from Algorithm 2 by having Option 2. Using Option 2 on a present receiver B, this receiver is preserved (that is, not removed during the pruning operation) in the graph G_D. With this additional receiver not removed (compared to Algorithm 2), there are additional directed edges flowing from the a *larger* message node to *smaller* message nodes with respect to the order $\preceq$, that is, from the message node $D(B)$ to message nodes $\{x \in B\}$ through the receiver node B. Clearly, all additional edges retained due to Option 2 in Algorithm 1 do not create any directed cycle. Hence, the proof for the lower bound (3) for Algorithm 2 can be modified accordingly to give the following:

Lemma 1: Consider a pliable-index-coding problem $\mathcal{P}_{m,\mathbb{U}}$. For a specific D, let S be the set of skipped messages for an instance of Algorithm 1. Then,

$$\beta_q(\mathcal{P}_{m,\mathbb{U}}) \geq m - \max_D |S|. \tag{4}$$

The lower bound is obtained by maximising $|S|$ over all decoding choices D. By optimising the choice of skipped messages for each D such that the minimum number of messages is skipped, we obtain the following lower bound:

$$\beta_q(\mathcal{P}_{m,\mathbb{U}}) \geq m - \max_D \min_S |S| = m - L^*, \tag{5}$$

where we define

$$L^* := \max_D \min_S |S|. \tag{6}$$

Remark 1: For any given D, although any instance of Algorithm 1 gives a lower bound for $\beta_q(\mathcal{P}_{m,\mathbb{U},D})$, skipping as few messages as possible gives tighter lower bounds.

Intuitively, Algorithm 1 says that the construction of decoding chain C can continue even if receiver C is absent, because if receiver $B \subsetneq C$ can decode $D(B) \notin C$, then knowing C, one is able to obtain $D(B)$ to extend the decoding chain.

Before formally deriving the second idea of "look ahead and skip" in Section VI, in the next section, we first improve upon an existing lower bound that can be obtained by simply looking at how the absent receivers are nested, that is, without needing an algorithm that constructs decoding chains.

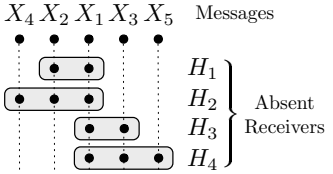


Fig. 1: Pliable-index-coding problem $\mathcal{P}_2$ for Example 1

V. An Improved Nested-Chain Lower Bound

From (5), we see that any upper bound on L^* provides a lower bound on β_q. For instance, see lower bound (2), where $L^* \leq L_{\max}$. The lower bound based on $L_{\max}$ may be loose, because we may be able to skip certain messages to avoid hitting some absent receivers in the longest chain. In this paper, we will prove a better* lower bound based on this idea. We now prove the following theorem:

Theorem 1: $L^* \leq L-1$ if the following condition holds: For every chain of absent receivers of length at least L, say, $H_1 \subsetneq \cdots \subsetneq H_{L'}$ for some $L' \geq L$, where $H_i \in \mathbb{U}^{\text{abs}}$, there exists $H_k \cup \{a\}$ (for some $k \in [1:L-1]$ and for some $a \notin H_k$) such that there is no chain of absent receivers of length $L-k$ where $(H_k \cup \{a\}) \subseteq \underbrace{H'_1 \subsetneq \cdots \subsetneq H'_{L-k}}_{\text{absent-receiver chain}}$, with $H'_i \in \mathbb{U}^{\text{abs}}$.

Proof of Theorem 1: Recall that each instance of Algorithm 1 (or Algorithm 2) returns a decoding chain $C = \{c_1, c_2, \ldots, c_m\}$, in the order $c_i \preceq c_j$ iff $i \leq j$, and a set of skipped messages $S \subseteq C$.

Let c_i by the kth skipped message. This means the algorithm must have hit an absent receiver $H \in \mathbb{U}^{\text{abs}}$, where

$$H = \begin{cases} \emptyset, & \text{if } i = 1, \\ \{c_1, \ldots, c_{i-1}\}, & \text{otherwise (i.e., } i \in [2:m]). \end{cases} \tag{7}$$

Suppose that ℓ is the maximum number of absent receivers that can form a chain $(H \cup \{c_i\}) \subseteq H'_1 \subsetneq H'_2 \subsetneq \cdots \subsetneq H'_\ell$, with each $H'_i \in \mathbb{U}^{\text{abs}}$. Then, at most ℓ more absent receivers can be hit. Consequently, the algorithm must terminate with $|S| \leq k + \ell$.

Now, for all nested receiver chains of length L or larger, suppose that the condition stated in the theorem is true, we can always skip message a after hitting H_k, such that $|S| < k + (L-k)$. As $|S|$ is an integer, $|S| \leq L-1$. Since this is true for all nested receiver chains of length L or larger, we can always avoid skipping L messages, giving $L^* \leq L-1$. ∎

We will show the utility of Theorem 1 using an example:

Example 1: Consider $\mathcal{P}_2$ with five messages and four absent receivers $H_1 = \{1,2\}, H_2 = \{1,2,4\}, H_3 = \{1,3\}$, and $H_4 = \{1,3,5\}$, as depicted in Figure 1. The length of the longest nested absent-receiver chain is 2. Our previous lower bound gives $\beta_q \geq 3$ (see (2)). Now, we invoke Theorem 1, and consider all chains of length $L \geq 2$, which are $H_1 \subsetneq H_2$ and $H_3 \subsetneq H_4$.

- When H_1 is hit, we skip message 3. $\{1,2,3\}$ is not contained in any absent receiver.
- When H_3 is hit, we skip message 4. $\{1,3,4\}$ is not contained in any absent receiver.

So, we have $L^* \leq 1$. Noting (6) and (5), we get $\beta_q \geq 5 - 1 = 4$. This lower bound can be achieved by the code $(X_3 + X_5, X_1, X_2, X_4)$.

While this new nested-chain lower bound improved on our previous longest-chain lower bound, it is still insufficient to solve $\mathcal{P}_1$ described in Section III. To solve $\mathcal{P}_1$, we will use the "look ahead and skip" technique detailed in the next section.

VI. Skipping Messages with Look Ahead

In this section, when we hit an absent receiver H, we will propose a method to skip a message in such a way to guarantee that we will subsequently not need to skip any message when we hit any absent receiver from a special subset, say $\mathbb{A}$, of absent receivers. This method is used in conjunction with Algorithm 1. Note that due to the algorithm, all members in $\mathbb{A}$ must be strict supersets of H. We now show a class of $\mathbb{A}$:

Theorem 2: Let $H \in \mathbb{U}^{\text{abs}}$ by an absent receiver, and $\mathbb{A} \subseteq \mathbb{U}^{\text{abs}} \setminus \{H\}$ be a subset of absent receivers that belongs to any of the following cases, where $H \subsetneq H'$ for all $H' \in \mathbb{A}$. Running Algorithms 1, suppose that H is hit. We can always choose to skip a message such that, if any $H' \in \mathbb{A}$ is hit subsequently, we can avoid skipping a message.

1) $\bigcup_{H' \in \mathbb{A}} H' \neq [1:m]$.
2) $\mathbb{A}$ is a minimal cover† of $[1:m]$, $T := \bigcap_{H' \in \mathbb{A}} H' \supsetneq H$, and $T \in \mathbb{U}$.
3) $\mathbb{A}$ is a minimal cover of $[1:m]$, and $\bigcap_{H' \in \mathbb{A}} H' = H$; furthermore, there exist‡ $H_1, H_2 \in \mathbb{A}$ such that $T := H_1 \cap H_2 \supsetneq H$ and $T \in \mathbb{U}$.

Proof of Theorem 2: For case 1, by skipping any $a \in [1:m] \setminus (\bigcup_{H' \in \mathbb{A}} H')$, we will not hit any absent receiver in $\mathbb{A}$.

For case 2, we *look ahead* and check $D(T)$. Since receiver T is present, $D(T)$ is defined. As $T := \bigcap_{H' \in \mathbb{A}} H'$ and $D(T) \notin \bigcap_{H' \in \mathbb{A}} H'$, there must exist an absent receiver $H_1 \in \mathbb{A}$ that does not contain $D(T)$. As $\mathbb{A}$ is a minimal cover, there exists some $a \in H_1$ that is not in all other sets in $\mathbb{A}$, that is, $a \notin \bigcup_{H' \in \mathbb{A} \setminus \{H_1\}} H'$. We choose to skip a, and by doing so, we will never hit any receiver in $\mathbb{A} \setminus \{H_1\}$. If we hit H_1, we can choose Option 2 in the algorithm without needing to skip any message, since $T \subseteq H_1$ and $D(T) \notin H_1$.

For case 3, we *look ahead* and check $D(T)$. As receiver T is present, $D(T)$ is defined. $D(T) \notin T = H_1 \cap H_2$. Without loss of generality, suppose $D(T) \notin H_1$. When we follow the same argument for case 2 by skipping some $a \in H_1$ that is not in all other sets in $\mathbb{A}$. By doing so, will can always avoid skipping a message due to hitting H_1. ∎

VII. Applications of Results

A. Optimal rates for the slightly imperfect L-nested setting

We have previously defined a class of pliable-index-coding problems as follows [6]:

*The new lower bound is strictly better for certain problems.

†A family of sets $\mathbb{A} = \{A_\ell : \ell \in L\}$ is a minimal cover of B iff $\bigcup_{\ell \in L} A_\ell = B$, and for any strict subset $L' \subsetneq L$, $\bigcup_{\ell \in L'} A_\ell \subsetneq B$.

‡If this is false, $\mathbb{A} \cup \{H\}$ forms 1-truncated L-nested absent receivers, which we will define in Definition 2 later.

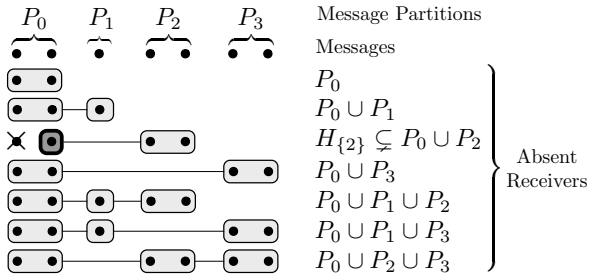


Fig. 2: Slightly imperfect 3-nested absent receivers, formed by shrinking the side-information set of one receiver among perfect 3-nested absent receivers.

Definition 1: A pliable-index-coding problem is said to have *perfect L-nested absent receivers* iff the messages $[1:m]$ can be partitioned into $L+1 \in [2:m]$ subsets $P_0, P_1, \ldots, P_L$ (that is, $\bigcup_{i=0}^{L} P_i = [1:m]$ and $P_i \cap P_j = \emptyset$ for all $i \neq j$), such that only P_0 can be an empty set, and there are exactly $2^L - 1$ *absent* receivers, which are defined as

$$H_Q := P_0 \cup \left(\bigcup_{i \in Q} P_i\right), \text{ for each } Q \subsetneq [1:L]. \tag{8}$$

For any pliable-index-coding problem $\mathcal{P}_{m,\mathbb{U}}$ with perfect L-nested absent receivers, $\beta_q(\mathcal{P}_{m,\mathbb{U}}) = m - L$ [6].

With Theorem 2, we can determine the optimal rate of problems deviating from the perfect L-nested setting. We now prove the optimal rate for pliable-index-coding problems with slightly imperfect L-nested absent receivers. Figure 2 depicts a example of slightly imperfect 3-nested absent receivers.

Theorem 3: Consider a pliable-index-coding problem $\mathcal{P}_{m,\mathbb{U}}$ that comprises perfect L-nested absent receivers with the following change: one absent receiver $H_Q = P_0 \cup \left(\bigcup_{i \in Q} P_i\right)$, for some $Q \subsetneq [1:L]$, is changed to the absent receiver $H_Q \subsetneq P_0 \cup \left(\bigcup_{i \in Q} P_i\right)$. Then, $\beta_q(\mathcal{P}_{m,\mathbb{U}}) = m - L + 1$.

Proof of Thm 3: See the extended version of this paper [8].

We can now prove Proposition 1 that we stated earlier.

Proof of Proposition 1: $\mathcal{P}_{m,\mathbb{U}}$ is formed by having perfect 2-nested absent receivers with $P_0 = H_2 \cap H_3$, $P_1 = H_2 \setminus H_3$, $P_2 = H_3 \setminus H_2$, and then replacing absent receiver P_0 with $H_1 \subsetneq P_0$. Using Theorem 3, we have $\beta_q(\mathcal{P}_{m,\mathbb{U}}) = m - 1$. ∎

B. Optimal rates for T-truncated L-nested absent receivers

We define another variation of perfect L-nested absent receivers.

Definition 2: A pliable-index-coding problem is said to have *T-truncated L-nested absent receivers* iff the messages $[1:m]$ can be partitioned into $L+1 \in [2:m]$ subsets $P_0, P_1, \ldots, P_L$ (that is, $\bigcup_{i=0}^{L} P_i = [1:m]$ and $P_i \cap P_j = \emptyset$ for all $i \neq j$), such that only P_0 can be an empty set, and there are $\sum_{i=0}^{T} \binom{L}{i}$ absent receivers, which are defined as

$$H_Q = P_0 \cup \left(\bigcup_{i \in Q} P_i\right), \ \forall Q \subsetneq [1:L], \text{ with } |Q| \in [0:T], \tag{9}$$

for some $T \in [0 : L-1]$.

Note that $(L-1)$-truncated L-nested absent receivers are equivalent to perfect L-nested absent receivers. Figure 3 depicts an example of 1-truncated 3-nested absent receivers.

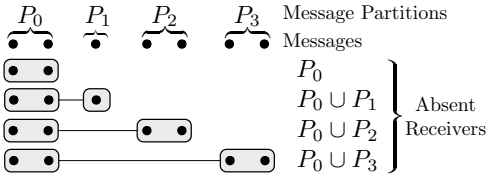


Fig. 3: 1-truncated 3-nested absent receivers, formed by keeping the top few groups of perfect 3-nested absent receivers

Theorem 4: For any pliable-index-coding problem $\mathcal{P}$ with T-truncated L-nested absent receivers, $\beta(\mathcal{P}) = \beta_q(\mathcal{P}) = m - T - 1$, for sufficiently large q.

Proof of Thm 4: See the extended version of this paper [8].

C. Optimal rates for a small number of absent receivers

We have established that $\beta_q = m$ if and only if there is no absent receiver, that is $|\mathbb{U}^{\text{abs}}| = 0$.

Corollary 1: If $1 \leq |\mathbb{U}^{\text{abs}}| \leq 2$, then $\beta_q = m - 1$.

Proof: For $|\mathbb{U}^{\text{abs}}| = 1$, by definition, the absent receiver $H \subsetneq [1:m]$, and hence $\bigcup_{H \in \mathbb{U}^{\text{abs}}} H \neq [1:m]$. So, the result follows from [6, Thm. 1]. For $|\mathbb{U}^{\text{abs}}| = 2$, there can be either no nested pair or one nested pair of absent receivers. The result follows from [6, Thm. 3]. ∎

While the optimal rate for up to two absent receivers can be determined using our previous results, we need the new results presented in this paper for more absent receivers.

Theorem 5: Suppose $|\mathbb{U}^{\text{abs}}| = 3$. Then

$$\beta_q = \begin{cases} m-2, & \text{if the absent receivers are perfect 2-nested,} \\ m-1, & \text{otherwise.} \end{cases}$$

Theorem 6: Suppose $|\mathbb{U}^{\text{abs}}| = 4$. Then

$$\beta_q = \begin{cases} m-2, & \text{if a subset of absent receivers is either} \\ & \text{perfect 2-nested or 1-truncated 3-nested,} \\ m-1, & \text{otherwise.} \end{cases}$$

Proofs of Thms 5 and 6: See the extended version [8].

References

[1] Z. Bar-Yossef, Y. Birk, T. S. Jayram, and T. Kol, "Index coding with side information," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1479–1494, Mar. 2011.

[2] L. Ong, "Optimal finite-length and asymptotic index codes for five or fewer receivers," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7116–7130, Nov. 2017.

[3] S. Brahma and C. Fragouli, "Pliable index coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6192–6203, Nov. 2014.

[4] T. Liu and D. Tuninetti, "Information theoretic converse proofs for some PICOD problems," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Kaohsiung, Taiwan, Nov. 6–10 2017, pp. 284–288.

[5] ——, "An information theoretic converse for the "consecutive complete-S" PICOD problem," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, Nov. 25–29 2018, pp. 165–169.

[6] L. Ong, B. N. Vellambi, and J. Kliewer, "Optimal-rate characterisation for pliable index coding using absent receivers," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, July 7-12 2019.

[7] M. J. Neely, A. S. Tehrani, and Z. Zhang, "Dynamic index coding for wireless broadcast networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7525–7540, Nov. 2013.

[8] L. Ong, B. N. Vellambi, J. Kliewer, and P. Sadeghi. (2019) Improved lower bounds for pliable index coding using absent receivers. [Online]. Available: https://arxiv.org/pdf/1909.11850.pdf

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

On the Capacity of Private Monomial Computation

Yauhen Yakimenka, Hsuan-Yin Lin, and Eirik Rosnes
Simula UiB, N–5008 Bergen, Norway
Email: {yauhen, lin, eirikrosnes}@simula.no

***Abstract*—In this work, we consider private monomial computation (PMC) for replicated noncolluding databases. In PMC, a user wishes to privately retrieve an arbitrary multivariate monomial from a candidate set of monomials in f messages over a finite field $\mathbb{F}_q$, where $q = p^k$ is a power of a prime p and $k \geq 1$, replicated over n databases. We derive the PMC capacity under a technical condition on p and for asymptotically large q. The condition on p is satisfied, e.g., for large enough p. Also, we present a novel PMC scheme for arbitrary q that is capacity-achieving in the asymptotic case above. Moreover, we present formulas for the entropy of a multivariate monomial and for a set of monomials in uniformly distributed random variables over a finite field, which are used in the derivation of the capacity expression.**

I. Introduction

The concept of private computation (PC) was introduced independently by Sun and Jafar [1] and Mirmohseni and Maddah-Ali [2]. In PC, a user wishes to compute a function of the messages stored in a set of databases without revealing any information about the function to any of the databases. PC can be seen as a generalization of private information retrieval (PIR). In PIR, a user wants to retrieve a single message from the set of databases privately. Applications of PC include, in principle, all scenarios where insights about certain actions of the user should be kept private. One practical motivation for considering arbitrary functions is that of *algorithmic privacy*, as protecting the identity of an algorithm running in the cloud could be even more critical than data privacy in some scenarios. Not only could the algorithm be valuable, but also in some cases, parameters of the algorithm carry lifetime secrets such as biological information of individuals [2].

The capacity in the linear case, i.e., the computation of arbitrary linear combinations of the stored messages, has been settled for both replicated [1] and coded [3], [4] databases. In the coded databases scenario, the messages are encoded by a linear code before being distributed and stored in a set of databases. Interestingly, the capacity in the linear case is equal to the corresponding PIR capacity for both replicated and coded databases. The monomial case was recently considered in [5], [6]. However, the presented achievable schemes have a PC rate, defined here as the ratio between the *smallest* desired amount of information and the total amount of downloaded information, that in general is strictly lower than the best known converse bound for a finite number of messages. PC schemes in the coded case for arbitrary polynomials were considered by Karpuk and Raviv in [7], [8], and recently improved in [5] when the number of messages is small. The capacity of private polynomial computation for coded databases remains open.

In this work, we first derive formulas for the entropy of a multivariate monomial and a set of monomials in uniformly distributed random variables over a finite field. We then present a novel PC scheme for multivariate monomials in the messages stored in a set of replicated noncolluding databases. The key ingredient of the scheme is the use of discrete logarithms. The discrete logarithm in the multiplicative group of a finite field of order $q = p^k$ (p is a prime and $k \geq 1$) is a bijection to the integer ring of size $q-1$, mapping multiplication to addition. Hence, the discrete logarithm maps multivariate monomial retrieval to linear function retrieval, given that none of the messages is the zero element. The latter holds with probability approaching one as q becomes large. The corresponding PC rate in this limiting case is derived using the entropy formulas from the first part of the paper. When the candidate set of multivariate monomials is fixed (i.e., independent of q), the PC rate converges to the PIR capacity for any number of messages stored in the databases, under a technical condition on p and as q goes to infinity. The condition on p is satisfied, e.g., for large enough p. Also, the presented monomial computation scheme is capacity-achieving in this asymptotic case.

II. Preliminaries

A. General Definitions and Notation

Throughout the paper, vectors are denoted by bold font and matrices are written as sans-serif capitals.

We work with different algebraic structures: the ring of integers $\mathbb{Z}$, rings of residuals $\mathbb{Z}_m$ for integers $m > 1$, and finite fields $\mathbb{F}_q$, where $q = p^k$ is a power of a prime p and $k \geq 1$. Occasionally, $\mathcal{R}$ denotes any of these structures. We often use the connection between $\mathbb{Z}$ and $\mathbb{Z}_m$. In principle, any element in $\mathbb{Z}$ can be considered as an element of $\mathbb{Z}_m$, with correspondence of addition and multiplication. If an expression consists of both integers and elements of $\mathbb{Z}_m$, we assume all operations are over $\mathbb{Z}_m$. When we need to stress that an element is in $\mathbb{Z}_m$, we write $a^{\langle m \rangle} \in \mathbb{Z}_m$ for $a \in \mathbb{Z}$. The same notation is used for matrices, e.g., $\mathsf{A}^{\langle m \rangle}$ has entries $a_{ij}^{\langle m \rangle} \in \mathbb{Z}_m$ for $a_{ij} \in \mathbb{Z}$.

Any $a \in \mathbb{Z}$ can be viewed as $a^{\langle p \rangle} \in \mathbb{Z}_p = \mathbb{F}_p \subseteq \mathbb{F}_q$. Operations on such elements of $\mathbb{F}_q$ are modulo p, as p is the *characteristic* of $\mathbb{F}_q$, i.e., the minimum positive integer l such that $l \cdot \alpha = 0$ for all $\alpha \in \mathbb{F}_q$. Analogously, $\mathsf{A} \in \mathbb{Z}^{s \times t}$ can be viewed as $\mathsf{A}^{\langle p \rangle} \in \mathbb{F}_q^{s \times t}$. Note the difference between $\mathsf{A}^{\langle p \rangle} \in \mathbb{F}_q^{s \times t}$ and $\mathsf{A}^{\langle q \rangle} \in \mathbb{Z}_q^{s \times t}$ for $q = p^k$ and $k > 1$.

The *multiplicative group* $\mathbb{F}_q^* = \mathbb{F}_q \setminus \{0\}$ is cyclic (cf. [9, Thm. 2.18]), and it is possible to define a *discrete logarithm*

function[1] $\mathrm{dlog} : \mathbb{F}_q^* \to \mathbb{Z}_{q-1}$, which is an isomorphism between $(\mathbb{F}_q^*, \times)$ and $(\mathbb{Z}_{q-1}, +)$.

We write $[a] \triangleq \{1, \ldots, a\}$ for a positive integer a. The *greatest common divisor (gcd)* of $a_1, \ldots, a_s \in \mathbb{Z}$ is denoted by $\gcd(a_1, \ldots, a_s)$, with the convention $\gcd(0, \ldots, 0) \triangleq 0$ and $\gcd(a_1^{\langle m \rangle}, \ldots, a_s^{\langle m \rangle}, m) \triangleq \gcd(a_1, \ldots, a_s, m)$. We write $a \mid b$ when a divides b, and $a \nmid b$ otherwise. The binomial coefficient of a over b (both nonnegative integers) is denoted by $\binom{a}{b}$ where $\binom{a}{b} = 0$ if $a < b$. The transpose of A is denoted by $\mathsf{A}^\intercal$.

A $k \times k$ *minor* in $\mathcal{R}$ of a matrix $\mathsf{A} \in \mathcal{R}^{s \times t}$, for a positive integer k, is the determinant of a $k \times k$ submatrix of A obtained by removing $s-k$ rows and $t-k$ columns from A. The largest integer r such that there is a nonzero $r \times r$ minor of A is called the *rank* of A in $\mathcal{R}$ and denoted by $\mathrm{rank}_{\mathcal{R}}\, \mathsf{A}$. A matrix $\mathsf{A} \in \mathcal{R}^{s \times s}$ is invertible in $\mathcal{R}$ if and only if the determinant of A is invertible as an element of $\mathcal{R}$ (cf. [9, Thm. 2.1]).

For $\mathsf{A} \in \mathbb{Z}^{s \times t}$, we denote the gcd of all $k \times k$ minors of A by $g_k(\mathsf{A})$. If $\delta \in \mathbb{Z}$ is some minor of A, the corresponding minor of $\mathsf{A}^{\langle m \rangle}$ is $\delta^{\langle m \rangle}$. Hence, $\mathrm{rank}_{\mathbb{Z}_m}\, \mathsf{A} = \mathrm{rank}_{\mathbb{Z}}\, \mathsf{A}$ for all $m \nmid g_r(\mathsf{A})$, where $r = \mathrm{rank}_{\mathbb{Z}}\, \mathsf{A}$.[2] Also,

$$\mathrm{rank}_{\mathbb{F}_q}\, \mathsf{A} = \mathrm{rank}_{\mathbb{Z}}\, \mathsf{A} \quad \Leftrightarrow \quad p \nmid g_r(\mathsf{A}). \tag{1}$$

It is known [10, Cor. 1.13, Cor. 1.20] that there exists a unique diagonal matrix $\mathsf{D} = \mathrm{diag}(d_1, \ldots, d_{\min(s,t)}) \in \mathbb{Z}^{s \times t}$ called the *Smith normal form* of A, with the following properties.

1) $\mathsf{D} = \mathsf{PAQ}$ for some matrices $\mathsf{P} \in \mathbb{Z}^{s \times s}$ and $\mathsf{Q} \in \mathbb{Z}^{t \times t}$ invertible in $\mathbb{Z}$,
2) $d_i \mid d_{i+1}$ for $i \in [\min(s,t) - 1]$,
3) $d_1 d_2 \cdots d_i = g_i(\mathsf{A})$ for $i \in [\min(s,t)]$.

The diagonal elements $d_1, \ldots, d_{\min(s,t)}$ are *invariant factors*, and $d_i = 0$ if and only if $i > \mathrm{rank}_{\mathbb{Z}}\, \mathsf{A}$. While D is unique, the matrices P and Q are not unique in the general case. It is also important to mention that the Smith normal form is defined for matrices over *principal ideal domains (PIDs)*. For example, $\mathbb{Z}$ is a PID while $\mathbb{Z}_m$ is not (in general).

Random variables are labeled by capital roman letters and we write $X \sim Y$ to indicate that X and Y are identically distributed. Moreover, $X \sim \mathfrak{U}(\mathcal{S})$ means that X is uniformly distributed over the set $\mathcal{S}$. We use $\log$ to denote logarithm base-2, although most statements hold for an arbitrary constant base. We denote the entropy in bits and q-ary units by $\mathrm{H}(\cdot)$ and $\mathrm{H}_q(\cdot)$, respectively, and $\mathrm{I}(\cdot;\cdot)$ denotes mutual information. The binary entropy function is denoted by $h(\cdot)$.

The notation $O(\phi(x))$ stands for any function $\psi(x)$ in x such that $|\psi(x)/\phi(x)| < \mathsf{B}$ for all large enough x and some constant $\mathsf{B} > 0$ independent of x. Also, $o(\phi(x))$ represents any $\psi(x)$ such that $\lim_{x\to\infty} \psi(x)/\phi(x) = 0$. In particular, $O(1)$ is any bounded function and $o(1)$ is any function that converges to zero as $x \to \infty$.

B. Private Computation

Suppose we have n noncommunicating databases, each storing duplicated data: f messages subpacketized into λ parts, each part denoted as $X_i^{(j)} \in \mathbb{F}_q$ for $i \in [f]$ and $j \in [\lambda]$. The subpackets are considered mutually independent and uniformly drawn from $\mathbb{F}_q$. There are μ public functions $\varphi_1, \ldots, \varphi_\mu$, where $\varphi_i : \mathbb{F}_q^f \to \mathbb{F}_q$ for $i \in [\mu]$. The user randomly chooses a secret index $V \sim \mathfrak{U}([\mu])$ and wants to retrieve

$$\boldsymbol{F}_V = \left(\varphi_V(\boldsymbol{X}^{(1)}), \ldots, \varphi_V(\boldsymbol{X}^{(\lambda)})\right) \in \mathbb{F}_q^\lambda,$$

where $\boldsymbol{X}^{(j)} \triangleq (X_1^{(j)}, \ldots, X_f^{(j)})$, $j \in [\lambda]$, without revealing any information about V. To achieve that, the user and the databases employ the following scheme.

1) The user generates secret randomness R, computes queries $Q_j = Q_j(V, R)$, $j \in [n]$, and sends the j-th query to the j-th database.
2) Based on Q_j and all the messages, the j-th database computes the response $A_j = A_j\left(Q_j, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(\lambda)}\right)$ and sends it back to the user.
3) Using all available information, the user can recover $\boldsymbol{F}_V$.

Formally, we require the scheme to satisfy

$$\begin{aligned} &\text{Privacy:} && \mathrm{I}(V; Q_j) = 0, \text{ for all } j \in [n], \\ &\text{Recovery:} && \mathrm{H}(\boldsymbol{F}_V \mid V, R, A_1, \ldots, A_n) = 0. \end{aligned}$$

Definition 1. The *download rate* of a PC scheme over the field $\mathbb{F}_q$, referred to as the PC rate, is defined as

$$\mathsf{R} = \mathsf{R}(n, f, \mu, \{\varphi_i\}, \lambda, \{Q_j\}, \{A_j\}, q) \triangleq \frac{\min_{v \in [\mu]} \mathrm{H}(\boldsymbol{F}_v)}{\Delta},$$

where Δ is the expected total number of downloaded bits, referred to as the *download cost*. The supremum of all achievable rates for all choices of λ, $\{Q_j\}$, and $\{A_j\}$ is the *PC capacity* over $\mathbb{F}_q$, $\mathsf{C}_{\mathrm{PC}}(n, f, \mu, \{\varphi_i\}, q)$.

In case $\mu = f$ and $\varphi_i(x_1, \ldots, x_f) = x_i$ for $i \in [f]$, PC reduces to PIR with capacity $\mathsf{C}_{\mathrm{PIR}}(n, f) \triangleq (1 + 1/n + 1/n^2 + \cdots + 1/n^{f-1})^{-1}$ [11]. Note that $\mathsf{C}_{\mathrm{PIR}}$ is independent of q.

The case when $\varphi_1, \ldots, \varphi_\mu$ are linear functions described by a matrix of coefficients $\mathsf{A} \in \mathbb{F}_q^{\mu \times f}$ without zero rows, is referred to as private linear computation (PLC). Its capacity $\mathsf{C}_{\mathrm{PLC}}$ only depends on n and $r = \mathrm{rank}_{\mathbb{F}_q}\, \mathsf{A}$, and it holds that $\mathsf{C}_{\mathrm{PLC}}(n, r) = \mathsf{C}_{\mathrm{PIR}}(n, r)$ [1].[3]

In this work, we consider private monomial computation (PMC), i.e., the case when $\varphi_i(x_1, \ldots, x_f) = x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_f^{a_{if}}$, $i \in [\mu]$, where $a_{ij} \in \mathbb{Z}$. The monomials can be described by a matrix of degrees $\mathsf{A} = (a_{ij}) \in \mathbb{Z}^{\mu \times f}$, and we assume there are no constant functions, i.e., no zero rows in A. The capacity of PMC is denoted by $\mathsf{C}_{\mathrm{PMC}}(n, f, \mu, \mathsf{A}, q)$.

III. Entropies of Linear Functions and Monomials

Lemma 1. *Let $a \in \mathbb{Z}$ and $Y \sim \mathfrak{U}(\mathbb{Z}_m)$. Then,*

$$\mathrm{H}(aY) = \mathrm{H}(a^{\langle m \rangle} Y) = \log m - \log \gcd(a, m).$$

Proof: From the theory of linear congruences [12, Sec. 5, Thm. 1], the equation $ay = b$ has $d = \gcd(a, m)$ solutions in

[1] Strictly speaking, dlog requires fixing a particular generator of $\mathbb{F}_q^*$.

[2] In particular, the requirement $a \nmid b$ is satisfied if $a > b$.

[3] In [1], the authors assume the messages are among the functions, e.g., $\varphi_i(x_1, \ldots, x_f) = x_i$ for $i \in [f]$. However, this is not required as we can define linearly independent functions as new variables and express other functions in these variables.

$\mathbb{Z}_m$ if $d \mid b$ and no solutions otherwise. Therefore, the random variable aY takes m/d different values from $\mathbb{Z}_m$ equiprobably, and the required statement follows. ∎

Lemma 2. *Let $\mathsf{A} \in \mathbb{Z}^{s\times t}$ be a fixed matrix whose invariant factors are $d_1, \dots, d_{\min(s,t)}$. Let $\boldsymbol{Y} = (Y_1, \dots, Y_t) \sim \mathfrak{U}(\mathbb{Z}_m^t)$, $r = \operatorname{rank}_{\mathbb{Z}} \mathsf{A}$, and $r' = \operatorname{rank}_{\mathbb{Z}_m} \mathsf{A}^{\langle m\rangle}$. Then,*

$$\mathrm{H}(\mathsf{A}\boldsymbol{Y}) = r\log m - \sum_{i=1}^{r} \log\gcd(d_i, m) \tag{2}$$

$$= r'\log m - \sum_{i=1}^{r'} \log\gcd(d_i, m). \tag{3}$$

Proof: Recall that, since $\boldsymbol{Y}$ is defined over $\mathbb{Z}_m^t$, the operations in $\mathsf{A}\boldsymbol{Y}$ are over $\mathbb{Z}_m$. In other words, $\mathsf{A}\boldsymbol{Y}$ is a shorthand for $\mathsf{A}^{\langle m\rangle}\boldsymbol{Y}$.

Let $\mathsf{D} = \mathsf{PAQ}$ be the Smith normal form of A, where both $\mathsf{P} \in \mathbb{Z}^{s\times s}$ and $\mathsf{Q} \in \mathbb{Z}^{t\times t}$ are invertible over $\mathbb{Z}$ (i.e., their determinants are ± 1) and $\mathsf{D} = \operatorname{diag}(d_1, \dots, d_r, 0, \dots, 0)$. After taking modulo m from both sides, we obtain $\mathsf{D}^{\langle m\rangle} = \mathsf{P}^{\langle m\rangle}\mathsf{A}^{\langle m\rangle}\mathsf{Q}^{\langle m\rangle}$, where $\mathsf{P}^{\langle m\rangle}$ and $\mathsf{Q}^{\langle m\rangle}$ are both invertible over $\mathbb{Z}_m$ (their determinants are ± 1 in $\mathbb{Z}_m$ too) and $\mathsf{D}^{\langle m\rangle} = \operatorname{diag}(d_1^{\langle m\rangle}, \dots, d_r^{\langle m\rangle}, 0, \dots, 0)$. Therefore,

$$\begin{aligned}\mathrm{H}(\mathsf{D}^{\langle m\rangle}\boldsymbol{Y}) &= \mathrm{H}(\mathsf{P}^{\langle m\rangle}(\mathsf{A}^{\langle m\rangle}\mathsf{Q}^{\langle m\rangle}\boldsymbol{Y})) = \mathrm{H}(\mathsf{A}^{\langle m\rangle}\mathsf{Q}^{\langle m\rangle}\boldsymbol{Y})\\ &= \mathrm{H}(\mathsf{A}^{\langle m\rangle}(\mathsf{Q}^{\langle m\rangle}\boldsymbol{Y})) = \mathrm{H}(\mathsf{A}^{\langle m\rangle}\boldsymbol{Y}) = \mathrm{H}(\mathsf{A}\boldsymbol{Y}),\end{aligned}$$

because $\mathsf{P}^{\langle m\rangle}$ and $\mathsf{Q}^{\langle m\rangle}$ are invertible over $\mathbb{Z}_m$, and multiplication from the left by an invertible matrix is a bijection. Thus, we can consider $\mathrm{H}(\mathsf{D}^{\langle m\rangle}\boldsymbol{Y})$ instead of $\mathrm{H}(\mathsf{A}\boldsymbol{Y})$. But $\mathsf{D}^{\langle m\rangle}\boldsymbol{Y} = (d_1^{\langle m\rangle}Y_1, \dots, d_r^{\langle m\rangle}Y_r, 0, \dots, 0)$ with mutually independent entries. Hence,

$$\begin{aligned}\mathrm{H}(\mathsf{D}^{\langle m\rangle}\boldsymbol{Y}) &= \sum_{i=1}^{r} \mathrm{H}(d_i^{\langle m\rangle}Y_i)\\ &\overset{\text{Lem. 1}}{=} r\log m - \sum_{i=1}^{r} \log\gcd(d_i, m).\end{aligned}$$

Finally, (3) holds because $m \mid d_i$ for $i > r'$ and hence $\gcd(d_i, m) = m$. ∎

Corollary 1. *In the setting of Lemma 2, $\mathrm{H}(\mathsf{A}\boldsymbol{Y}) = r\log m + O(1)$, as $m \to \infty$, where $r = \operatorname{rank}_{\mathbb{Z}} \mathsf{A}$.*

Proof: For all $m > d_r$ and all $i \in [\min(s,t)]$, it holds that $d_i^{\langle m\rangle} = d_i$. In this case, $r' = r$ and

$$\begin{aligned}\mathrm{H}(\mathsf{A}\boldsymbol{Y}) &= r\log m - \sum_{i=1}^{r} \log\gcd(d_i, m)\\ &\geq r\log m - \log\prod_{i=1}^{r} d_i = r\log m - \log g_r(\mathsf{A}).\end{aligned} \tag{4}$$

On the other hand,

$$\mathrm{H}(\mathsf{A}\boldsymbol{Y}) = r\log m - \sum_{i=1}^{r} \log\gcd(d_i, m) \leq r\log m. \tag{5}$$

We note that both (4) and (5) are attained for infinitely many values of m, e.g., for $m = ug_r(\mathsf{A})$ and $m = 1 + ug_r(\mathsf{A})$, respectively (for any positive integer u). In other words, $\mathrm{H}(\mathsf{A}\boldsymbol{Y})$ does not converge as $m \to \infty$.

Finally, as $\log g_r(\mathsf{A})$ does not depend on m, we have

$$\mathrm{H}(\mathsf{A}\boldsymbol{Y}) = r\log m + O(1), \text{ as } m \to \infty.$$ ∎

Next, we present some results on entropies of monomials over finite fields. The key idea is to use the bijection of dlog and treat a special case of zero separately.

Lemma 3. *Let $a_1, \dots, a_t \in \mathbb{Z}$, $X_1, \dots, X_t \sim \mathfrak{U}(\mathbb{F}_q)$ be mutually independent, τ be the number of nonzeros among $a_1, \dots, a_t$, and $\pi = (1 - 1/q)^\tau$. Then,*

$$\mathrm{H}(X_1^{a_1}X_2^{a_2}\cdots X_t^{a_t}) = h(\pi) + \pi\log\frac{q-1}{\gcd(a_1, \dots, a_t, q-1)}.$$

Moreover, if not all $a_1, \dots, a_t$ are zeros,

$$\mathrm{H}_q(X_1^{a_1}X_2^{a_2}\cdots X_t^{a_t}) \xrightarrow[q\to\infty]{} 1.$$

Proof: If $a_i = 0$, the variable X_i is not present in the monomial. Hence, we can exclude such variables and assume $a_1, \dots, a_\tau \in \mathbb{Z} \setminus \{0\}$. Dropping zero arguments of the gcd above does not change its value either.

Let $M = X_1^{a_1}X_2^{a_2}\cdots X_\tau^{a_\tau}$. Define $Z = 0$ if $M = 0$ and $Z = 1$ otherwise. Then, $\pi = \mathbb{P}\{M \neq 0\} = \mathbb{P}\{Z = 1\}$ and

$$\begin{aligned}\mathrm{H}(M) &= \mathrm{H}(Z) + \mathrm{H}(M \mid Z) - \mathrm{H}(Z \mid M)\\ &= h(\pi) + \mathrm{H}(M \mid Z = 0)(1-\pi) + \mathrm{H}(M \mid Z = 1)\pi\\ &= h(\pi) + \pi\,\mathrm{H}(M \mid M \neq 0).\end{aligned}$$

Now, $M \neq 0$ if and only if none of $X_1, \dots, X_\tau$ is zero. In this case, all $X_1, \dots, X_\tau \in \mathbb{F}_q^*$ and we can define $Y_j = \operatorname{dlog} X_j \in \mathbb{Z}_{q-1}$ for $j \in [\tau]$ and $L' = \operatorname{dlog} M = a_1Y_1 + \cdots + a_\tau Y_\tau \in \mathbb{Z}_{q-1}$. Since dlog is bijective, $Y_1, \dots, Y_\tau \sim \mathfrak{U}(\mathbb{Z}_{q-1})$ and $\mathrm{H}(M \mid M \neq 0) = \mathrm{H}(L')$. By applying Lemma 2 with $m = q-1$, $s = 1$, $r = 1$, and $d_1 = \gcd(a_1, \dots, a_\tau)$, we get

$$\mathrm{H}(L') = \log\frac{q-1}{\gcd(a_1, \dots, a_\tau, q-1)}.$$

Further, as $q \to \infty$, $\pi \to 1$ and therefore $h(\pi) \to 0$. Additionally, $\gcd(a_1, \dots, a_\tau, q-1) \leq \min(|a_1|, \dots, |a_\tau|) = O(1)$, as $q \to \infty$. Finally,

$$\mathrm{H}_q(X_1^{a_1}X_2^{a_2}\cdots X_t^{a_t}) = \frac{\mathrm{H}(X_1^{a_1}X_2^{a_2}\cdots X_t^{a_t})}{\log q} \xrightarrow[q\to\infty]{} 1.$$ ∎

Theorem 1. *Let $\mathsf{A} \in \mathbb{Z}^{s\times t}$ be a fixed matrix of coefficients with rank $r = \operatorname{rank}_{\mathbb{Z}} \mathsf{A}$. Let $X_1, \dots, X_t \sim \mathfrak{U}(\mathbb{F}_q)$ be mutually independent. For $i \in [s]$, define $M_i = X_1^{a_{i1}}X_2^{a_{i2}}\cdots X_t^{a_{it}} \in \mathbb{F}_q$ and $\boldsymbol{M} = (M_1, \dots, M_s)$. Then,*

$$\mathrm{H}(\boldsymbol{M}) = r\log q + O(1), \text{ as } q \to \infty.$$

Proof: First, if there is a zero column in A, we can drop the corresponding variable, as it does not influence either the values of any of the monomials or $\operatorname{rank}_{\mathbb{Z}} \mathsf{A}$. Thus, for the remainder of the proof, we assume there are no zero columns

in A, and we also consider values of q large enough so that there are no zero columns in $\mathsf{A}^{\langle q-1\rangle}$ as well.

Define $Z=0$ if $X_1X_2\cdots X_t=0$ and $Z=1$ otherwise. It holds that $\pi=\mathbb{P}\{Z=1\}=(1-1/q)^t$. Moreover, $Z=0$ if and only if any of the monomials $M_1,\dots,M_s$ is zero. Hence, $\mathrm{H}(Z\mid \boldsymbol{M})=0$ and we have

$$\begin{aligned}\mathrm{H}(\boldsymbol{M}) &= \mathrm{H}(Z)+\mathrm{H}(\boldsymbol{M}\mid Z)-\mathrm{H}(Z\mid \boldsymbol{M})\\ &= h(\pi)+(1-\pi)\,\mathrm{H}(\boldsymbol{M}\mid Z=0)+\pi\,\mathrm{H}(\boldsymbol{M}\mid Z=1).\end{aligned}$$

Next, $Z=1$ if and only if none of $X_1,\dots,X_t$ is zero, i.e., all $X_1,\dots,X_t\in\mathbb{F}_q^*$. In this case, we can define $Y_j=\operatorname{dlog}X_j\in\mathbb{Z}_{q-1}$, for $j\in[t]$, $L'_i=\operatorname{dlog}M_i=a_{i1}Y_1+\cdots+a_{it}Y_t\in\mathbb{Z}_{q-1}$, for $i\in[s]$, and $\boldsymbol{L}'=(L'_1,\dots,L'_s)$. Then, $\mathrm{H}(\boldsymbol{L}')=\mathrm{H}(\boldsymbol{M}\mid Z=1)$ and

$$\begin{aligned}&|\,\mathrm{H}(\boldsymbol{M})-\mathrm{H}(\boldsymbol{L}')| = |\,\mathrm{H}(\boldsymbol{M})-\mathrm{H}(\boldsymbol{M}\mid Z=1)|\\ &= |h(\pi)+(1-\pi)\,\mathrm{H}(\boldsymbol{M}|Z=0)+(\pi-1)\,\mathrm{H}(\boldsymbol{M}\mid Z=1)|\\ &\le h(\pi)+(1-\pi)|\,\mathrm{H}(\boldsymbol{M}|Z=0)-\mathrm{H}(\boldsymbol{M}|Z=1)|\\ &\le h(\pi)+s(1-\pi)\log q=o(1),\ \text{as } q\to\infty.\end{aligned}$$

From Corollary 1 with $m=q-1$, we have $\mathrm{H}(\boldsymbol{L}')=r\log(q-1)+O(1)=r\log q+O(1)$, as $q\to\infty$. Finally,

$$\mathrm{H}(\boldsymbol{M})=\mathrm{H}(\boldsymbol{L}')+o(1)=r\log q+O(1),\ \text{as } q\to\infty.\quad\blacksquare$$

Corollary 2. *In the setting of Theorem 1, consider $q=p^k$ with $p\nmid g_r(\mathsf{A})$. Then,*

$$|\mathrm{H}_q(\boldsymbol{M})-\mathrm{H}_q(\boldsymbol{L})|=o(1),\ \textit{as } q\to\infty,$$

where $L_i=a_{i1}X_1+\cdots+a_{it}X_t\in\mathbb{F}_q$ for $i\in[s]$, and $\boldsymbol{L}=(L_1,\dots,L_s)$.[4]

Proof: As A defines a linear transformation of a vector space over $\mathbb{F}_q$, $\mathrm{H}(\boldsymbol{L})=\operatorname{rank}_{\mathbb{F}_q}\mathsf{A}\cdot\log q$. From (1) and since $p\nmid g_r(\mathsf{A})$, we obtain $\operatorname{rank}_{\mathbb{F}_q}\mathsf{A}=\operatorname{rank}_{\mathbb{Z}}\mathsf{A}=r$. Next, from Theorem 1, as $q\to\infty$,

$$|\mathrm{H}_q(\boldsymbol{M})-\mathrm{H}_q(\boldsymbol{L})|=\frac{|\mathrm{H}(\boldsymbol{M})-\mathrm{H}(\boldsymbol{L})|}{\log q}=\frac{O(1)}{\log q}=o(1).\quad\blacksquare$$

Note that we do not require p to be either fixed or infinitely large. However, all primes $p>g_r(\mathsf{A})$ satisfy the requirement $p\nmid g_r(\mathsf{A})$. Corollary 2 states that the entropy of any fixed set of monomials is equal to the entropy of the corresponding set of linear functions (i.e., defined by the same matrix A), both over $\mathbb{F}_q$, when $p\nmid g_r(\mathsf{A})$ and as q approaches infinity. Moreover, this also holds for conditional entropies consisting of various sets of monomials because they can be expressed as a difference of two unconditional entropies. This key observation is further used in Section IV-B.

IV. Achievable Scheme

A. Sun–Jafar Scheme for Private Linear Computation

We build our PMC achievable scheme based on the Sun–Jafar scheme for PLC ([1, Alg. 1], referred to as *PC* there). Due to lack of space, we do not present their scheme in all details and refer the reader to [1] for a full description and analysis. Here, we briefly repeat the facts (in our notation) essential for further discussion.

The Sun–Jafar scheme uses $\lambda=n^\mu$ subpackets. From each of the n databases, the user downloads symbols in μ blocks. The b-th block, $b\in[\mu]$, of each database consists of $(n-1)^{b-1}\binom{\mu}{b}$ symbols, and each symbol is a linear combination (using only coefficients ±1) of b judiciously chosen pieces $\varphi_u(\boldsymbol{X}^{(j)})$ for different values of $u\in[\mu]$ and $j\in[\lambda]$. Since all φ_u are linear combinations, each symbol the user downloads is some linear combination of $\{X_i^{(j)}\}$. The user's randomized queries define which linear combinations the databases will reply with. The queries enforce symmetry across databases and function evaluation symmetry within symbols downloaded from each database. This ensures privacy of the user.

A crucial observation is that $(n-1)^{b-1}\binom{\mu-r}{b}$ of the symbols in block b of each database are redundant based on side information downloaded from other databases. More precisely, these redundant symbols are linear combinations of other symbols in block b from the same database as well as symbols downloaded from other databases. Hence, they need not to be downloaded, as the user can reconstruct them offline. This preserves the user's privacy while reducing the download cost to the value corresponding to the PLC capacity. A distinctive property of the Sun–Jafar scheme is that it is oblivious to the coefficients of the linear functions φ_v. It is only the number of them, μ, that matters. Furthermore, the scheme can be used for PIR if $\mu=f$ and the linear functions are the messages, i.e., $\varphi_i(x_1,\dots,x_f)=x_i$ for $i\in[f]$. In this case, there are no redundant symbols in any block.

B. Private Monomial Computation

Let $\lambda=n^\mu$ and suppose that none of $\{X_i^{(j)}\}$ equals zero. Then we can construct a *multiplicative* scheme by substituting each linear combination of $\{\varphi_v\}$ in the Sun–Jafar scheme with a corresponding multiplicative combination. For example, if at some step the user downloads the symbol $\varphi_1(\boldsymbol{X}^{(j_1)})+\varphi_2(\boldsymbol{X}^{(j_2)})-\varphi_3(\boldsymbol{X}^{(j_3)})$, $j_1,j_2,j_3\in[\lambda]$, then the corresponding multiplicative combination is $\varphi_1(\boldsymbol{X}^{(j_1)})\varphi_2(\boldsymbol{X}^{(j_2)})\big(\varphi_3(\boldsymbol{X}^{(j_3)})\big)^{-1}$, where the functions φ_v now denote the corresponding monomials. Since there are no zeros among $\{X_i^{(j)}\}$, all operations are valid and ensure correct reconstruction of the monomial of interest. Moreover, from Corollary 2, when $p\nmid g_r(\mathsf{A})$ and as $q\to\infty$, the entropies of all the symbols as well as the entropy of each block b conditioned on the side information received from other databases converge to those of the Sun–Jafar scheme. This means that in the multiplicative scheme above, a database can also encode the whole b-th block into no more than $(n-1)^{b-1}\left(\binom{\mu}{b}-\binom{\mu-r}{b}\right)$ q-ary symbols, resulting in the same download cost as in the Sun–Jafar scheme. Since there is only a finite number of entropies involved, we can satisfy the requirement on p from Corollary 2 for all of them simultaneously, e.g., by requiring p to be large enough (but not necessarily approaching infinity).

[4] In contrast to Lemma 2 and Corollary 1, $\boldsymbol{L}$ is defined over the field.

Now, in case any of $\{X_i^{(j)}\}$ equals zero, we can ignore dependencies between the monomials and run a PIR scheme, for example, the same Sun–Jafar scheme in PIR mode for μ messages. Altogether, our scheme is as follows.

Algorithm 1: PMC Scheme

1 **if** *there are no zeros among* $\{X_i^{(j)}\}$ *and* $\mu > r$ **then**
2 Each database replies according to the multiplicative scheme.
3 **else**
4 Each database replies according to the Sun–Jafar scheme in PIR mode oblivious to the dependencies between the monomials.

Note that the queries of both schemes need to be uploaded since the user does not know if there are zeros among $\{X_i^{(j)}\}$. Moreover, the user can determine which scheme is used (Line 2 or Line 4) from (r, μ) and the size of the responses (the size is smaller for the multiplicative scheme provided $r < \mu$).

We note that privacy of the user in the suggested PMC scheme is inherited from the privacy of the Sun–Jafar scheme.

Theorem 2. *For PMC with n databases, f messages, and μ monomials defined by a degree matrix $\mathsf{A} \in \mathbb{Z}^{\mu \times f}$ of rank $r = \operatorname{rank}_{\mathbb{Z}} \mathsf{A}$, for $p \nmid g_r(\mathsf{A})$ and as $q \to \infty$, the PMC capacity converges to that of PIR: $\mathrm{C}_{\mathrm{PMC}}(n, f, \mu, \mathsf{A}, q) \to \mathrm{C}_{\mathrm{PIR}}(n, r)$.*

Proof: First, we show that the PC rate $\mathrm{C}_{\mathrm{PIR}}(n, r)$ is achievable by Algorithm 1. For Line 2, for $p \nmid g_r(\mathsf{A})$ and as $q \to \infty$, the download cost measured in q-ary units converges to $n^\mu / \mathrm{C}_{\mathrm{PLC}}(n, r) = n^\mu / \mathrm{C}_{\mathrm{PIR}}(n, r)$. The download cost at Line 4 is $n^\mu / \mathrm{C}_{\mathrm{PIR}}(n, \mu)$.

The probability that none of $\{X_i^{(j)}\}$ equals zero is $\pi = (1-1/q)^{n^\mu f} \to 1$, as $q \to \infty$. Therefore, the average download cost of Algorithm 1 becomes

$$n^\mu \left(\frac{\pi}{\mathrm{C}_{\mathrm{PIR}}(n, r)} + \frac{1-\pi}{\mathrm{C}_{\mathrm{PIR}}(n, \mu)} \right) \xrightarrow[q \to \infty]{} \frac{n^\mu}{\mathrm{C}_{\mathrm{PIR}}(n, r)}.$$

On the other hand, from Lemma 3, it follows that

$$\min_{v \in [\mu]} \mathrm{H}_q(\boldsymbol{F}_v) = n^\mu \cdot \min_{v \in [\mu]} \mathrm{H}_q(\varphi_v(\boldsymbol{X}^{(1)})) \xrightarrow[q \to \infty]{} n^\mu.$$

Altogether, we have that the download rate of our PMC scheme converges to the PIR capacity for r messages.

It remains to prove the converse, i.e., showing that $\mathrm{C}_{\mathrm{PIR}}(n, r)$ is an upper (or outer) bound on the PC capacity. For that, we consider the general converse in [6, Thm. 1] and show that, for $q \to \infty$ and provided $p \nmid g_r(\mathsf{A})$, the upper bounds from [6, Thm. 1] coincide for the monomial and linear cases with the same matrix A. Note that [6, Thm. 1] gives $\mu!$ upper bounds on the PC capacity (according to the number of permutations of μ functions). For the linear case, the outer bounds in [6, Thm. 1] reduce to $\mathrm{C}_{\mathrm{PIR}}(n, r)$, independent of q. In general, for a fixed permutation, the bound depends on $\min_{v \in [\mu]} \mathrm{H}_q(\varphi_v(\boldsymbol{X}^{(1)}))$ and joint entropies of different subsets of function evaluations. Then, it follows from the key observation in Section III that this bound is coinciding for the monomial and linear cases as $q \to \infty$, provided $p \nmid g_r(\mathsf{A})$ (details omitted for brevity). ∎

Corollary 3. *In the setting of Theorem 2, the scheme in Algorithm 1 is capacity-achieving for $p \nmid g_r(\mathsf{A})$ and as $q \to \infty$.*

Note that we prove that the scheme in Algorithm 1 is capacity-achieving only for asymptotic q and provided $p \nmid g_r(\mathsf{A})$. As an example, take $\mu = f = 2$, $n = 2$, $\varphi_1(x_1, x_2) = x_1^2 x_2$, and $\varphi_2(x_1, x_2) = x_1 x_2^2$. Then the asymptotic PC rate of Corollary 3 is $\mathrm{C}_{\mathrm{PIR}}(2, 2) = 2/3$, since $r = \operatorname{rank}_{\mathbb{Z}} \mathsf{A} = 2$. On the other hand, the PC capacity C_{PC} for two arbitrary functions for any finite field is known [1, Sec. VII, Eq. (82)]. For this example, $\mathrm{C}_{\mathrm{PC}} = 2\,\mathrm{H} / (\mathrm{H}(X_1^2 X_2, X_1 X_2^2) + \mathrm{H})$, where $\mathrm{H} \triangleq \mathrm{H}(X_1^2 X_2) = \mathrm{H}(X_1 X_2^2)$ and the superscripts on the X's have been suppressed for brevity. Finally, Algorithm 1 defaults to PIR mode and achieves the PC rate $2\,\mathrm{H}/3$, which can be shown to be smaller than C_{PC} for any finite q.

V. Conclusion

We derived the PMC capacity for replicated noncolluding databases, by considering the case of an arbitrary large field and under a technical condition on the size p of the base field, which is satisfied, e.g., for p large enough. A PMC scheme that is capacity-achieving in the above asymptotic case was also outlined. Furthermore, we presented formulas for the entropy of a multivariate monomial and for a set of monomials in uniformly distributed random variables over a finite field.

Acknowledgment

The authors would like to thank Srimathi Varadharajan and Alessandro Melloni for useful discussions.

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] M. Mirmohseni and M. A. Maddah-Ali, "Private function retrieval," in *Proc. Iran Workshop Commun. Inf. Theory (IWCIT)*, Tehran, Iran, Apr. 25–26, 2018, pp. 1–6.
[3] 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.
[4] S. A. Obead, H.-Y. Lin, E. Rosnes, and J. Kliewer, "Capacity of private linear computation for coded databases," in *Proc. 56th Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Oct. 2–5, 2018, pp. 813–820.
[5] ——, "Private polynomial computation for noncolluding coded databases," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1677–1681.
[6] ——, "On the capacity of private nonlinear computation for replicated databases," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 25–28, 2019, pp. 1–5.
[7] 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.
[8] N. Raviv and D. A. Karpuk, "Private polynomial computation from Lagrange encoding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1672–1676.
[9] N. Jacobson, *Basic Algebra I*, 2nd ed. Freeman and Company, 1985.
[10] C. Norman, *Finitely Generated Abelian Groups and Similarity of Matrices over a Field*. Springer Science & Business Media, 2012.
[11] 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.
[12] U. Dudley, *Elementary Number Theory*, 2nd ed. Freeman and Company, 1978.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Asymptotic Absorbing Set Enumerators for Irregular LDPC Code Ensembles

Emna Ben Yacoub
Technical University of Munich
Institute for Communications Engineering
Munich, Germany
Email: emna.ben-yacoub@tum.de

Gianluigi Liva
German Aerospace Center (DLR)
Institute of Communications and Navigation
Wessling, Germany
Email: gianluigi.liva@dlr.de

***Abstract*—The normalized logarithmic asymptotic distribution of elementary and general absorbing sets for irregular low-density parity-check (LDPC) code ensembles is derived and evaluated. The method is based on enumerating binary matrices with specified column and row weight vectors and solving a system of equations.**

I. Introduction

The performance under iterative decoding of LDPC codes [1] is frequently dominated, in the error floor region, by the presence of specific graphical configurations in the code Tanner graphs [2], [3], [4]. Such structures are typically referred to as *stopping sets* over the binary erasure channel (BEC) [2], and as *trapping sets* [3], [4] over more general channel models. As pointed out in [5], not all trapping sets may cause decoding failures. Nevertheless, a characterization (e.g., enumeration) of trapping sets for LDPC code ensembles is of paramount importance to gain a deeper understanding of the error floor phenomenon. This is especially relevant to applications demanding very low error floors [6], [7], where a Monte Carlo simulation approach to the error floor estimation is impractical. A sub-class of trapping sets that is particularly harmful is the one of *absorbing sets* [8].

With some notable exceptions (see, e.g., [9], [10], [11]), the impact of stopping/trapping/absorbing sets on the performance of a code has been often addressed from a code ensemble perspective. In [2] the average stopping set enumerators for finite-length LDPC code ensembles were introduced. An asymptotic analysis of the stopping set distribution for unstructured irregular LDPC code ensembles was devised in [12]. The analysis was later extended to protograph-based LDPC and generalized LDPC code ensembles [13] and to unstructured doubly-generalized LDPC code ensembles. An elegant derivation of the asymptotic trapping set enumerators for regular/irregular LDPC code ensembles was provided in [14] and is based on random matrix enumeration methods. The approach was adopted in [5] to obtain the asymptotic absorbing set enumerators of regular LDPC code ensembles. Trapping set enumerators for protograph-based LDPC codes were derived in [15], whereas pseudocodeword weight enumerators for protograph-based generalized LDPC code ensembles were introduced in [13].[1]

In this work, we follow the code ensemble perspective to analyze absorbing set enumerators of LDPC code ensembles. In particular, we extend the analysis of [5] to unstructured irregular LDPC code ensembles. The paper is organized as follows. In Section II, we review the main definitions and known results. Section III provides the derivation of the asymptotic absorbing sets enumerators for irregular LDPC code ensembles. Numerical results are presented and discussed in Section IV. Conclusions follow in Section V.

II. Preliminaries

A. LDPC Codes

Binary LDPC codes are binary linear block codes defined by an $m \times n$ sparse parity-check matrix $\boldsymbol{H}$. The code dimension is $k \geq n - m$. The Tanner graph of an LDPC code is a bipartite graph $G = (\mathcal{V} \cup \mathcal{C}, \mathscr{E})$ consisting of n variable nodes (VNs) and m check nodes (CNs). The set $\mathscr{E}$ of edges contains the elements e_{ij}, where e_{ij} is an edge between VN $\mathtt{v}_j \in \mathcal{V}$ and CN $\mathtt{c}_i \in \mathcal{C}$. Note that e_{ij} belongs to the set $\mathcal{E}$ if and only if the parity-check matrix element h_{ij} is equal to 1. The sets $\mathcal{N}(\mathtt{v}_j)$ and $\mathcal{N}(\mathtt{c}_i)$ denote the neighbors of VN $\mathtt{v}_j$ and CN $\mathtt{c}_i$, respectively. The degree of a VN $\mathtt{v}_j$ (CN $\mathtt{c}_i$) is the cardinality of the set $\mathcal{N}(\mathtt{v}_j)$ ($\mathcal{N}(\mathtt{c}_i)$). The node-oriented degree distribution polynomials of an LDPC code graph are given by $\Lambda(x) = \sum_i \Lambda_i x^i$ and $\mathrm{P}(x) = \sum_i \mathrm{P}_i x^i$, where Λ_i corresponds to the fraction of VNs with degree i and P_i corresponds to the fraction of CNs with degree i. We further define by $d_{\mathtt{v}}^{\max}$ ($d_{\mathtt{c}}^{\max}$) is the maximum VN (CN) degree. We denote by $\bar{d}_{\mathtt{v}} = \sum_i i\Lambda_i$ the average VN degree. An unstructured irregular LDPC code ensemble $\mathscr{C}_n^{\Lambda,\mathrm{P}}$ is the set of all LDPC codes with block length n and node-oriented degree distributions $\Lambda(x)$ and $\mathrm{P}(x)$.

B. Absorbing Sets

For a set $\mathcal{S} \subseteq \mathcal{V}$ of VNs, we denote by $\mathcal{N}(\mathcal{S})$ the set of its neighboring CNs. Further, we denote by $\mathcal{O}(\mathcal{S})$ the set of CNs in $\mathcal{N}(\mathcal{S})$ that are connected to $\mathcal{S}$ an odd number of times and $\mathcal{E}(\mathcal{S})$ the set of CNs in $\mathcal{N}(\mathcal{S})$ that are connected to $\mathcal{S}$ an even number of times.

[1] Recently, a notion of absorbing sets for generalized LDPC codes has been introduced, together with an initial absorbing set enumeration analysis [16].

Definition 1. An (a, b) trapping set (TS) $\mathcal{T}_{a,b}$ is set $\mathcal{S}$ of a VNs such that $\mathcal{O}(\mathcal{S})$ contains b CNs [14].

Definition 2. An (a, b) absorbing set (AS) $\mathcal{A}_{a,b}$ is a trapping set with the additional property that each VN $\mathsf{v} \in \mathcal{S}$ has strictly fewer neighboring CNs from $\mathcal{O}(\mathcal{S})$ than from $\mathcal{E}(\mathcal{S})$ [5].

Moreover, an elementary absorbing set (EAS) $\mathcal{A}^{\mathrm{E}}_{a,b}$ is an AS where each CN $\mathsf{c} \in \mathcal{E}(\mathcal{S})$ is connected to two VNs in $\mathcal{S}$ and each CN $\mathsf{c} \in \mathcal{O}(\mathcal{S})$ is connected to exactly one VN in $\mathcal{S}$.

C. Random Matrix Enumeration

Definition 3. Let $x(n)$ and $y(n)$ be two real-valued sequences, where $y(n) \neq 0\ \forall n$. The sequence $x(n)$ is called exponentially equivalent to $y(n)$ as $n \to \infty$ if and only if $\lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{x(n)}{y(n)}\right) = 0$. We will use the notation $x(n) \doteq y(n)$ to specify that $x(n)$ is exponentially equivalent to $y(n)$.

Lemma 1. It holds for every sequence $y(w)$

$$\sum_{w} \exp(ny(w)) \doteq \exp(n \max_{w} y(w)). \tag{1}$$

Theorem 1. Let $\mathcal{H}^{\boldsymbol{R},\boldsymbol{L}}_{m,n}$ be the set of all $m \times n$ binary matrices with row weight vector $\boldsymbol{R} = (R_1, \ldots, R_m)$ and column weight vector $\boldsymbol{L} = (L_1, \ldots, L_n)$, where R_i, $1 \leq i \leq m$, is the weight of the i-th row and L_j, $1 \leq j \leq n$, is the weight of the j-th column. The cardinality of $\mathcal{H}^{\boldsymbol{R},\boldsymbol{L}}_{m,n}$ for constant ratio $\xi = m/n$ and $\max\{\max_i R_i, \max_j L_j\} \leq (\ln(n))^{1/4-\epsilon}, \epsilon > 0$, as $n \to \infty$ is given by [17]

$$|\mathcal{H}^{\boldsymbol{R},\boldsymbol{L}}_{m,n}| = \frac{f!}{\prod_{j=1}^{n} L_j! \prod_{i=1}^{m} R_i!} \cdot (1 + o(n^{-1+\delta})) \times \exp\left[-\frac{1}{2f^2}\left(\sum_{i=1}^{m} R_i(R_i-1)\right)\left(\sum_{j=1}^{n} L_j(L_j-1)\right)\right]$$

and for $\delta > 0$, with $f = \sum_{j=1}^{n} L_j = \sum_{i=1}^{m} R_i$.

III. Asymptotic Distribution of Absorbing Sets

In this section, we derive the asymptotic distribution of ASs and EASs for the ensemble $\mathscr{C}^{\Lambda,\mathrm{P}}_n$ for $a = \theta n$ and $b = \gamma n$. We write the transpose of the parity-check matrix as

$$\boldsymbol{H}^T = \left[\begin{array}{c|c} \boldsymbol{M}_1 & \boldsymbol{M}_2 \\ \hline \multicolumn{2}{c}{\boldsymbol{M}_3} \end{array}\right] \tag{2}$$

where $\boldsymbol{M}_1$ is a $a \times (m-b)$ binary matrix representing the subgraph of the Tanner graph containing the VNs in $\mathcal{A}_{a,b}$ ($\mathcal{A}^{\mathrm{E}}_{a,b}$) and the CNs that are connected to $\mathcal{A}_{a,b}$ ($\mathcal{A}^{\mathrm{E}}_{a,b}$) an even number of times (including zero), $\boldsymbol{M}_2$ is a $a \times b$ binary matrix corresponding to the subgraph of the Tanner graph containing the VNs in $\mathcal{A}_{a,b}$ ($\mathcal{A}^{\mathrm{E}}_{a,b}$) and the CNs that are connected to $\mathcal{A}_{a,b}$ ($\mathcal{A}^{\mathrm{E}}_{a,b}$) an odd number of times, and $\boldsymbol{M}_3$ is a $(n-a) \times m$ binary matrix representing the remainder of the Tanner graph [5]. Note that the columns of $\boldsymbol{M}_1$ have even weights and the ones of $\boldsymbol{M}_2$ have odd weights.

The parity-check matrix of each code from $\mathscr{C}^{\Lambda,\mathrm{P}}_n$ contains $\Lambda_j n$ columns of weight j and $\mathrm{P}_i m$ rows of weight i. The cardinality of the set containing all $m \times n$ binary matrices with these row and column weights is given by

$$|\mathcal{H}^{\boldsymbol{R},\boldsymbol{L}}_{m,n}| = \frac{N!}{\prod_{i=1}^{d_{\mathsf{c}}^{\max}} (i!)^{\mathrm{P}_i m} \prod_{j=1}^{d_{\mathsf{v}}^{\max}} (j!)^{\Lambda_j n}} (1 + o(n^{-1+\delta})) \times \exp\left[-\frac{mn \sum_{i=1}^{d_{\mathsf{c}}^{\max}} i(i-1)\mathrm{P}_i \sum_{j=1}^{d_{\mathsf{v}}^{\max}} j(j-1)\Lambda_j}{2N^2}\right] \tag{3}$$

for $\delta > 0$, with $N = n\bar{d}_{\mathsf{v}}$.

We denote by $\alpha^{(i)}_k$ the number of columns in $\boldsymbol{H}^T$ of weight i whose first a entries sum to k, where $k \in \{0, 1, 2\}$ for $\mathcal{A}^{\mathrm{E}}_{a,b}$ and $k \in \{0, \ldots, i\}$ for $\mathcal{A}_{a,b}$. Similarly, $\beta^{(j)}_k$ represents the number of rows in $\boldsymbol{H}^T$ of weight j whose first $m-b$ entries sum to $k \in \{\lfloor \frac{j}{2} \rfloor + 1, \ldots, j\}$. Further, we introduce $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d_{\mathsf{v}}^{\max}})$, where $n\theta_j$ represents the number of VNs of degree j in $\mathcal{A}_{a,b}$ ($\mathcal{A}^{\mathrm{E}}_{a,b}$), i.e., the number of rows of weight j in the submatrix $[\boldsymbol{M}_1 \,|\, \boldsymbol{M}_2]$. Note that $\sum_j \theta_j = \theta$. We define $\mathcal{M}_l$ as the set of binary matrices with the same weight vectors as $\boldsymbol{M}_l$ for $l = 1, 2, 3$ and the set $\mathcal{M}$ containing all $n \times m$ binary matrices with the structure shown in (2) and where $\boldsymbol{M}_l \in \mathcal{M}_l$ for $l = 1, 2, 3$.

A. Elementary Absorbing Sets

Consider the matrix $\boldsymbol{M}_1$. It contains, for each $j \in \{1, \ldots, d_{\mathsf{v}}^{\max}\}$, $\beta^{(j)}_k$ rows of weight $k \in \{\lfloor \frac{j}{2} \rfloor + 1, \ldots, j\}$ and, for each $i \in \{1, \ldots, d_{\mathsf{c}}^{\max}\}$, $\alpha^{(i)}_0$ columns of weight 0 and $\alpha^{(i)}_2$ columns of weight 2. The number of ones in the matrix $\boldsymbol{M}_1$ is $f_1 = \sum_{j=1}^{d_{\mathsf{v}}^{\max}} \sum_{k=\lfloor \frac{j}{2} \rfloor + 1}^{j} k\beta^{(j)}_k = \sum_{i=1}^{d_{\mathsf{c}}^{\max}} 2\alpha^{(i)}_2$. From Theorem 1, the cardinality of $\mathcal{M}_1$, for $\delta_1 > 0$, is given by

$$|\mathcal{M}_1| = \frac{f_1!}{\prod_{i=1}^{d_{\mathsf{c}}^{\max}} (2!)^{\alpha^{(i)}_2} \prod_{j=1}^{d_{\mathsf{v}}^{\max}} \prod_{k=\lfloor \frac{j}{2} \rfloor + 1}^{j} (k!)^{\beta^{(j)}_k}} \times (1 + o(n^{-1+\delta_1})) \times \exp\left[-\frac{1}{2f_1}\left(\sum_{j=1}^{d_{\mathsf{v}}^{\max}} \sum_{k=\lfloor \frac{j}{2} \rfloor + 1}^{j} (k-1)k\beta^{(j)}_k\right)\right].$$

Consider now the matrix $\boldsymbol{M}_2$. For each $j \in \{1, \ldots, d_{\mathsf{v}}^{\max}\}$, there are $\beta^{(j)}_k$ rows of weight $j-k$ and all columns have weight 1. The number of ones in $\boldsymbol{M}_2$ is given by $f_2 = \sum_{j=1}^{d_{\mathsf{v}}^{\max}} \sum_{k=\lfloor \frac{j}{2} \rfloor + 1}^{j} (j-k)\beta^{(j)}_k = \sum_{i=1}^{d_{\mathsf{c}}^{\max}} \alpha^{(i)}_1 = \gamma n$. The cardinality of $\mathcal{M}_2$, for $\delta_2 > 0$, is then

$$|\mathcal{M}_2| = \frac{(\gamma n)!}{\prod_{j=1}^{d_{\mathsf{v}}^{\max}} \prod_{k=\lfloor \frac{j}{2} \rfloor + 1}^{j} ((j-k)!)^{\beta^{(j)}_k}} (1 + o(n^{-1+\delta_2})).$$

Note that f_1+f_2 is the total number of ones in the submatrix $[\boldsymbol{M}_1 \mid \boldsymbol{M}_2]$, which is equal to $n\tilde{\theta}$, where $\tilde{\theta}=\sum_j j\theta_j$. Thus,we have $\sum_{i=1}^{d_{\mathsf{c}}^{\max}}\sum_{k=0}^{2} k\alpha_k^{(i)} = n\tilde{\theta}$.

The matrix $\boldsymbol{M}_3$, has $n(\Lambda_j-\theta_j)$ rows of weight j for each $j\in\{1,\dots,d_{\mathsf{v}}^{\max}\}$ and $\alpha_k^{(i)}$ columns of weight $i-k$, where $k\in\{0,1,2\}$ and $i\in\{1,\dots,d_{\mathsf{c}}^{\max}\}$. The number of ones in $\boldsymbol{M}_3$ is given by $f_3 = n\sum_{j=1}^{d_{\mathsf{v}}^{\max}} j(\Lambda_j-\theta_j) = \sum_{i=1}^{d_{\mathsf{c}}^{\max}}\sum_{k=0}^{2}\alpha_k^{(i)}(i-k) = N-n\tilde{\theta}$. From Theorem 1, the cardinality of $\mathcal{M}_3$ is shown in (4) for $\delta_3>0$.

We define for $i=1,\dots,d_{\mathsf{c}}^{\max}$, $\tilde{\boldsymbol{\alpha}}^{(i)}=\boldsymbol{\alpha}^{(i)}/n$ and for $j=1,\dots,d_{\mathsf{v}}^{\max}$, $\tilde{\boldsymbol{\beta}}^{(j)}=\boldsymbol{\beta}^{(j)}/n$, where $\boldsymbol{\alpha}^{(i)}=(\alpha_0^{(i)},\alpha_1^{(i)},\alpha_2^{(i)})$ and $\boldsymbol{\beta}^{(j)}=(\beta^{(j)}_{\lfloor\frac{j}{2}\rfloor+1},\dots,\beta_j^{(j)})$. The cardinality of $\mathcal{M}$ can be expressed as in (5)

$$|\mathcal{M}| = \sum_{\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}}\prod_{i=1}^{d_{\mathsf{c}}^{\max}}\binom{n\xi \mathrm{P}_i}{n\tilde{\alpha}_0^{(i)},n\tilde{\alpha}_1^{(i)},n\tilde{\alpha}_2^{(i)}}\times \prod_{j=1}^{d_{\mathsf{v}}^{\max}}\binom{n\theta_j}{n\tilde{\beta}^{(j)}_{\lfloor\frac{j}{2}\rfloor+1},\dots,n\tilde{\beta}_j^{(j)}}|\mathcal{M}_1||\mathcal{M}_2||\mathcal{M}_3| \tag{5}$$

where $\xi=m/n$ and the sum is over the vectors $\tilde{\boldsymbol{\alpha}}=(\tilde{\boldsymbol{\alpha}}^{(1)},\dots,\tilde{\boldsymbol{\alpha}}^{(d_{\mathsf{c}}^{\max})})$ and $\tilde{\boldsymbol{\beta}}=(\tilde{\boldsymbol{\beta}}^{(1)},\dots,\tilde{\boldsymbol{\beta}}^{(d_{\mathsf{v}}^{\max})})$ that satisfy

$$\sum_{k=0}^{2}\tilde{\alpha}_k^{(i)} = \xi\mathrm{P}_i,\quad i=1,\dots,d_{\mathsf{c}}^{\max} \tag{7}$$

$$\sum_{i=1}^{d_{\mathsf{c}}^{\max}}\tilde{\alpha}_1^{(i)} = \gamma,\quad \sum_{i=1}^{d_{\mathsf{c}}^{\max}}\sum_{k=0}^{2}k\tilde{\alpha}_k^{(i)} = \tilde{\theta}, \tag{8}$$

$$\sum_{k=\lfloor\frac{j}{2}\rfloor+1}^{j}\tilde{\beta}_k^{(j)} = \theta_j,\quad j=1,\dots,d_{\mathsf{v}}^{\max}, \tag{9}$$

$$\sum_{j=1}^{d_{\mathsf{v}}^{\max}}\sum_{k=\lfloor\frac{j}{2}\rfloor+1}^{j}k\tilde{\beta}_k^{(j)} = 2\sum_{i=1}^{d_{\mathsf{c}}^{\max}}\tilde{\alpha}_2^{(i)}. \tag{10}$$

For each $i\in\{1,\dots,d_{\mathsf{c}}^{\max}\}$ and for each $j\in\{1,\dots,d_{\mathsf{v}}^{\max}\}$, we define the probability vectors $\boldsymbol{p}^{(i)}=(p_0^{(i)},p_1^{(i)},p_2^{(i)})$ and $\boldsymbol{z}^{(j)}=(z^{(j)}_{\lfloor\frac{j}{2}\rfloor+1},\dots,z_j^{(j)})$ with

$$p_k^{(i)} = \frac{1}{U_i}\binom{i}{k},\qquad U_i=\sum_{k=0}^{2}\binom{i}{k},$$
$$z_k^{(j)} = \frac{1}{U'_j}\binom{j}{k},\qquad U'_j=\sum_{k=\lfloor\frac{j}{2}\rfloor+1}^{j}\binom{j}{k}. \tag{11}$$

The normalized logarithmic asymptotic distribution of $\mathcal{A}^{\mathrm{E}}_{a,b}$ for $a=\theta n$ and $b=\gamma n$ for $\mathscr{C}_n^{\Lambda,\mathrm{P}}$ is defined as

$$G_{\mathrm{E}}^{\Lambda,\mathrm{P}}(\theta,\gamma) := \lim_{n\to\infty}\frac{1}{n}\ln\left(P_{\mathrm{E}}^{\Lambda,\mathrm{P}}(\theta n,\gamma n)\right) \tag{12}$$

where $P_{\mathrm{E}}^{\Lambda,\mathrm{P}}(a,b)$ is the average number of size (a,b) EASs in the Tanner graph of a code drawn randomly from the ensemble $\mathscr{C}_n^{\Lambda,\mathrm{P}}$. We have

$$P_{\mathrm{E}}^{\Lambda,\mathrm{P}}(\theta n,\gamma n) = \sum_{\boldsymbol{\theta}}\prod_{j=1}^{d_{\mathsf{v}}^{\max}}\binom{n\Lambda_j}{n\theta_j}\frac{|\mathcal{M}|}{|\mathcal{H}_{m,n}^{\boldsymbol{R},\boldsymbol{L}}|} \doteq \sum_{\boldsymbol{\theta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}}\exp(nT(\boldsymbol{\theta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}})) \tag{13}$$

where $|\mathcal{H}_{m,n}^{\boldsymbol{R},\boldsymbol{L}}|$ and $|\mathcal{M}|$ are given in (3) and (5) and $T(\boldsymbol{\theta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}})$ is defined in (6). From Lemma 1 and by using the Lagrangian multiplier method, it can be shown that the normalized logarithmic asymptotic distribution $G_{\mathrm{E}}^{\Lambda,\mathrm{P}}(\theta,\gamma)$ can be expressed as shown in (14), where $A_1,A_2,A_3,A_4,\theta^\star$ satisfy

$$A_1A_2\sum_{i=1}^{d_{\mathsf{c}}^{\max}}\frac{i\mathrm{P}_i}{1+iA_1A_2+\binom{i}{2}A_2^2} = \frac{\gamma}{\xi},\quad \sum_{j=1}^{d_{\mathsf{v}}^{\max}}\theta_j^\star=\theta$$

$$\sum_{i=1}^{d_{\mathsf{c}}^{\max}}\frac{\mathrm{P}_i(iA_1A_2+i(i-1)A_2^2)}{1+iA_1A_2+\binom{i}{2}A_2^2} = \frac{\tilde{\theta}^\star}{\xi},\quad \sum_{j=1}^{d_{\mathsf{v}}^{\max}}j\theta_j^\star=\tilde{\theta}^\star$$

$$\sum_{j=1}^{d_{\mathsf{v}}^{\max}}\theta_j\frac{\sum_{k=\lfloor\frac{j}{2}\rfloor+1}^{j}\binom{j}{k}kA_3^k}{\sum_{k'=\lfloor\frac{j}{2}\rfloor+1}^{j}\binom{j}{k'}A_3^{k'}} = \tilde{\theta}^\star-\gamma$$

$$|\mathcal{M}_3| = \frac{(N-n\tilde{\theta})!}{\prod_{i=1}^{d_{\mathsf{c}}^{\max}}\prod_{k=0}^{2}((i-k)!)^{\alpha_k^{(i)}}\prod_{j=1}^{d_{\mathsf{v}}^{\max}}(j!)^{n(\Lambda_j-\theta_j)}}\exp\left[-\frac{n}{2f_3^2}\left(\sum_{j=1}^{d_{\mathsf{v}}^{\max}}(\Lambda_j-\theta_j)j(j-1)\right)\left(\sum_{i=1}^{d_{\mathsf{c}}^{\max}}\sum_{k=0}^{2}(i-k)(i-k-1)\alpha_k^{(i)}\right)\right]\times (1+o(n^{-1+\delta_3})). \tag{4}$$

$$T(\boldsymbol{\theta},\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\ln\left(\frac{\prod_{j=1}^{d_{\mathsf{v}}^{\max}}\binom{n\Lambda_j}{n\theta_j}}{\binom{N}{n\tilde{\theta}}\binom{n\tilde{\theta}}{n\gamma}}\prod_{i=1}^{d_{\mathsf{c}}^{\max}}U_i^{\mathrm{P}_i\xi n}\binom{\mathrm{P}_i\xi n}{n\tilde{\alpha}_0^{(i)},n\tilde{\alpha}_1^{(i)},n\tilde{\alpha}_2^{(i)}}\prod_{k=0}^{2}(p_k^{(i)})^{n\tilde{\alpha}_k^{(i)}}\prod_{j=1}^{d_{\mathsf{v}}^{\max}}U_j'^{n\theta_j}\binom{n\theta_j}{n\tilde{\beta}^{(j)}_{\lfloor\frac{j}{2}\rfloor+1},\dots,n\tilde{\beta}_j^{(j)}}\prod_{k=\lfloor\frac{j}{2}\rfloor+1}^{j}(z_k^{(j)})^{n\tilde{\beta}_k^{(j)}}\right) \tag{6}$$

$$G_{\mathrm{E}}^{\Lambda,\mathrm{P}}(\theta,\gamma) = -\bar{d}_{\mathrm{v}} H\left(\frac{\tilde{\theta}^{\star}}{\bar{d}_{\mathrm{v}}}, 1-\frac{\tilde{\theta}^{\star}}{\bar{d}_{\mathrm{v}}}\right) - \tilde{\theta}^{\star} H\left(\frac{\gamma}{\tilde{\theta}^{\star}}, 1-\frac{\gamma}{\tilde{\theta}^{\star}}\right) + \sum_{j=1}^{d_{\mathrm{v}}^{\max}} \theta_j^{\star} \ln\left(\sum_{k=\lfloor \frac{j}{2} \rfloor+1}^{j} \binom{j}{k} A_3^k\right) + \sum_{i=1}^{d_{\mathrm{c}}^{\max}} \xi \mathrm{P}_i \ln\left(1 + iA_1A_2 + \binom{i}{2} A_2^2\right) - \gamma \ln(A_1/A_3) - \tilde{\theta}^{\star} \ln(A_2A_3) + \sum_{j=1}^{d_{\mathrm{v}}^{\max}} \Lambda_j H\left(\frac{\theta_j}{\Lambda_j}, 1-\frac{\theta_j}{\Lambda_j}\right). \tag{14}$$

with

$$\theta_j^{\star} = \frac{\Lambda_j \sum_{k=\lfloor \frac{j}{2} \rfloor+1}^{j} \binom{j}{k} A_3^k}{\sum_{k=\lfloor \frac{j}{2} \rfloor+1}^{j} \binom{j}{k} A_3^k + A_4 \left(A_2 A_3 \frac{\bar{d}_{\mathrm{v}} - \tilde{\theta}^{\star}}{\tilde{\theta}^{\star} - \gamma}\right)^j}$$

and $H(p_1, \ldots, p_N) = -\sum_i p_i \ln(p_i)$, with $\sum_i p_i = 1$, denotes the entropy function.

B. General Absorbing Sets

We use the same notation of subsection III-A. The asymptotic cardinalities of $\mathcal{M}_l$ for $l = 1, 2, 3$ are given in (15), (16) and (17), where we omit the exponential terms in Theorem 1. We have $\tilde{q} = \frac{1}{n} \sum_{i=1}^{d_{\mathrm{c}}^{\max}} \sum_{\substack{k=0 \\ k \text{ even}}}^{i} k\alpha_k^{(i)} = \frac{1}{n} \sum_{j=1}^{d_{\mathrm{v}}^{\max}} \sum_{k=\lfloor \frac{j}{2} \rfloor+1}^{j} k\beta_k^{(j)}$

$$|\mathcal{M}_1| \doteq \frac{(n\tilde{q})!}{\prod_{i=1}^{d_{\mathrm{c}}^{\max}} \prod_{\substack{k=0 \\ k \text{ even}}}^{i} (k!)^{\alpha_k^{(i)}} \prod_{j=1}^{d_{\mathrm{v}}^{\max}} \prod_{k=\lfloor \frac{j}{2} \rfloor+1}^{j} (k!)^{\beta_k^{(j)}}} \tag{15}$$

$$|\mathcal{M}_2| \doteq \frac{(n\tilde{\theta} - n\tilde{q})!}{\prod_{i=1}^{d_{\mathrm{c}}^{\max}} \prod_{\substack{k=0 \\ k \text{ odd}}}^{i} (k!)^{\alpha_k^{(i)}} \prod_{j=1}^{d_{\mathrm{v}}^{\max}} \prod_{k=\lfloor \frac{j}{2} \rfloor+1}^{j} ((j-k)!)^{\beta_k^{(j)}}} \tag{16}$$

$$|\mathcal{M}_3| \doteq \frac{(N - n\tilde{\theta})!}{\prod_{i=1}^{d_{\mathrm{c}}^{\max}} \prod_{k=0}^{i} ((i-k)!)^{\alpha_k^{(i)}} \prod_{j=1}^{d_{\mathrm{v}}^{\max}} (j!)^{n(\Lambda_j - \theta_j)}}. \tag{17}$$

We extend the probability vector defined in (11) to $\boldsymbol{p}^{(i)} = (p_0^{(i)}, \ldots, p_i^{(i)})$ with $p_k^{(i)} = \binom{i}{k}/U_i$ where $U_i = \sum_{k=0}^{i} \binom{i}{k}$. The normalized logarithmic asymptotic distribution of $\mathcal{A}_{a,b}$ for $a = \theta n$ and $b = \gamma n$ for $\mathscr{C}_n^{\Lambda,\mathrm{P}}$ is defined as

$$G^{\Lambda,\mathrm{P}}(\theta,\gamma) := \lim_{n\to\infty} \frac{1}{n} \ln\left(P^{\Lambda,\mathrm{P}}(\theta n, \gamma n)\right) \tag{18}$$

where $P^{\Lambda,\mathrm{P}}(a,b)$ is the average number of size (a,b) ASs in the Tanner graph of a code drawn randomly from the ensemble $\mathscr{C}_n^{\Lambda,\mathrm{P}}$. We proceed in the same manner as in subsection III-A and get the expression in (19).

IV. Numerical Results

In this section, we evaluate the expressions of the normalized logarithmic asymptotic distribution of $\mathcal{A}_{a,b}^{\mathrm{E}}$ and $\mathcal{A}_{a,b}$ derived in the previous section. In Fig 1 and 2, we fix the ratio $\Delta = \gamma/\theta$ and compute $G_{\mathrm{E}}^{\Lambda,\mathrm{P}}(\theta, \Delta\theta)$, $G^{\Lambda,\mathrm{P}}(\theta, \Delta\theta)$ for the rate $1/2$ ensembles $\mathscr{C}_n^{\Lambda^{(1)},\mathrm{P}^{(1)}}$ and $\mathscr{C}_n^{\Lambda^{(2)},\mathrm{P}^{(2)}}$ with $\Lambda^{(1)}(x) = 0.5x^3 + 0.5x^4$, $\mathrm{P}^{(1)}(x) = x^7$, $\Lambda^{(2)}(x) = 0.5x^4 + 0.5x^5$, $\mathrm{P}^{(2)}(x) = x^9$ and $\Delta \in \{0.005, 0.05, 0.1, 0.5, 1\}$. Following [15], for a fixed Δ, the second zero crossing of $(G_{\mathrm{E}}^{\Lambda,\mathrm{P}}(\theta, \Delta\theta))$ $G^{\Lambda,\mathrm{P}}(\theta, \Delta\theta)$ (the first one is zero), if it exists, is called the typical Δ-(elementary) absorbing set number. We denote by $(d_{\min}^{\mathrm{EAS}})$ $d_{\min}^{\mathrm{AS}}$ the Δ-(elementary) absorbing set number, which is the size of the smallest (elementary) absorbing set with $\Delta = b/a$. Having a strictly positive typical Δ-(elementary) absorbing set number is a desired property of the LDPC code ensemble. We can observe that the typical Δ-(elementary) absorbing set numbers decrease as Δ increases. We also remark that $\mathscr{C}_n^{\Lambda^{(2)},\mathrm{P}^{(2)}}$ has better absorbing set properties than $\mathscr{C}_n^{\Lambda^{(1)},\mathrm{P}^{(1)}}$.

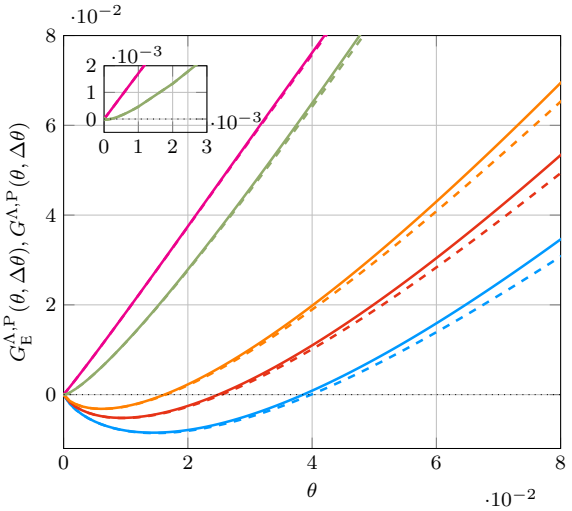


Fig. 1. Normalized logarithmic asymptotic distribution of elementary (- - -, - - -, - - -, - - -, - - -) and general absorbing sets (——, ——, ——, ——, ——) for $\Lambda^{(1)}(x) = 0.5x^3 + 0.5x^4$, $\mathrm{P}^{(1)}(x) = x^7$ and $\Delta = 1$ (——), $\Delta = 0.5$ (——), $\Delta = 0.1$ (——), $\Delta = 0.05$ (——), $\Delta = 0.005$ (——).

V. Conclusion

We derived asymptotic distributions of elementary and general absorbing sets for unstructured LDPC code ensembles. The method is similar to the approaches for the trapping sets proposed in [14] and for absorbing sets in [5], but [5] does not consider irregular LDPC ensembles. Moreover, the simplified expressions derived in [5] are valid only for regular LDPC code ensembles with VN degree 3 and 4. Following [15], we defined the typical Δ-(elementary) absorbing set number, which can be used to evaluate the absorbing set properties of an LDPC code ensemble.

$$G^{\Lambda,\mathrm{P}}(\theta,\gamma) = \sum_{i=1}^{d_{\mathrm{c}}^{\max}} \xi \mathrm{P}_i \ln\left((1+B_2B_3)^i + (1-B_2B_3)^i + B_1[(1+B_2)^i - (1-B_2)^i]\right) - \xi\ln(2) - \gamma\ln(B_1) + \tilde{\theta}^\star \ln\left(\frac{\tilde{\theta}^\star - \tilde{q}^\star}{B_2\tilde{\theta}^\star}\right)$$
$$+ \sum_{j=1}^{d_{\mathrm{v}}^{\max}} \theta_j^\star \ln\left(\sum_{k=\lfloor \frac{j}{2}\rfloor+1}^{j} \binom{j}{k} \frac{1}{B_3^k} \left(\frac{\tilde{q}^\star}{\tilde{\theta}^\star - \tilde{q}^\star}\right)^k\right) - \bar{d}_{\mathrm{v}} H\left(\frac{\tilde{\theta}^\star}{\bar{d}_{\mathrm{v}}}, 1 - \frac{\tilde{\theta}^\star}{\bar{d}_{\mathrm{v}}}\right) + \sum_{j=1}^{d_{\mathrm{v}}^{\max}} \Lambda_j H\left(\frac{\theta_j}{\Lambda_j}, 1 - \frac{\theta_j}{\Lambda_j}\right) \tag{19}$$

where $\theta_j^\star = \Lambda_j \left(1 + B_4 \left(B_2 \frac{\bar{d}_{\mathrm{v}} - \tilde{\theta}^\star}{\tilde{\theta}^\star - \tilde{q}^\star}\right)^j \frac{1}{\sum_{k=\lfloor \frac{j}{2}\rfloor+1}^{j} \binom{j}{k} \frac{1}{B_3^k} \left(\frac{\tilde{q}^\star}{\tilde{\theta}^\star - \tilde{q}^\star}\right)^k}\right)^{-1}$ $\tilde{q}^\star = \xi B_2 B_3 \sum_{i=1}^{d_{\mathrm{c}}^{\max}} i\mathrm{P}_i \frac{(1+B_2B_3)^{i-1} - (1-B_2B_3)^{i-1}}{(1+B_2B_3)^i + (1-B_2B_3)^i + B_1[(1+B_2)^i - (1-B_2)^i]}$

and $B_1, B_2, B_3, B_4, \tilde{\theta}$ are the positive roots of

$$B_1 \sum_{i=1}^{d_{\mathrm{c}}^{\max}} \mathrm{P}_i \frac{(1+B_2)^i - (1-B_2)^i}{(1+B_2B_3)^i + (1-B_2B_3)^i + B_1[(1+B_2)^i - (1-B_2)^i]} = \frac{\gamma}{\xi}, \quad \sum_{j=1}^{d_{\mathrm{v}}^{\max}} \theta_j^\star = \theta, \quad \sum_{j=1}^{d_{\mathrm{v}}^{\max}} j\theta_j^\star = \tilde{\theta}^\star$$

$$B_2 \sum_{i=1}^{d_{\mathrm{c}}^{\max}} i\mathrm{P}_i \frac{B_3[(1+B_2B_3)^{i-1} - (1-B_2B_3)^{i-1}] + B_1[(1+B_2)^{i-1} + (1-B_2)^{i-1}]}{(1+B_2B_3)^i + (1-B_2B_3)^i + B_1[(1+B_2)^i - (1-B_2)^i]} = \frac{\tilde{\theta}^\star}{\xi}$$

$$\xi B_2 B_3 \sum_{i=1}^{d_{\mathrm{c}}^{\max}} i\mathrm{P}_i \frac{(1+B_2B_3)^{i-1} - (1-B_2B_3)^{i-1}}{(1+B_2B_3)^i + (1-B_2B_3)^i + B_1[(1+B_2)^i - (1-B_2)^i]} = \sum_{j=1}^{d_{\mathrm{v}}^{\max}} \theta_j \frac{\sum_{k=\lfloor \frac{j}{2}\rfloor+1}^{j} \binom{j}{k} \frac{k}{B_3^k} \left(\frac{\tilde{q}^\star}{\tilde{\theta}^\star - \tilde{q}^\star}\right)^k}{\sum_{k'=\lfloor \frac{j}{2}\rfloor+1}^{j} \binom{j}{k'} \frac{1}{B_3^{k'}} \left(\frac{\tilde{q}^\star}{\tilde{\theta}^\star - \tilde{q}^\star}\right)^{k'}}$$

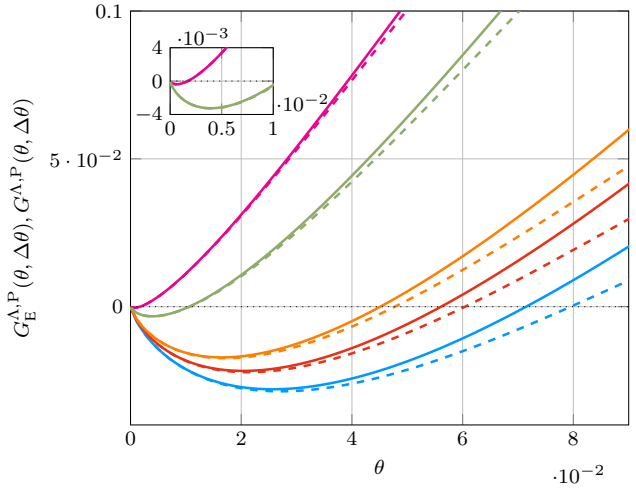


Fig. 2. Normalized logarithmic asymptotic distribution of elementary (– – –, – – –, – – –, – – –) and general absorbing sets (———, ———, ———, ———, ———) for $\Lambda^{(2)}(x) = 0.5x^4 + 0.5x^5$, $\mathrm{P}^{(2)}(x) = x^9$ and $\Delta = 1$ (———), $\Delta = 0.5$ (———), $\Delta = 0.1$ (———), $\Delta = 0.05$ (———), $\Delta = 0.005$ (———).

References

[1] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[2] C. Di, D. Proietti, I. Telatar, T. Richardson, and R. Urbanke, "Finite-length analysis of low-density parity-check codes on the binary erasure channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1570–1579, Jun. 2002.

[3] D. J. MacKay and M. S. Postol, "Weaknesses of Margulis and Ramanujan-Margulis low-density parity-check codes," *Electronic Notes in Theoretical Computer Science*, vol. 74, pp. 97–104, 2003.

[4] T. Richardson, "Error floors of LDPC codes," in *Proc. Allerton Conf. on Communication, Control and Computing*, Monticello, USA, Oct. 2003.

[5] B. Amiri, C. Lin, and L. Dolecek, "Asymptotic distribution of absorbing sets and fully absorbing sets for regular sparse code ensembles," *IEEE Trans. Commun.*, vol. 61, no. 2, pp. 455–464, February 2013.

[6] E. Kurtas and B. Vasi, *Advanced error control techniques for data storage systems*. New York: CRC Press, 2005.

[7] B. P. Smith, A. Farhood, A. Hunt, F. R. Kschischang, and J. Lodge, "Staircase Codes: FEC for 100 Gb/s OTN," *J. Lightw. Technol.*, vol. 30, no. 1, pp. 110–117, Jan. 2012.

[8] L. Dolecek, Z. Zhang, V. Anantharam, M. J. Wainwright, and B. Nikolic, "Analysis of absorbing sets and fully absorbing sets of array-based LDPC codes," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 181–201, Jan. 2009.

[9] S. K. Chilappagari, S. Sankaranarayanan, and B. Vasić, "Error floors of LDPC codes on the binary symmetric channel," in *IEEE Int. Conf. Commun.*, vol. 3, June 2006, pp. 1089–1094.

[10] S. K. Chilappagari and B. Vasić, "Error-correction capability of column-weight-three LDPC codes," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2055–2061, May 2009.

[11] B. Vasić, S. K. Chilappagari, D. V. Nguyen, and S. K. Planjery, "Trapping set ontology," in *Proc. Allerton Conf. on Commun., Control and Computing*, Monticello, USA, Oct. 2009.

[12] A. Orlitsky, K. Viswanathan, and J. Zhang, "Stopping set distribution of LDPC code ensembles," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 929–953, Mar. 2005.

[13] S. Abu-Surra, D. Divsalar, and W. E. Ryan, "Enumerators for protograph-based ensembles of LDPC and generalized LDPC codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 858–886, Feb 2011.

[14] O. Milenkovic, E. Soljanin, and P. Whiting, "Asymptotic spectra of trapping sets in regular and irregular LDPC code ensembles," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 39–55, Jan 2007.

[15] S. Abu-Surra, W. Ryan, and D. Divsalar, "Ensemble trapping set enumerators for protograph-based LDPC codes," *Proc. 45th Annual Allerton Conf. on Commun., Control and Computing*, pp. 201–210, Sep 2007.

[16] M. Ferrari, A. Tomasoni, L. Barletta, and S. Bellini, "Absorbing sets of generalized LDPC codes," in *Proc. International Zurich Seminar on Information and Communication*, Zurich, Switzerland, 2018, pp. 108–112.

[17] S. Litsyn and V. Shevelev, "On ensembles of low-density parity-check codes: asymptotic distance distributions," *IEEE Trans. Inf. Theory*, vol. 48, no. 4, pp. 887–908, Apr 2002.

A Recursive Algorithm for Quantizer Design for Binary-Input Discrete Memoryless Channels

Mehdi Dabirnia
Universitat Pompeu Fabra
mehdi.dabirnia@upf.edu

Alfonso Martinez
Universitat Pompeu Fabra
alfonso.martinez@ieee.org

Albert Guillén i Fàbregas
ICREA and Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—The quantization of the outputs of a binary-input discrete memoryless channel is considered. A new recursive method for finding all optimal quantizers for all output cardinalities is proposed. Two different versions of the newly proposed method for top-down and bottom-up approaches are developed which provide an improved understanding of the quantization problem under consideration. Also, an efficient algorithm based on dynamic programing is proposed and shown to have a comparable complexity with the state of the art.**

I. Introduction

Quantization has practical applications in hardware implementations of communication systems, e.g., from channel output quantization to message passing decoders [1] and polar code construction [2]. In such applications, there is a trade-off between performance and complexity of the system represented by the number of quantization levels. Therefore, it is of interest to use as few quantization levels as possible while maintaining reliable communication with a given transmission rate.

Recently we studied channel output quantization from a mismatched-decoding perspective [3]. This study showed that the best mismatched decoder coincides with maximum likelihood decoding for the channel between the channel input and the quantizer output. This result supports the approach of optimizing the quantizer based on a performance metric for the quantized channel, e.g., mutual information [4] or error exponent [5].

Consider a discrete memoryless channel (DMC) followed by a quantizer at the output, as shown in Fig. 1. The channel input X takes values in $\mathcal{X} = \{1, \dots, J\}$ with probability distribution $p_x = \Pr(X = x)$, and the channel output Y takes values in $\mathcal{Y} = \{1, \dots, M\}$, with channel transition probabilities $W_{y|x} = \Pr(Y = y|X = x)$. The channel output is quantized to $Z^{(K)}$, which takes values in $\mathcal{Z}^{(K)} = \{z_1^{(K)}, \dots, z_K^{(K)}\}$, by a possibly stochastic quantizer $Q_{z|y} = \Pr(Z^{(K)} = z|Y = y)$. The conditional probability distribution of the quantizer output given the channel input is $T_{z|x} = \Pr(Z^{(K)} = z|X = x) = \sum_{y\in\mathcal{Y}} Q_{z|y} W_{y|x}$.

The mutual information between X and $Z^{(K)}$ is

$$I(X; Z^{(K)}) = \sum_{z\in\mathcal{Z}^{(K)}} \sum_{x\in\mathcal{X}} p_x T_{z|x} \log \frac{T_{z|x}}{\sum_{x'} p_{x'} T_{z|x'}}. \quad (1)$$

This work has been funded in part by the European Research Council under grant 725411, and by the Spanish Ministry of Economy and Competitiveness under grant TEC2016-78434-C3-1-R.

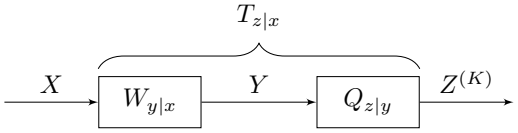


Fig. 1: A discrete memoryless channel followed by a quantizer.

Let us denote the set of all possible quantizers Q with K outputs, including stochastic quantizers, with $\mathcal{Q}^{(K)}$. We formulate the quantizer optimization as follows: for a given constant $0 \leq \alpha \leq 1$, we want to find an optimal quantizer Q_α^* with the smallest cardinality K from the set $\mathcal{S}$ defined as

$$\mathcal{S} \triangleq \{Q \in \mathcal{Q}^K : 1 \leq K \leq M, I(X; Z^{(K)}) \geq \alpha I(X; Y)\}. \quad (2)$$

The optimal quantizer Q_α^* preserves at least an α-portion of the original mutual information with the smallest number of quantization levels K.

II. Background and Contribution

For a fixed output cardinality K, Kurkoski and Yagi showed that there is a deterministic quantizer that maximizes the mutual information (1) between channel input and quantized output [4]. Therefore, considering only deterministic quantizers is sufficient to find the optimal quantizer Q_α^*. A deterministic quantizer Q maps each output y to only one quantized output $z_k^{(K)}$, $Q : \{1, \dots, M\} \to \{z_1^{(K)}, \dots, z_K^{(K)}\}$, therefore, the corresponding probabilistic map $Q_{z|y}$ takes only values 0 or 1. We define the pre-image of $z_k^{(K)}$ as

$$\mathcal{A}(z_k^{(K)}) = \left\{y \in \mathcal{Y} : Q^{-1}(z_k^{(K)}) = y\right\} \quad (3)$$

which is the set of channel outputs that are mapped to $z_k^{(K)}$. Hence, the deterministic quantizer $Q_{z|y}$ partitions $\mathcal{Y}$ to K subsets $\{\mathcal{A}(z_1^{(K)}), \dots, \mathcal{A}(z_K^{(K)})\}$.

Let $P_{x|y} = \Pr(X = x|Y = y)$ be the posterior conditional probability distribution on the channel input which depends on the input distribution p_x and the channel conditional distribution $W_{y|x}$. For each channel output y, we define a vector v_y

$$v_y = \left[P_{1|y}, P_{2|y}, \dots, P_{J-1|y}\right] \quad (4)$$

with $v_y \in \mathcal{U} = [0, 1]^{J-1}$. Define an equivalent quantizer $\tilde{Q}$ on the vectors $\{v_1, \dots, v_M\}$ as $\tilde{Q}(v_y) = Q(y) = z$.

Kurkoski and Yagi in [4, Lemma 2], using the results of [6], study a condition to find an optimal quantizer $\tilde{Q}^*$. They show the existence of an optimal quantizer $\tilde{Q}^*$ for which two distinct preimages $\tilde{Q}^{*-1}(z)$ and $\tilde{Q}^{*-1}(z')$ are separated by a hyperplane in Euclidean space $\mathcal{U}$. Unfortunately, this condition does not offer a practical search method for quantizer design problem in general; however, as suggested in [4], it simplifies the problem for binary-input case.

The problem of finding Q^*_α can be tackled by either a *bottom-up* or *top-down* approach. The former starts with trivial partition into $K = M$ subsets, where each subset $\mathcal{A}(z_k^{(K)})$, $1 \le k \le K$ contains exactly one element of $\mathcal{Y}$. At each step, we decrease the cardinality K by one and design a quantizer with output size K. We stop when the corresponding mutual information goes below the desired threshold. The latter approach starts with the other trivial solution with single partition containing all the elements, i.e., $\mathcal{A}(z_1^{(1)}) = \mathcal{Y}$. At each step, we increment the cardinality K by one and design a quantizer with output size K. We stop when the corresponding mutual information reaches (or exceeds) the desired threshold. In both approaches, the quantizer design at each step can be performed either *recursively*, namely by starting from the result of previous step, or *independently*, which means that the design is performed independent of the previous step result.

An example of a recursive bottom-up approach is the *agglomerative information bottleneck* [7] which has been rediscovered multiple times in the literature with different names such as *greedy merging* or *greedy combining* [1], [2]. This algorithm iteratively reduces the cardinality by merging two outputs into a new single output. At each iteration, the greedy algorithm evaluates all the possible pairwise merges and selects the one that minimizes the mutual information loss. Although this algorithm finds the optimal pairwise merge at each step, it is globally suboptimal, since it fixes all the previously performed merges. This algorithm has complexity $O(M^2)$ for a bottom-up design, resulting in a quantizer for each cardinality $1 \le K \le M$.

As for the independent approach, several quantizer design algorithms from the literature can be utilized. For binary-input DMCs, Kurkoski and Yagi developed an algorithm based on dynamic programming that finds an optimal quantizer with time complexity $O(K(M-K)^2)$ [4]. Iwata and Ozawa [8] improved the complexity to $O(K(M-K))$ using the SMAWK algorithm. For the non-binary-input case, finding the optimal quantizer is an NP-hard problem [9], however several suboptimal algorithms are proposed in the literature. An example is KL-means quantizer [10] which is a variation of the K-means clustering algorithm by replacing Euclidean distance metric with Kullback-Leibler divergence. This algorithm has complexity $O(KMT)$ where T is the number of iterations that algorithm is run to converge to a local optimum. The complexity of top-down (or bottom-up) approach with independent design at each step is K (or $M-K$) times the complexity of a single-step run, respectively.

In this paper, we focus on the binary-input case and we propose a recursive method for quantization of binary-input DMCs that finds all the optimal quantizers. We develop two versions of the new method, one for top-down and the other for bottom-up approach. In addition, we propose an algorithm based on dynamic programming that has comparable complexity to the best known algorithm from the literature.

III. Optimal Recursive Quantizer

For the binary-input case, the posterior conditional probabilities $v_y = P(1|y)$ are in one-dimensional space $\mathcal{U} = [0,1]$. Denote the output probabilities by $\pi_i = \Pr(Y = i)$. We assume that the outputs are relabelled to satisfy

$$P(1|1) < P(1|2) < \cdots < P(1|M). \tag{5}$$

According to the [4, Lemma 3], there is an optimal quantizer Q^* such that preimages of the quantizer outputs consist of contiguous set of integers,

$$\mathcal{A}^*(z_k^{(K)}) = \{a_{k-1}^* + 1, \ldots, a_k^*\} \tag{6}$$

for $z_k^{(K)} \in \mathcal{Z}^{(K)}$, with $a_0^* = 0$ and $a_{k-1}^* < a_k^*$ and $a_K^* = M$. The a_k^*'s are optimal quantizer boundaries which satisfy

$$0 < a_1^* < a_2^* < \cdots < a_{K-1}^* < M. \tag{7}$$

Here we show that this condition is necessary for any optimal quantizer. Denote the mutual information loss corresponding to merging outputs j and l with $\Delta\imath(j,l)$ which is given by

$$\begin{aligned}\Delta\imath(j,l) = &\sum_{x\in\{1,2\}} \pi_j \Phi(P(x|j)) + \pi_l \Phi(P(x|l)) \\ &- (\pi_j + \pi_l)\Phi(P(x|y_{jl})),\end{aligned} \tag{8}$$

where $\Phi(x) = x\log(x)$.

Lemma 1. *For binary-input DMC, assuming that the outputs are relabelled to satisfy* (5)*, then for any choice of* $1 \le j < k < l \le M$ *at least one of the following is true,*

$$\begin{cases} \Delta\imath(j,k) < \Delta\imath(j,l) & \text{if } \frac{\pi_j}{\pi_l} \le \frac{v_l - v_k}{v_k - v_j}, \\ \Delta\imath(k,l) < \Delta\imath(j,l) & \text{if } \frac{\pi_j}{\pi_l} \ge \frac{v_l - v_k}{v_k - v_j}. \end{cases} \tag{9}$$

The proof is in the Appendix. Lemma 1 shows that for any quantizer that does not satisfy the condition in (6), there is another quantizer satisfying this condition that has a higher mutual information. Therefore, based on this necessary condition, the quantizer design reduces to searching for the optimal boundaries a_k^* as in (7).

A. Modified Greedy Merging

The greedy merging algorithm [1], [2] reduces the output cardinality by performing the best pairwise merge at each step. It finds the optimal single-step quantizer by a greedy search, i.e.,

$$Q_{\mathrm{m}}^{(i)} = \arg\min_{Q\in\mathcal{Q}_{\mathrm{m}}^{(i)}} I(X;Z^{(i+1)}) - I(X;Z^{(i)}), \tag{10}$$

where $\mathcal{Q}_{\mathrm{m}}^{(i)}$ is set of all possible single-step deterministic quantizers (pairwise merges) from $Z^{(i+1)}$ to $Z^{(i)}$.

In this section, we propose a new greedy algorithm which considers all pairwise merges and also another set of single-step quantizers which we denote them as *contractions*. A *contraction* is a single-step quantizer that consists of splits and merges. Next, we denote the definitions of split and merge and afterwards we define a *contraction*.

Definition 1 (Splitting an output)**.** *A quantizer output z_k with preimage $\mathcal{A}(z_k) = \{a_{k-1}+1, \ldots, a_k\}$ of size $b_k = |\mathcal{A}(z_k)| \geq 2$, splits into two non-empty parts z_{kL} (left) and z_{kR} (right) with preimages $\mathcal{A}(z_{kL}) = \{a_{k-1}+1, \ldots, s\}$ and $\mathcal{A}(z_{kR}) = \{s+1, \ldots, a_k\}$. This split can be done in $b_k - 1$ different ways, $a_{k-1} + 1 \leq s \leq a_k - 1$.*

Definition 2 (Merging an split output)**.** *An split output z_k with two non-empty parts z_{kL} (left) and z_{kR} (right) is merged as: 1- z_{kL} merges with z_{k-1} (or $z_{(k-1)R}$ if it has been split too) 2- z_{kR} merges with z_{k+1} (or $z_{(k+1)L}$)*

Contraction from K-level to $(K-1)$-level:

1) *Input: a K-level quantizer with output boundaries $\{a_1, a_2, \ldots, a_{K-1}\}$*
2) *Select a set of consecutive non-boundary outputs $\{z_j, z_{j+1}, \ldots, z_l\}$ with $j > 1$, $l < K$ and $b_k = |\mathcal{A}(z_k)| \geq 2$ for all $j \leq k \leq l$.*
3) *Split each z_k according to Definition 1. This step can be done in $\prod_{k=j}^{l}(b_k - 1)$ different ways.*
4) *Merge z_{kR} with $z_{(k+1)L}$ for all $j \leq k \leq l-1$, also merge z_{j-1} with z_{jL} and z_{lR} with z_{l+1}.*
5) *Output: a $(K-1)$-level quantizer with output boundaries $\{a'_1, \ldots, a'_{K-2}\}$ for which $a_{k-1} < a'_{k-1} < a_k$ for all $j \leq k \leq l-1$.*

Let us denote the set of all quantizers obtained by *contraction* as $\mathcal{Q}_{\mathrm{c}}^{(K-1)}$.

As an example to illustrate *contraction*, consider a quantizer with 3 outputs with preimages $\mathcal{A}(z_1) = \{1, \ldots, a_1\}$, $\mathcal{A}(z_2) = \{a_1+1, \ldots, a_2\}$ and $\mathcal{A}(z_3) = \{a_2+1, \ldots, M\}$. According to step 2 of *contraction*, the only possibility for a set of consecutive non-boundary outputs is $\{z_2\}$ if $b_2 = |\mathcal{A}(z_2)| \geq 2$. In step 3, we split z_2 into two parts $\mathcal{A}(z_{2\mathrm{L}}) = \{a_1+1, \ldots, s\}$ and $\mathcal{A}(z_{2\mathrm{R}}) = \{s+1, \ldots, a_2\}$ where $a_1+1 \leq s \leq a_2-1$. We merge $z_{2\mathrm{L}}$ with z_1 and $z_{2\mathrm{R}}$ with z_3 according to step 4. The output of this *contraction* is a quantizer with 2 outputs that has the boundary $a'_1 = s$. The set of all b_2-1 possible *contractions* for this example are specified by $a_1 + 1 \leq s \leq a_2 - 1$.

Modified greedy merging starts from the trivial solution with M outputs and at each step performs a greedy search over all possible *contractions* $\mathcal{Q}_{\mathrm{c}}^{(i)}$ and all pairwise merges $\mathcal{Q}_{\mathrm{m}}^{(i)}$, selecting the one with lowest mutual information loss. At each step it keeps all the quantizers that have the highest mutual information and uses them as a seed for the next step.

Theorem 1. *For the binary-input DMC, the modified greedy merging algorithm finds all optimal quantizers Q^* for all output cardinalities $1 \leq K \leq M$.*

Due to space limitations, we omit the proof.

B. Modified Greedy Splitting

Modified greedy splitting is a top-down algorithm that is the dual of modified greedy merging. It starts from the trivial solution with a single output and at each step it increases the output cardinality by one, performing a greedy search over all possible *expansions*. It keeps all the quantizers that have the highest mutual information at each step and uses them as a seed for next step. In the following we define an *expansion* which consists of splits and merges.

Assume that we have a K-level quantizer which is specified by its boundaries $\{a_1, a_2, \ldots, a_K\}$, we obtain a $(K+1)$-level quantizer by set of splits and merges according to following steps.

Expansion from K-level to $(K+1)$-level:

1) *Input: a K-level quantizer with output boundaries $\{a_1, a_2, \ldots, a_{K-1}\}$*
2) *Select a set of consecutive outputs $\{z_j, z_{j+1}, \ldots, z_l\}$ with $j \geq 1$, $l \leq K$ and $b_k = |\mathcal{A}(z_k)| \geq 2$ for all $j \leq k \leq l$.*
3) *Split each z_k according to Definition 1. This step can be done in $\prod_{k=j}^{l}(b_k - 1)$ different ways.*
4) *If the size of selected set in Step 2 is one, omit this otherwise merge z_{kR} with $z_{(k+1)L}$ for all $j \leq k \leq l-1$.*
5) *Output: a $(K+1)$-level quantizer with output boundaries $\{a'_1, \ldots, a'_K\}$ for which $a_{k-1} < a'_k < a_k$ for all $j \leq k \leq l$.*

Let us denote the set of all quantizers obtained by *expansions* as $\mathcal{Q}_{\mathrm{e}}^{(K+1)}$.

As an example to illustrate *expansion*, consider a quantizer with 2 outputs with preimages $\mathcal{A}(z_1) = \{1, \ldots, a_1\}$, $\mathcal{A}(z_2) = \{a_1+1, \ldots, M\}$. An *expansion* for this example can be obtained in two different ways. The first one is simply by splitting one of the outputs z_1 or z_2 which can be performed in $b_1 - 1$ and $b_2 - 1$ different ways. The second one is by splitting both z_1 and z_2 and merging $z_{1\mathrm{R}}$ with $z_{2\mathrm{L}}$. The latter can be performed in $(b_1-1)(b_2-1)$ different ways. The output of any such *expansion* is a quantizer with 3 outputs that has the boundaries $\{a'_1, a'_2\}$.

Theorem 2. *For the binary-input DMC, the modified greedy splitting finds all optimal quantizers Q^* for all output cardinalities $1 \leq K \leq M$.*

This theorem can be easily proved by showing the duality between *expansions* and *contractions* plus pairwise merges.

Note that the number of possible *contractions* and *expansions* increases polynomially as the number of outputs with large preimages increase. Therefore, the complexity of the modified greedy algorithms also grows polynomially. In the following we provide an algorithm based on dynamic programming which has quadratic complexity in the worst case.

C. Dynamic Programming Based Algorithm

This algorithm is a modified version of the Quantizer Design Algorithm [4] which is an instance of dynamic programming.

The assumption for this algorithm is that we already know the optimal K-level quantizer (which is specified by its boundaries $\{a_i\}_{i=0}^{K}$) and we want to find the optimal $(K+1)$-level quantizer employing the constraints imposed by *expansion* procedure on the resulting boundaries $\{a'_i\}_{i=0}^{K+1}$. The algorithm has a state value $S_z(y)$, which is the maximum partial mutual information when channel outputs 1 to y are quantized to quantizer outputs 1 to z. This can be computed recursively by conditioning on the state value at time index $z-1$:

$$S_z(a) = \max_{a'} \Big(S_{z-1}(a') + \imath(a' \to a) \Big), \tag{11}$$

where $\imath(a' \to a)$ is the contribution that the quantizer output $z = \{a' \to a\}$ makes to the mutual information. It is called partial mutual information and is given by

$$\imath(a' \to a) = \sum_{x \in \mathcal{X}} P_x \sum_{y=a'+1}^{a} P_{y|x} \log \frac{\sum_{y'=a'+1}^{a} P_{y'|x}}{\sum_{x'} \sum_{y'=a'+1}^{a} P_{y'|x'}}. \tag{12}$$

There are constraints imposed by the *expansion* procedure on the set of states a' that needs to be considered in the maximization in (11). These constraints have a key role in simplifying the original Quantizer Design Algorithm [4].

Splitting Algorithm

1) Inputs
   - Binary-input discrete memoryless channel $P_{y|x}$ relabelled to satisfy (5).
   - Input distribution P_x.
   - Set of boundaries $\{a_i\}_{i=0}^{K}$ corresponding to the optimal K-level quantizer.
2) Precompute the partial mutual information. For each $0 \le i \le K-1$,
   - For $a' = a_i + 1$ and for each $a \in a_i + 1, \dots, a_{i+1}$, compute $\imath(a' \to a)$ according to (12).
   - For each $a' \in \{a_i + 2, \dots, a_{i+1}\}$ and for each $a \in \{a_{i+1}, \dots, t\}$, (where $t = M$ for $i = K-1$ and $t = a_{i+2} - 1$ otherwise) compute $\imath(a' \to a)$ according to (12).
3) Recursion
   - $S_1(a) = \imath(1 \to a)$ for $a \in \{1, \dots, a_1\}$.
   - Store the local decision $h_1(a) = 0$ for $a \in \{1, \dots, a_1\}$.
   - For each $1 \le i \le K-1$,
     - Compute

$$S_{i+1}(a_i) = \max_{a'} S_i(a') + \imath(a' \to a_i),$$
$$h_{i+1}(a_i) = \arg\max_{a'} S_i(a') + \imath(a' \to a_i),$$

       where the maximization is over $a' \in \{a_{i-1} + 1, \dots, a_i - 1\}$.
     - For each $a \in \{a_i + 1, \dots, a_{i+1} - 1\}$ compute

$$S_{i+1}(a) = \max_{a'} S_i(a') + \imath(a' \to a),$$
$$h_{i+1}(a) = \arg\max_{a'} S_i(a') + \imath(a' \to a),$$

       where the maximization is over $a' \in \{a_{i-1} + 1, \dots, a_i\}$.
     - Compute

$$S_{i+1}(a_{i+1}) = S_i(a_i) + \imath(a_i \to a_{i+1}),$$
$$h_{i+1}(a_{i+1}) = a_i.$$

4) Find the optimal quantizer by traceback. Let $a^*_{K+1} = M$. For each $i \in \{K, K-1, \dots, 1\}$,

$$a^*_i = h_{i+1}(a^*_{i+1}).$$

Theorem 2 guarantees finding all the optimal quantizers at each step provided that the algorithm is run with all seeds from the previous step and that a tie-preserving implementation collects all locally optimal decisions and tracebacks.

Note that the dual of this algorithm can be developed for the bottom-up approach, based on the *contraction* procedure. Namely, with the assumption of already knowing the optimal K-level quantizer, all the optimal $(K-1)$-level quantizers are found using similar dynamic programming approach.

D. Complexity

The splitting algorithm developed here has complexity $O(M^2)$ in the worst case, and more generally it has complexity $O(\sum_{i=1}^{K} b_i b_{i+1})$ where $\sum_{i=1}^{K} b_i = M$. The worst case complexity is in the same order as the best known state of the art algorithm in [8].

E. Example: Additive White Gaussian Noise (AWGN) Channel

We consider a binary-input AWGN channel with equally likely ± 1 inputs and noise variance of $\sigma^2 = 0.5$. We first uniformly quantize the output of the AWGN channel y between -2 and 2 with $M = 1000$ levels. The natural order of the outputs of the resulting DMC satisfies (5). Later we apply the splitting algorithm to find a quantizer with minimum output levels which preserves $\alpha = 0.99$ of the mutual information of the original AWGN. Fig. 2 shows the quantization boundaries for the optimal quantizers (of underlying DMC) with 2 to 8 outputs. The results match with those obtained by the algorithm in [4]. We observe that the optimal quantizer with $K = 8$ outputs satisfies the mutual information constraint (Fig. 3).

Appendix A
Proof of Lemma 1

Let us denote the new output resulting from merging j and l as y'_{jl} and its conditional posterior probability as v_{jl}

$$v_{jl} = P_{1|y'_{jl}} = \frac{(\pi_j v_j + \pi_l v_l)}{\pi_j + \pi_l} \quad \to \quad \frac{\pi_j}{\pi_l} = \frac{v_l - v_{jl}}{v_{jl} - v_j} \tag{13}$$

$$\bar{v}_{jl} = P_{2|y'_{jl}} = 1 - v_{jl} \quad \to \quad \frac{\pi_j}{\pi_l} = \frac{\bar{v}_{jl} - \bar{v}_l}{\bar{v}_j - \bar{v}_{jl}}. \tag{14}$$

Now let us assume that

$$\frac{\pi_j}{\pi_l} = \frac{v_l - v_{jl}}{v_{jl} - v_j} \ge \frac{v_l - v_k}{v_k - v_j}, \tag{15}$$

therefore, $v_{jl} \le v_k$.

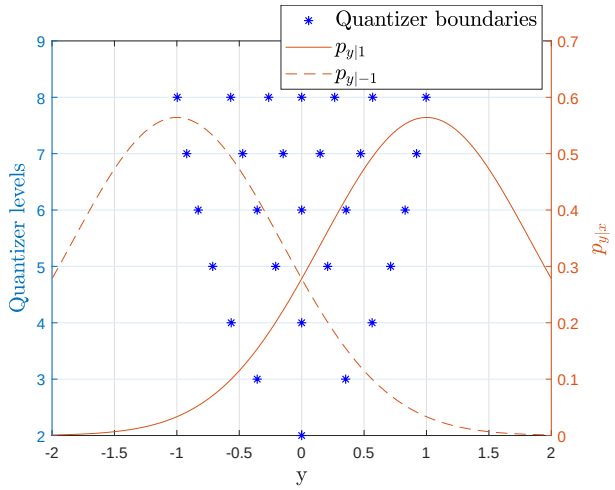


Fig. 2: Optimal quantization of a DMC derived from a finely quantized AWGN channel with $M = 1000$ to $K = 2$ to $K = 8$ levels using the top-down splitting algorithm.

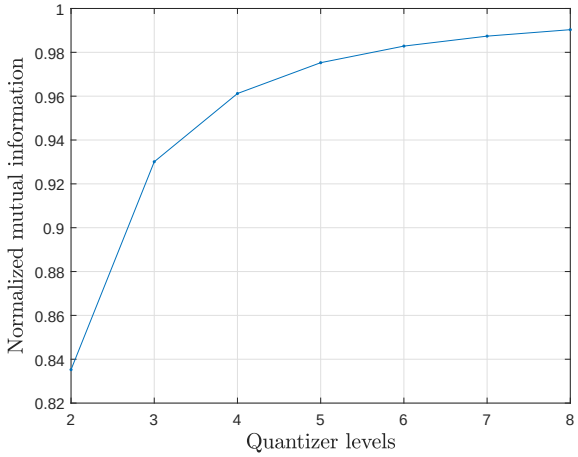


Fig. 3: Normalized mutual information of the Optimal quantizers with $K = 2$ to $K = 8$ levels.

With this assumption, we will show that both terms of the summation in (8) is larger for a (j, l) merge than for a (k, l) merge.

$$\Delta\imath_1(j,l) = \pi_j\Phi(v_j)+\pi_l\Phi(v_l)-(\pi_j+\pi_l)\Phi(v_{jl}) > \Delta\imath_1(k,l), \tag{16}$$

$$\Delta\imath_2(j,l) = \pi_j\Phi(\bar{v}_j)+\pi_l\Phi(\bar{v}_l)-(\pi_j+\pi_l)\Phi(\bar{v}_{jl}) > \Delta\imath_2(k,l). \tag{17}$$

Fig. 4 illustrates (16) where,

$$\delta_1 = \frac{\Delta\imath_1(j,l)}{\pi_j+\pi_l}, \quad \delta_2 = \frac{\Delta\imath_1(k,l)}{\pi_k+\pi_l}. \tag{18}$$

We have the following relations on the triangles in Fig. 4,

$$\frac{\delta_1}{\Delta_1+\Delta_2} = \frac{v_{jl}-v_j}{v_l-v_j} = \frac{\pi_l}{\pi_j+\pi_l}, \tag{19}$$

$$\frac{\delta_2}{\Delta_2} = \frac{v_{kl}-v_k}{v_l-v_k} = \frac{\pi_l}{\pi_k+\pi_l}, \tag{20}$$

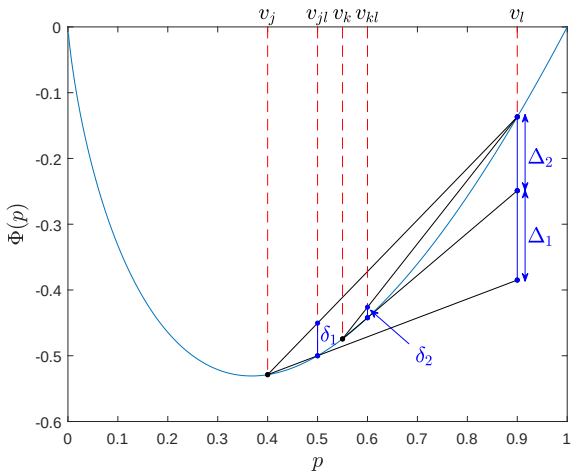


Fig. 4: Illustration of $\Delta\imath_1(j,l)$ and $\Delta\imath_1(k,l)$.

where the second equality comes from (13). Notice that $\Delta_1 > 0$, since $v_{jl} \le v_k$ and $\Phi(\cdot)$ is a strictly convex function. Using (19) and (20) in (18) we have

$$\Delta\imath_1(j,l) = \pi_l(\Delta_1+\Delta_2) > \pi_l\Delta_2 = \Delta\imath_1(k,l), \tag{21}$$

which proves (16). We can prove (17) in a similar way since from the assumption in (15) we have $\bar{v}_{jl} \ge \bar{v}_k$.

If we assume other side of inequality from (15), namely

$$\frac{\pi_j}{\pi_l} = \frac{v_l-v_{jl}}{v_{jl}-v_j} \le \frac{v_l-v_k}{v_k-v_j}, \tag{22}$$

we can similarly prove that $\Delta\imath(j,l) > \Delta\imath(j,k)$. This completes the proof.

References

[1] B. M. Kurkoski, K. Yamaguchi, K. Kobayashi, "Noise thresholds for discrete LDPC decoding mappings", in *Proc. IEEE Global Telecommun. Conf.*, pp. 1–5, Nov./Dec. 2008.

[2] I. Tal, A. Vardy, "How to construct polar codes", *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6562–6582, 2013.

[3] M. Dabirnia, A. Martinez, A. Guillén i Fàbregas, "A Mismatched Decoding Perspective of Channel Output Quantization," in *Proc. Inform. Theory Workshop*, Visby, Sweden, Aug. 2019.

[4] B. M. Kurkoski, H. Yagi, "Quantization of Binary-Input Discrete Memoryless Channels", *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4544–4552, Aug. 2014.

[5] H. Yagi, B. M. Kurkoski, "Channel Quantizers that Maximize Random Coding Exponents for Binary-Input Memoryless Channels", in *Proc. IEEE Int. Conf. Commun*, pp. 2256-2260, Jun. 2012.

[6] D. Burshtein, V. Della Pietra, D. Kanevsky, A. Nádas, "Minimum impurity partitions", *Ann. Statist.*, vol. 20, no. 3, pp. 1637–1646, Sep. 1992.

[7] N. Slonim and N. Tishby, "Agglomerative Information Bottleneck", in *Proc. of Neural Information Processing Systems (NIPS-99)*, pp. 617–623, 1999.

[8] K. Iwata and S. Ozawa, "Quantizer design for outputs of binary-input discrete memoryless channels using SMAWK algorithm", in *Proc. IEEE Int. Symp. Information Theory* , Honolulu, HI, 2014, pp. 191–195.

[9] E. Laber, M. Molinaro, F. M. Pereiram, "Binary Partitions with Approximate Minimum Impurity" in *Proc. of Machine Learning Research*, Jul. 2018.

[10] J. A. Zhang and B. M. Kurkoski, "Low-complexity quantization of discrete memoryless channels," in *Proc. Int. Symp. on Information Theory and Its Applications*, Monterey, CA, 2016, pp. 448–452.

An Upgrading Algorithm with Optimal Power Law

Or Ordentlich, Ido Tal

***Abstract*—Consider a channel W along with a given input distribution P_X. In certain settings, such as in the construction of polar codes, the output alphabet of W is often 'too large', and hence we replace W by a channel Q having a smaller output alphabet. We say that Q is upgraded with respect to W if W is obtained from Q by processing its output. In this case, the mutual information $I(P_X, W)$ between the input and output of W is upper-bounded by the mutual information $I(P_X, Q)$ between the input and output of Q. In this paper, we present an algorithm that produces an upgraded channel Q from W, as a function of P_X and the required output alphabet size of Q, denoted L. We show that the difference in mutual informations is 'small'. Namely, it is $O(L^{-2/(|\mathcal{X}|-1)})$, where $|\mathcal{X}|$ is the size of the input alphabet. This power law of L is optimal.**

I. Introduction

In his seminal paper on polar codes, Arıkan introduced synthetic channels [1, equation (5)], also called bit-channels. These synthetic channels have a binary input alphabet and an intractably large output alphabet. Namely, the output alphabet size of such a channel is at least 2^N, where N is the length of the polar code. When *constructing* a polar code, the vast size of the output alphabet is very much an issue. We note that in many settings more general than the seminal one, we search for channels that are 'very noisy'. A crucial observation is that instead of considering the original synthetic channel, one may approximate it by another channel having a much smaller output alphabet size [2]. Specifically, if the approximating channel is upgraded with respect to the original channel and shown to be 'very noisy', then this must also be the case for the original channel.

II. Setting

We are given a channel $W : \mathcal{X} \to \mathcal{Y}$ along with an input distribution P_X. We denote the mutual information between the input and output of W as $I(P_X, W) \triangleq I(X;Y)$, where X and Y are random variables with joint distribution

$$P_{X,Y}(x, y) = P_X(x)W(y|x) \ . \tag{1}$$

Let $Q : \mathcal{X} \to \mathcal{Z}$ be a channel with the same input alphabet as $W : \mathcal{X} \to \mathcal{Y}$. We say that Q is *upgraded* with respect to W if we can obtain W by processing the output of Q. That is, if there exists a third channel $\Phi : \mathcal{Z} \to \mathcal{Y}$ such that, for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$W(y|x) = \sum_{z \in \mathcal{Z}} Q(z|x)\Phi(y|z) \ .$$

Put another way, we want X, Z, and Y to form a Markov chain, in that order.

O. Ordentlich is with the School of Computer Science and Engineering, Hebrew University of Jerusalem, Israel (email: or.ordentlich@mail.huji.ac.il).

I. Tal is with the Department of Electrical Engineering, Technion, Haifa 32000, Israel (email: idotal@ee.technion.ac.il).

Our goal in this paper, given a fixed input alphabet size $|\mathcal{X}|$, an input distribution P_X, a channel $W : \mathcal{X} \to \mathcal{Y}$, and a parameter L, is to construct a channel $Q : \mathcal{X} \to \mathcal{Z}$ that is upgraded with respect to W and whose output alphabet size satisfies $|\mathcal{Z}| \le L$. Our method produces such a Q for which

$$I(P_X, Q) - I(P_X, W) = O(L^{-2/(|\mathcal{X}|-1)}) \ . \tag{2}$$

By [3, Section IV], the above power law of L is optimal.

III. The algorithm

Similarly to the method in [4], we use the 'one-hot' representation of $x \in \mathcal{X}$ to affect a reduction from the non-binary alphabet $\mathcal{X}$ to the binary alphabet $\mathcal{X}'$. Namely, w.l.o.g. let us assume that $\mathcal{X} = \{1, 2, \ldots, q\}$. We will replace $x \in \mathcal{X}$ by a length $q-1$ vector $f(x) = (x_1, x_2, \ldots, x_{q-1})$, such that

$$x_i = \begin{cases} 1 & \text{if } x = i \\ 0 & \text{otherwise} \end{cases}$$

For each $1 \le i \le q-1$, we apply the binary-input upgrading algorithm in [3, Section VI] to $P_{X_i Y | X_1^{i-1} = 0_1^{i-1}}$, and require that the output alphabet size of the upgrading channel satisfy $|\mathcal{Z}^{(i)}| \le \Lambda$, where $\Lambda = \left\lfloor L^{1/(q-1)} \right\rfloor$. Denote the resulting joint distribution $\beta^{(i)}_{X_i, Z_i, Y}(x', z', y)$. From these $q-1$ distributions, we define our final distribution on (X, Z, Y). The output alphabet is

$$\mathcal{Z} = \mathcal{Z}^{(1)} \times \mathcal{Z}^{(2)} \times \cdots \times \mathcal{Z}^{(q-1)} \ ,$$

and the joint distribution is

$$P^*_{X,Z,Y}(x, z, y) = P_Y(y) \cdot \left(\prod_{i=1}^{q-1} \beta^{(i)}_{Z_i|Y}(z_i|y) \right) \cdot \left(\prod_{i=1}^{q-1} \gamma^{(i)}_{X_i|Z_i, X_1^{i-1}}(x_i|z_i, x_1^{i-1}) \right) \ , \tag{3}$$

where, for $1 \le i \le q$,

$$\gamma^{(i)}_{X_i|Z_i, X_1^{i-1}}(x_i|z_i, x_1^{i-1}) = \begin{cases} \beta^{(i)}_{X_i|Z_i}(x_i|z_i) & \text{if } x_1^{i-1} = 0_1^{i-1} \ , \\ 1 & \text{if } x_1^{i-1} \ne 0_1^{i-1} \text{ and } x_i = 0 \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{4}$$

References

[1] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inform. Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.

[2] R. Mori and T. Tanaka, "Performance and construction of polar codes on symmetric binary-input memoryless channels," in *Proc. IEEE Int'l Symp. Inform. Theory (ISIT'2009)*, Seoul, South Korea, 2009, pp. 1496–1500.

[3] A. Kartowsky and I. Tal, "Greedy-merge degrading has optimal power-law," *IEEE Trans. Inform. Theory*, vol. 65, no. 2, pp. 917–934, February 2019.

[4] A. Bhatt, B. Nazer, O. Ordentlich, and Y. Polyanskiy, "Information-distilling quantizers," *arXiv preprint arXiv:1812.03031*, 2018.

Efficient Compression of Long Arbitrary Sequences with No Reference at the Encoder

Yuval Cassuto and Jacob Ziv
Viterbi Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa Israel 32000
Email: {ycassuto, jz}@ee.technion.ac.il

***Abstract*—In a distributed information application an encoder compresses an arbitrary vector while a similar reference vector is available to the decoder as side information. For the Hamming-distance similarity measure, and when guaranteed perfect reconstruction is required, we present two contributions to the solution of this problem. One potential application of the results is the compression of DNA sequences, where similar (but not identical) reference vectors are shared among senders and receivers.**

I. Introduction

This paper[1] continues the line of work on guaranteed-success compression with Hamming-bounded side information [1]. In the first part of the paper (Section II), we study the case where the encoder as usual does not know the decoder's reference vector $\boldsymbol{z}$, but it does have a set Z of vectors that contains $\boldsymbol{z}$ (among many other vectors). Our results in this part show that if the vectors in Z have a certain well-defined "clustering" property, then it is possible to reduce the compression rate below the best known. This can be achieved without any probabilistic assumptions on the set Z, and without directly enforcing a bound on its size. Our results in this part are for guaranteed-decoding *average* compression rate, where the average is taken over the random hash function used, and *not* over the input $\boldsymbol{y}$ (which has no probability distribution). For the same model our results also include a lower bound on compression rate for any scheme that uses random hashing. In the second part of the paper (Section III), we return to the classical model of [1] (no Z in the encoder), and propose coding schemes with low complexity of encoding and decoding. For guaranteed decoding of length-n vectors with a constant fractional distance bound p, existing schemes require decoding complexity that is exponential in n due to the complexity of decoding an error-correcting code. Our proposed schemes have $O(n\sqrt{n})$ decoding complexity, which is low enough for practical implementation even for long input sequences. For low distance fractions p, our scheme has low compression rates, although not as low as the prior schemes that do not consider the decoding complexity. We use codes with structure similar to *generalized concatenation* (GC) codes [2].

[1]A full version of this paper is currently under review for the IEEE Transactions on Information Theory.

II. Structured side information

Let $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M\}$ be a set of vectors, where each vector $\boldsymbol{z}_i$ is a binary vector of length n. The set Z is known to the encoder, and it contains the reference vector $\boldsymbol{z}$ available at the decoder (but the encoder does not know which one it is). The structure of Z is defined through the *p-spread parameter*: $p'(Z,p) \triangleq \frac{D_p(Z)}{2n}$, where $D_p(Z)$ is the maximal distance between a pair of vectors in Z whose distance is at most $2pn$. Given those definitions, we have an achievability result

Theorem 1. *Let Z be a set of reference vectors with p-spread parameter p'. Then there exists a coding scheme where for any input vector $\boldsymbol{y}$,*

$$|ENC(\boldsymbol{y})| \leq n\left[H(p) + H(p') + \epsilon\right], \tag{1}$$

as $n \to \infty$ and on average over the random hash functions.

$H(\cdot)$ is the entropy function. We also have the converse

Theorem 2. *Given the parameters p and p', any compression scheme that encodes $\boldsymbol{y}$ as $u(\boldsymbol{y})$, where $u : \{0,1\}^n \to \{0,1\}^m$ is a random hash function, requires asymptotically for any $\boldsymbol{y}$, on average over the random hash functions*

$$|ENC(\boldsymbol{y})| \geq n\left[H(p' + p)\right]. \tag{2}$$

III. Unstructured side information

For the case of unstructured side information that only assumes that the Hamming distance between $\boldsymbol{y}$ and $\boldsymbol{z}$ is at most pn, we propose a deterministic (guaranteed success) compression scheme built on a GC code construction with the following compression rate

Theorem 3. *For any constant integer l the compression rate of the GC-based construction is*

$$H\left(3p + \frac{1}{l}\left(\frac{1}{2} - 3p\right)\right) + \sum_{i=2}^{l}\left[H\left(3p + \frac{i}{l}\left(\frac{1}{2} - 3p\right)\right) - H\left(3p + \frac{i-1}{l}\left(\frac{1}{2} - 3p\right)\right)\right] \cdot \frac{3p}{3p + \frac{i-1}{l}\left(\frac{1}{2} - 3p\right)}. \tag{3}$$

References

[1] A. Orlitsky and K. Viswanathan, "One-way communication and error-correcting codes," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1781–1788, 2003.
[2] E. L. Blokh and V. V. Zyablov, *Generalized Concatenated Codes*. Moscow, Sviaz' (in Russian), 1976.

Locally Repairable Codes from Sum-Rank Codes

Umberto Martínez-Peñas and Frank R. Kschischang
Dept. of Electrical & Computer Engineering
University of Toronto
Toronto, Ontario, M5S 3G4, Canada
Email: {umberto, frank}@ece.utoronto.ca

***Abstract*—Locally repairable codes (LRCs) are considered with equal or unequal localities, local distances, and local field sizes. An explicit two-layer architecture with a sum-rank outer code is obtained, having disjoint local groups and achieving maximal recoverability (MR) for all families of local linear codes (MDS or not) simultaneously, up to a specified maximum locality r. Furthermore, the local linear codes (thus the localities, local distances, and local fields) can be efficiently and dynamically modified without global recoding or changes in architecture or outer code, while preserving the MR property, easily adapting to new configurations in storage or new hot and cold data. In addition, local groups and file components can be added, removed or updated without global recoding. The construction requires global fields of size roughly g^r, for g local groups and maximum or specified locality r. For equal localities, these global fields are smaller than those of previous MR-LRCs when $r \leq h$ (global parities). For unequal localities, they provide an exponential field size reduction on all previous best known MR-LRCs. For bounded localities and a large number of local groups, the global erasure-correction complexity of the given construction is comparable to that of Tamo–Barg codes or Reed–Solomon codes with local replication, while local repair is as efficient as for the Cartesian product of the local codes. Reed–Solomon codes with local replication and Cartesian products are recovered from the given construction when $r = 1$ and $h = 0$, respectively. The given construction can also be adapted to provide hierarchical MR-LRCs for all types of hierarchies and parameters. Finally, subextension subcodes and sum-rank alternant codes are introduced to obtain further exponential field size reductions, at the expense of lower information rates. This work is reported in [1].**

REFERENCES

[1] U. Martínez-Peñas and F. R. Kschischang, "Universal and Dynamic Locally Repairable Codes With Maximal Recoverability via Sum-Rank Codes," *IEEE Trans. Info. Theory*, vol. 65, no. 12, pp. 7790–7805, Dec. 2019.

Efficient Evaluation of Asymptotic Trapping Set Enumerators for Irregular LDPC Code Ensembles

Emna Ben Yacoub*, Gianluigi Liva†, Gerhard Kramer*
* Institute for Communications Engineering, Technical University of Munich, Munich, Germany
† Institute of Communications and Navigation, German Aerospace Center (DLR), Wessling, Germany

Email: {emna.ben-yacoub,gerhard.kramer}@tum.de, gianluigi.liva@dlr.de

***Abstract*—The normalized logarithmic asymptotic distribution of elementary and general trapping sets for irregular low-density parity-check code ensembles is derived based on the generating functions approach. A numerical technique for its evaluation is presented that requires solving a system of equations.**

I. Introduction

Trapping sets [1], [2] and absorbing sets [3] play a fundamental role in the error floor performance (under iterative decoding) of low-density parity-check (LDPC) codes [4]. An enumeration of the trapping sets present within a specific LDPC code graph is a formidable problem (see, e.g., [5]–[7]). The difficulty can be circumvented by analyzing the average trapping set enumerators of an LDPC code ensemble, rather than analyzing a specific code. This path was followed in [8] where a characterization of the (asmptotic) trapping set properties of regular/irregular unstructured LDPC ensembles was obtained based on random matrix enumeration methods.

In this paper, we provide an alternative derivation of the normalized logarithmic asymptotic distribution of elementary and general trapping sets for irregular LDPC code ensembles. The derivation relies on the generating function approach, already adopted for the analysis of weight and stopping set enumerators of unstructured (generalized) LDPC ensembles [9]–[13], and it requires solving a system of equations.

The paper is organized as follows. In Section II, we review definitions and known results. Section III provides the derivation of the asymptotic trapping set enumerators for irregular LDPC code ensembles. Numerical results are presented and discussed in Section IV. Conclusions follow in Section V.

II. Preliminaries

A. LDPC Codes

Binary LDPC codes are binary linear block codes defined by an $m \times n$ sparse parity-check matrix $\boldsymbol{H}$. The code dimension is $k \geq n - m$. The Tanner graph of an LDPC code is a bipartite graph $G = (\mathcal{V} \cup \mathcal{C}, \mathscr{E})$ consisting of n variable nodes (VNs) and m check nodes (CNs). The set $\mathscr{E}$ of edges contains the elements e_{ij}, where e_{ij} is an edge between VN $\mathsf{v}_j \in \mathcal{V}$ and CN $\mathsf{c}_i \in \mathcal{C}$. Note that e_{ij} belongs to the set $\mathcal{E}$ if and only if the parity-check matrix element h_{ij} is equal to 1. The sets $\mathcal{N}(\mathsf{v}_j)$ and $\mathcal{N}(\mathsf{c}_i)$ denote the neighbors of VN v_j and CN c_i, respectively. The degree of a VN v_j (CN c_i) is the cardinality of the set $\mathcal{N}(\mathsf{v}_j)$ ($\mathcal{N}(\mathsf{c}_i)$). The node-oriented degree distribution polynomials of an LDPC code graph are

$$\Lambda(x) = \sum_i \Lambda_i x^i, \quad \mathrm{P}(x) = \sum_i \mathrm{P}_i x^i \tag{1}$$

where Λ_i, P_i correspond, respectively, to the fraction of VNs and CNs with degree i. We further define by $\mathsf{d}_\mathsf{v}^{\max}$ ($\mathsf{d}_\mathsf{c}^{\max}$) the maximum VN (CN) degree. We denote by

$$\bar{\mathsf{d}}_\mathsf{v} = \sum_i i\Lambda_i, \quad \bar{\mathsf{d}}_\mathsf{c} = \sum_i i\mathrm{P}_i \tag{2}$$

the average VN and CN degrees, respectively. Note that $n\bar{\mathsf{d}}_\mathsf{v} = m\bar{\mathsf{d}}_\mathsf{c}$ represents the total number of edges. We define ξ as

$$\xi = \frac{m}{n} = \frac{\bar{\mathsf{d}}_\mathsf{v}}{\bar{\mathsf{d}}_\mathsf{c}}. \tag{3}$$

An unstructured irregular LDPC code ensemble $\mathscr{C}_n^{\Lambda,\mathrm{P}}$ is the set of all LDPC codes with block length n defined by a bipartite graph with degree distributions $\Lambda(x)$ and $\mathrm{P}(x)$.

B. Trapping Sets

For a set $\mathcal{S} \subseteq \mathcal{V}$ of VNs , we denote by $\mathcal{N}(\mathcal{S})$ the set of its neighboring CNs. Further, we denote by $\mathcal{O}(\mathcal{S})$ the set of CNs in $\mathcal{N}(\mathcal{S})$ that are connected to $\mathcal{S}$ an odd number of times and $\mathcal{E}(\mathcal{S})$ the set of CNs in $\mathcal{N}(\mathcal{S})$ that are connected to $\mathcal{S}$ an even number of times.

Definition 1. An (a, b) trapping set (TS) $\mathcal{T}_{a,b}$ is set $\mathcal{S}$ of a VNs such that $\mathcal{O}(\mathcal{S})$ contains b CNs [8].

Definition 2. An elementary trapping set (ETS) $\mathcal{T}_{a,b}^{\mathrm{E}}$ is a TS where each CN $\mathsf{c} \in \mathcal{E}(\mathcal{S})$ is connected to two VNs in $\mathcal{S}$ and each CN $\mathsf{c} \in \mathcal{O}(\mathcal{S})$ is connected to exactly one VN in $\mathcal{S}$.

C. Useful Results

Definition 3. Let $x(n)$ and $y(n)$ be two real-valued sequences, where $y(n) \neq 0\ \forall n$, $x(n)$ is exponentially equivalent to $y(n)$ as $n \to \infty$ if and only if

$$\lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{x(n)}{y(n)}\right) = 0.$$

We will use the notation $x(n) \doteq y(n)$ to specify that $x(n)$ is exponentially equivalent to $y(n)$.

Lemma 1. We have

$$\binom{\alpha n}{\beta n} \doteq \exp\left\{ n\alpha H\left(\frac{\beta}{\alpha}\right)\right\} \tag{4}$$

where $H(p) = -p\ln(p) - (1-p)\ln(1-p)$ is the entropy function.

Lemma 2. For every sequence $y(w)$, we have

$$\sum_{w} \exp(ny(w)) \doteq \exp\left(n \max_{w} y(w)\right). \tag{5}$$

For $\boldsymbol{z} = (z_1, z_2, \ldots, z_d)$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, we define

$$\boldsymbol{z}^{\boldsymbol{\alpha}} = \prod_{t=1}^{d} z_t^{\alpha_t}. \tag{6}$$

Lemma 3. [Hayman Formula for Multivariate Polynomials] Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_d)$ and let $p(\boldsymbol{z})$ be a multivariate polynomial with $p(\boldsymbol{0}) \neq 0$. Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ where $0 \leq \alpha_t \leq 1$ and $\alpha_t n$ is an integer for all $t \in \{1, 2, \ldots, d\}$. Then we have [14, Appendix A.2]

$$\operatorname{coeff}(p(\boldsymbol{z})^n, \boldsymbol{z}^{n\boldsymbol{\alpha}}) \doteq \exp\left\{ n\left[\ln\left(p(\boldsymbol{x})\right) - \sum_{t=1}^{d} \alpha_t \ln(x_t)\right]\right\} \tag{7}$$

where $\operatorname{coeff}(p(\boldsymbol{z})^n, \boldsymbol{z}^{n\boldsymbol{\alpha}})$ represents the coefficient of $\boldsymbol{z}^{n\boldsymbol{\alpha}}$ in the polynomial $p(\boldsymbol{z})^n$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ and $x_1, x_2, \ldots, x_d$ are the unique positive solutions to

$$x_t \frac{\partial p(\boldsymbol{x})}{\partial x_t} = \alpha_t p(\boldsymbol{x}), \qquad \forall t \in \{1, 2, \ldots, d\}. \tag{8}$$

III. Asymptotic Distribution of Trapping Sets

The average number of size (a, b) TSs in the Tanner graph of a code drawn randomly from the ensemble $\mathscr{C}_n^{\Lambda,\mathrm{P}}$ is

$$\mathsf{E}_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(a,b) = \sum_{w} \frac{\operatorname{coeff}\left(g(x,y)^n, x^w y^b\right)}{\binom{n\bar{\mathsf{d}}_{\mathsf{v}}}{w}} \operatorname{coeff}\left(f(t,s)^n, t^a s^w\right) \tag{9}$$

where we introduced the generating functions

$$f(t,s) = \prod_{j=1}^{\mathsf{d}_{\mathsf{v}}^{\max}} (1 + ts^j)^{\Lambda_j} \tag{10}$$

and

$$g(x,y) = \prod_{i=1}^{\mathsf{d}_{\mathsf{c}}^{\max}} \left[\frac{(1+x)^i + (1-x)^i}{2} + y\frac{(1+x)^i - (1-x)^i}{2}\right]^{\xi \mathrm{P}_i}. \tag{11}$$

Proof. Consider the Tanner graph of a code drawn randomly from the ensemble $\mathscr{C}_n^{\Lambda,\mathrm{P}}$. The number of ways to choose a VNs such that exactly w edges emanate from them is

$$\operatorname{coeff}\left(f(t,s)^n, t^a s^w\right).$$

Moreover, the number of ways to choose w check sockets such that exactly b CNs each have an odd number of sockets and the other CNs each have an even number of check sockets is given by

$$\operatorname{coeff}\left(g(x,y)^n, x^w y^b\right).$$

We randomly choose a set $\mathcal{S}$ of a VNs with a uniform distribution over all $\binom{n}{a}$ possibilities. Let Z_1 be a random variable (RV) indicating the number of edges emanating from the set $\mathcal{S}$. Further, let Z_2 be a RV that is equal to 1 if there are exactly b CNs each connected an odd number of times to $\mathcal{S}$ and the other CNs each have an even number (including zero) of connections to $\mathcal{S}$, and to 0 otherwise. Thus

$$\mathsf{E}_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(a,b) = \binom{n}{a} \Pr\{Z_2 = 1\} \tag{12}$$

and

$$\begin{aligned} \Pr\{Z_2 = 1\} &= \sum_{w} \Pr\{Z_1 = w\} \Pr\{Z_2 = 1 | Z_1 = w\} \\ &= \sum_{w} \frac{\operatorname{coeff}(f(t,s)^n, t^a s^w)}{\binom{n}{a}} \frac{\operatorname{coeff}\left(g(x,y)^n, x^w y^b\right)}{\binom{n\bar{\mathsf{d}}_{\mathsf{v}}}{w}}, \end{aligned} \tag{13}$$

□

The normalized logarithmic asymptotic distributions of TSs for the ensemble $\mathscr{C}_n^{\Lambda,\mathrm{P}}$ for $a = \theta n$ and $b = \gamma n$ is defined by

$$G_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(\theta,\gamma) := \lim_{n\to\infty} \frac{1}{n} \ln\left(\mathsf{E}_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(\theta n, \gamma n)\right) \tag{14}$$

where $\mathsf{E}_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(\theta n, \gamma n)$ is the average number of $(\theta n, \gamma n)$ TSs in the Tanner graph of a random code in $\mathscr{C}_n^{\Lambda,\mathrm{P}}$.

Theorem 1. The normalized asymptotic distribution of $(\theta n, \gamma n)$ TSs is given by

$$\begin{aligned} G_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(\theta,\gamma) = &-\bar{\mathsf{d}}_{\mathsf{v}} \ln(1+xs) - \theta\ln(t) - \gamma\ln(y) \\ &+ \ln\left(f(t,s)\right) + \ln\left(g(x,y)\right) \end{aligned} \tag{15}$$

where t, s, x, y are the unique positive solutions of

$$t\frac{\partial f(t,s)}{\partial t} = \theta f(t,s) \tag{16}$$

$$s\frac{\partial f(t,s)}{\partial s} = \tilde{w}^\star f(t,s) \tag{17}$$

$$x\frac{\partial g(x,y)}{\partial x} = \tilde{w}^\star g(x,y) \tag{18}$$

$$y\frac{\partial g(x,y)}{\partial y} = \gamma g(x,y) \tag{19}$$

where $f(t,s)$ and $g(x,y)$ are defined in (10) and (11) respectively and

$$\tilde{w}^\star = \bar{\mathsf{d}}_{\mathsf{v}} \frac{xs}{1+xs}. \tag{20}$$

The proof can be found in Appendix A. Note that to compute the normalized asymptotic distribution of ETSs, we simply need to replace $g(x,y)$ given in (11) with

$$g(x,y) = \prod_{i=1}^{\mathsf{d}_{\mathsf{c}}^{\max}} \left(1 + \binom{i}{2}x^2 + ixy\right)^{\xi \mathrm{P}_i}. \tag{21}$$

Definition 4. For fixed ratio $\Delta = b/a$, the second zero crossing of $G_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(\theta, \Delta\theta)$ (the first one is zero), if it exists, is

called the *typical minimum* Δ*-trapping set size* that we denote by $\theta^{\star}_{\mathrm{TS}}$ [15].

To determine $\theta^{\star}_{\mathrm{TS}}$ we add another equation to the system of equations of Theorem 1, namely

$$G^{\Lambda,\mathrm{P}}_{\mathrm{TS}}(\theta, \Delta\theta) = 0 \tag{22}$$

with $\theta > 0$.

Lemma 4. For fixed $\Delta = \gamma/\theta$ and $\theta \to 0$ the derivative of $G^{\Lambda,\mathrm{P}}_{\mathrm{TS}}(\theta, \Delta\theta)$ is given by

$$\frac{\partial G^{\Lambda,\mathrm{P}}_{\mathrm{TS}}(\theta, \Delta\theta)}{\partial \theta} = -\ln(t) - \Delta \ln(y). \tag{23}$$

Proof. The solutions of the system of equations in (16)–(19) are implicit functions of θ. From (15) and (20) we get the expression in (24). The terms in the brackets are equal to zero due to (16)-(19). This yields the result of Lemma 4. □

Note that for a fixed Δ a positive $\theta^{\star}_{\mathrm{TS}}$ exists whenever the derivative of $G^{\Lambda,\mathrm{P}}_{\mathrm{TS}}(\theta, \Delta\theta)$ is negative as $\theta \to 0$.

IV. Numerical Results

Example 1. Consider a rate $1/2$ ensemble $\mathscr{C}^{\Lambda,\mathrm{P}}_n$ with $\Lambda(x) = 0.8x^3 + 0.2x^4$, $\mathrm{P}(x) = 0.6x^6 + 0.4x^7$. The normalized logarithmic asymptotic distribution of ETSs and TSs of this ensemble are depicted in Fig. 1 for fixed ratio $\Delta = \gamma/\theta \in \{0.005, 0.05, 0.1, 0.3, 0.5\}$. Observe that the gap between TSs and ETS is very small for small θ.

Example 2. Consider a rate $1/2$ ensemble $\mathscr{C}^{\Lambda,\mathrm{P}}_n$ with $\Lambda(x) = 0.8x^4 + 0.2x^5$, $\mathrm{P}(x) = 0.6x^8 + 0.4x^9$. The normalized logarithmic asymptotic distribution of (elementary) TSs of this ensemble are depicted in Fig. 2 for $\Delta \in \{0.005, 0.05, 0.1, 0.3, 0.5\}$. We remark that this ensemble has better trapping set properties than the one in Example 1.

V. Conclusion

Expressions of the asymptotic distributions of elementary and general trapping sets for unstructured LDPC code ensembles have been derived. The evaluation of the expressions requires solving a system of equations. Using the proposed method, we reproduced the results in [8], where the derivation of the asymptotic TS distribution is based on asymptotic enumeration techniques for matrices with specified column and row weight vectors.

Appendix A
Proof of Theorem 1

From Lemma 3, we have

$$\begin{aligned}\operatorname{coeff}\left(f(t,s)^n, t^{n\theta} s^{n\tilde{w}}\right) \doteq \exp\{n[\ln(f(t,s)) - \theta\ln(t) \\ -\tilde{w}\ln(s)]\}\end{aligned} \tag{25}$$

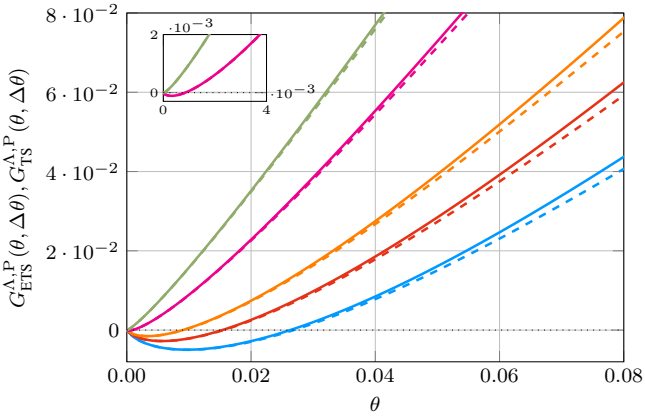


Fig. 1. Normalized logarithmic asymptotic distribution of elementary (– – –, – – –, – – –, – – –, – – –) and general trapping sets (——, ——, ——, ——, ——) of the ensemble in Example 1 for $\Delta = 0.005$ (——), $\Delta = 0.05$ (——), $\Delta = 0.1$ (——), $\Delta = 0.3$ (——), $\Delta = 0.5$ (——).

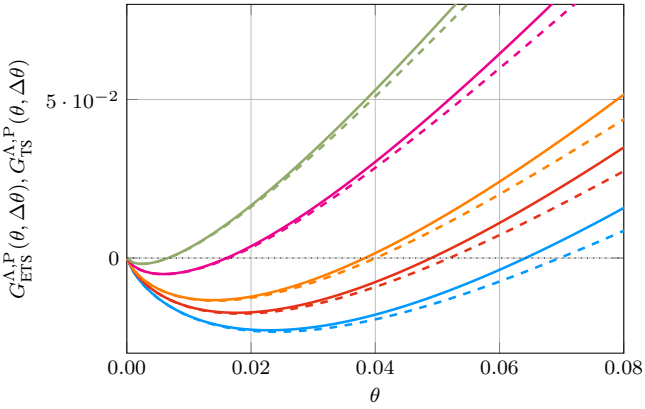


Fig. 2. Normalized logarithmic asymptotic distribution of elementary (– – –, – – –, – – –, – – –, – – –) and general trapping sets (——, ——, ——, ——, ——) of the ensemble in Example 2 for $\Delta = 0.005$ (——), $\Delta = 0.05$ (——), $\Delta = 0.1$ (——), $\Delta = 0.3$ (——), $\Delta = 0.5$ (——).

and

$$\begin{aligned}\operatorname{coeff}\left(g(x,y)^n, x^{n\tilde{w}} y^{n\gamma}\right) \doteq \exp\{n[\ln(g(x,y)) \\ -\gamma\ln(y) - \tilde{w}\ln(x)]\}\end{aligned} \tag{26}$$

where $\tilde{w} = w/n$ and t, s, x, y are the unique positive solutions of

$$\begin{aligned} t\frac{\partial f(t,s)}{\partial t} &= \theta f(t,s) \\ s\frac{\partial f(t,s)}{\partial s} &= \tilde{w} f(t,s) \\ x\frac{\partial g(x,y)}{\partial x} &= \tilde{w} g(x,y) \\ y\frac{\partial g(x,y)}{\partial y} &= \gamma g(x,y). \end{aligned} \tag{27}$$

$$\frac{\mathrm{d}G^{\Lambda,\mathrm{P}}_{\mathrm{TS}}(\theta, \Delta\theta)}{\mathrm{d}\theta} = -\ln(t) - \Delta\ln(y) + \frac{\mathrm{d}t}{\mathrm{d}\theta}\left[-\frac{\theta}{t} + \frac{\frac{\partial f(t,s)}{\partial t}}{f(t,s)}\right] + \frac{\mathrm{d}s}{\mathrm{d}\theta}\left[-\frac{\tilde{w}^{\star}}{s} + \frac{\frac{\partial f(t,s)}{\partial s}}{f(t,s)}\right] + \frac{\mathrm{d}x}{\mathrm{d}\theta}\left[-\frac{\tilde{w}^{\star}}{x} + \frac{\frac{\partial g(x,y)}{\partial x}}{g(x,y)}\right] + \frac{\mathrm{d}y}{\mathrm{d}\theta}\left[-\frac{\Delta\theta}{y} + \frac{\frac{\partial g(x,y)}{\partial y}}{g(x,y)}\right] \tag{24}$$

Lemma 1 gives

$$\binom{n\bar{\mathsf{d}}_{\mathsf{v}}}{n\tilde{w}} \doteq \exp\left\{n\bar{\mathsf{d}}_{\mathsf{v}} H\left(\frac{\tilde{w}}{\bar{\mathsf{d}}_{\mathsf{v}}}\right)\right\} \tag{28}$$

and from (25), (26) and (28), we have

$$\mathsf{E}_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(\theta,\gamma) \doteq \sum_{\tilde{w}} \exp(nS(\tilde{w})) \tag{29}$$

with

$$\begin{aligned} S(\tilde{w}) = &-\bar{\mathsf{d}}_{\mathsf{v}} H\left(\frac{\tilde{w}}{\bar{\mathsf{d}}_{\mathsf{v}}}\right) + \ln(f(t,s)) - \theta\ln(t) \\ &- \tilde{w}\ln(xs) + \ln(g(x,y)) - \gamma\ln(y). \end{aligned} \tag{30}$$

Thus, we have

$$G_{\mathrm{TS}}^{\Lambda,\mathrm{P}}(\theta,\gamma) = \max_{\tilde{w}} S(\tilde{w}). \tag{31}$$

It can be shown that

$$\tilde{w}^{\star} = \operatorname*{argmax}_{\tilde{w}} S(\tilde{w}) = \bar{\mathsf{d}}_{\mathsf{v}} \frac{xs}{1+xs}. \tag{32}$$

By substituting (32) in (30) and (27), we obtain (15)-(19), as desired.

REFERENCES

[1] D. J. MacKay and M. S. Postol, "Weaknesses of Margulis and Ramanujan-Margulis low-density parity-check codes," *Electronic Notes in Theoretical Computer Science*, vol. 74, pp. 97–104, 2003.

[2] T. Richardson, "Error floors of LDPC codes," in *Proc. Allerton Conf. on Commun., Control and Computing*, Monticello, USA, Oct. 2003.

[3] B. Amiri, C. Lin, and L. Dolecek, "Asymptotic distribution of absorbing sets and fully absorbing sets for regular sparse code ensembles," *IEEE Trans. Commun.*, vol. 61, no. 2, pp. 455–464, February 2013.

[4] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[5] S. K. Chilappagari, S. Sankaranarayanan, and B. Vasić, "Error floors of LDPC codes on the binary symmetric channel," in *IEEE Int. Conf. Commun.*, vol. 3, June 2006, pp. 1089–1094.

[6] S. K. Chilappagari and B. Vasić, "Error-correction capability of column-weight-three LDPC codes," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2055–2061, May 2009.

[7] B. Vasić, S. K. Chilappagari, D. V. Nguyen, and S. K. Planjery, "Trapping set ontology," in *Proc. Allerton Conf. on Commun., Control and Computing*, Monticello, USA, Oct. 2009.

[8] O. Milenkovic, E. Soljanin, and P. Whiting, "Asymptotic spectra of trapping sets in regular and irregular LDPC code ensembles," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 39–55, Jan 2007.

[9] D. Burshtein and G. Miller, "Asymptotic enumeration methods for analyzing LDPC codes," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1115–1131, June 2004.

[10] A. Orlitsky, K. Viswanathan, and J. Zhang, "Stopping set distribution of LDPC code ensembles," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 929–953, Mar. 2005.

[11] C. Di, T. Richardson, and R. Urbanke, "Weight Distribution of Low-Density Parity-Check Codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4839–4855, Nov. 2006.

[12] M. F. Flanagan, E. Paolini, M. Chiani, and M. P. C. Fossorier, "On the Growth Rate of the Weight Distribution of Irregular Doubly Generalized LDPC Codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3721–3737, June 2011.

[13] ——, "Spectral shape of doubly-generalized LDPC codes: Efficient and exact evaluation," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7212–7228, Nov 2013.

[14] C. Di, "Asymptotic and finite-length analysis of low-density parity-check codes," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2004.

[15] S. Abu-Surra, W. Ryan, and D. Divsalar, "Ensemble trapping set enumerators for protograph-based LDPC codes," *Proc. 45th Annual Allerton Conf. on Commun., Control and Computing*, pp. 201–210, Sep 2007.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Joint Decoding of Distribution Matching and Error Control Codes

Patrick Schulte[1], Wafa Labidi[2], Gerhard Kramer[1]
[1]Institute for Communications Engineering, Technical University of Munich, 80333, Munich, Germany
[2]Institute of Theoretical Information Technology, Technical University of Munich, 80333, Munich, Germany

***Abstract*—An improved decoder for low-density parity-check (LDPC) codes and for probabilistic amplitude shaping with constant composition distribution matching (CCDM) is presented. The decoder combines standard LDPC belief propagation with a soft-input soft-output processor that exploits the constraints imposed by CCDM and it gains up to 0.5 dB at a frame error rate of 10^{-4} for a block-length $n = 192$ 5G code.**

I. Introduction

Probabilistic amplitude shaping (PAS) [1] is a block based probabilistic shaping (PS) technique that induces a non-uniform distribution on a signal constellation. A distribution matcher (DM) encodes a message into a non-linear set that satisfies a constraint on the average symbol distribution. A systematic forward error correction (FEC) encoder preserves the distribution in the systematic part.

A constant composition distribution matcher (CCDM) [2] is a DM that imposes a common empirical distribution on the constellation points' amplitudes within a block. The CCDM thus introduces dependencies over all symbols in a block. For very long blocks, the PAS rate is not affected by these dependencies, but systems with short length DMs suffer in transmission rate [3]. In [4]–[8], DMs with smaller rate-loss are proposed. In [9] the dependencies introduced by an extremely short 4-D shell mapping (SMDM) [4]–[6] are resolved by a 4-D demodulator. The authors of [10] use polar codes with list decoding and check if the codeword candidates fulfill the constant composition (CC) constraint.

PAS uses a systematic FEC encoder in a manner similar to the Bliss scheme [11] for constrained sequence coding. To improve the Bliss scheme's performance, [12] and [13] use a supplementary soft input soft output (SISO) decoder and iterate with the usual FEC decoder. We adopt this approach for PAS and let a low-density parity-check (LDPC) decoder iterate with a SISO CC code decoder based on the forward backward (BCJR) algorithm to improve performance. For this purpose, we introduce the trellis of a CC code. The resulting decoder is a generalized LDPC (GLDPC) decoder [14] with a non-linear constraint.

This paper is structured as follows. In Sec. II we introduce notation and the basic components of PAS. In Sec. III we introduce the interface of the BCJR algorithm and construct a trellis for CC codes. In Sec. IV we show combinations of BCJR and LDPC-belief propagation (BP) decoders. Simulation results are presented in Sec. V. We draw conclusions in Sec. VI.

II. Preliminaries and Notation

A. Notation

We write matrices in capital bold letters $\mathbf{L}$, random variables with uppercase sans-serif letters X, and their realizations with lowercase letters x. Let A be a discrete random variable with probability mass function (pmf) P_{A} defined on the set $\mathcal{A}$. The *entropy* of a random variable A is

$$\mathbb{H}(\mathsf{A}) = \sum_{a \in \mathrm{supp}(P_{\mathsf{A}})} -P_{\mathsf{A}}(a) \log_2 \left(P_{\mathsf{A}}(a)\right) \tag{1}$$

where $\mathrm{supp}(P_{\mathsf{A}}) \subseteq \mathcal{A}$ is the support of P_{A}, i.e., the subset of a in $\mathcal{A}$ with positive probability. We denote a length n vector of random variables as $\mathsf{A}^n = \mathsf{A}_1\mathsf{A}_2 \cdots \mathsf{A}_n$ with realization $a^n = a_1 a_2 \cdots a_n$, and the number of occurrences of letter $\alpha \in \mathcal{A}$ in a^n as $n_\alpha(a^n)$. Next, we describe the channel model and the components of the PAS transceiver.

B. Channel Model

For transmission we consider M-amplitude shift keying (ASK), i.e., transmission symbols X take on values in $\mathcal{X} = \{-M+1, -M+3, \cdots, M-3, M-1\}$. Each symbol can be factored into a sign and amplitude

$$\mathsf{X} = \mathsf{A} \cdot \mathsf{S}. \tag{2}$$

The corresponding amplitude set is

$$\mathcal{A} = \{\alpha_1, \alpha_2, \cdots, \alpha_{M/2}\} = \{1, 3, \cdots, M-1\}. \tag{3}$$

We consider additive white Gaussian noise (AWGN), i.e., the output symbols of the channel are obtained via

$$\mathsf{Y} = \mathsf{X} + \mathsf{Z} \tag{4}$$

where Z is a Gaussian random variable with zero mean and variance σ^2. The signal-to-noise ratio (SNR) is

$$\mathrm{SNR} = \frac{\mathbb{E}\left[\mathsf{X}^2\right]}{\sigma^2}. \tag{5}$$

C. Probabilistic Amplitude Shaping

PAS [1] is a coded modulation scheme that can approach the Shannon capacity for the AWGN channel [15], [16] and is rate adaptive. An important building block is the DM which encodes messages into sequences of amplitudes with a desired average distribution. One can use any DM, and common choices are CCDM and SMDM [6]. A systematic LDPC encoder generates parity bits from a binary representation of the amplitudes. The parities serve as signs for the amplitudes.

For high rate codes, additional source bits are encoded without distribution matching. We refer to [1] for a detailed review of PAS.

D. Labeling Function

An invertible labeling function β converts m bits to an $M = 2^m$-ary symbol $x \in \mathcal{X}$:

$$\beta(b_1, \cdots, b_m) = x. \tag{6}$$

The inverse function is

$$\beta^{-1}(x) = [b_1, \cdots, b_m]. \tag{7}$$

We refer to the j-th bit of the label by $\beta_j^{-1}(x)$. We use a binary reflected Gray code (BRGC) [17] where b_1 decides the symbol's sign, i.e., we have

$$\beta_A(b_2, \cdots, b_m) = |\beta(0, b_2, \cdots, b_m)| = |\beta(1, b_2, \cdots, b_m)|. \tag{8}$$

The notation $b_{i,j}$ refers to the j-th bit of the i-th symbol x_i. We write $\boldsymbol{B}$ to refer to all bits $b_{i,j}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

E. Demodulation

We consider a symbol-wise demodulator that is aware of the signal statistics P_{A}, P_{B_j}. The log-likelihoods (LLs) $\tilde{L}_i(x)$ of the i-th transmitted symbol are

$$\tilde{L}_i(x) = \log(p_{\mathsf{Y}|\mathsf{X}}(\mathsf{Y}_i|\mathsf{X}_i = x) \cdot P_{\mathsf{X}}(x)), \forall x \in \mathcal{X}. \tag{9}$$

The demodulator calculates the bit-wise LLs

$$\tilde{L}_{i,j}(b) = \log(p_{\mathsf{Y}|\mathsf{B}_j}(\mathsf{Y}_i|\mathsf{B}_{i,j} = b) \cdot P_{\mathsf{B}_j}(b)), \forall b \in \{0, 1\}. \tag{10}$$

Thus, one symbol-channel splits into m parallel bit-channels [1]. The log-likelihood ratio (LLR) of the j-th bit in the i-th transmitted symbol is

$$L_{i,j} = \tilde{L}_{i,j}(0) - \tilde{L}_{i,j}(1). \tag{11}$$

For convenience, we collect LLs and LLRs in the matrices $\tilde{\mathbf{L}}$ and $\mathbf{L}$, respectively. The (i, j)-th entry of the LLR matrix $\mathbf{L}$ corresponds to $L_{i,j}$. The (i, j)-th entry of the LL matrix $\tilde{\mathbf{L}}$ corresponds to $\tilde{L}_i(\xi_j)$, $\xi_j \in \mathcal{X}$.

F. LDPC Codes and BP Decoding

A (n, k) LDPC code [18] is a binary linear block code described by an $r \times n$ parity-check matrix $\mathbf{H}$ with entries $h_{i,j}$, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, n$, where $r \geq n - k$. LDPC codes can be visualized through a bipartite graph also known as the Tanner graph $\mathcal{G}$. This graph consists of a set $\mathcal{V} = \{V_1, V_2, \cdots, V_n\}$ of n variable nodes, a set $\mathcal{C} = \{C_1, C_2, \cdots, C_r\}$ of r check nodes and a set $\mathcal{E} = \{e_{j,i}\}$ of edges. The check node C_j is connected to the variable node V_i through the edge $e_{j,i}$ if the entry $h_{i,j}$ of the parity-check matrix is one.

An LDPC BP decoder operates on LLRs [1]. Based on the channel observations $\mathbf{L}^{\mathsf{CH}}$, the LDPC decoder outputs the APP LLRs:

$$\mathbf{L}^{\mathsf{APP}} = \mathbf{L}^{\mathsf{CH}} + \mathbf{L}^{\mathsf{E,LDPC}} \tag{12}$$

where $\mathbf{L}^{\mathsf{E,LDPC}}$ denotes the extrinsic information.

G. Constant Composition Distribution Matching

The type $\boldsymbol{t}$ of a sequences a^n expresses how many times each letter $\alpha \in \mathcal{A}$ appears in a^n, i.e., we have

$$\boldsymbol{t} = (n_{\alpha_1}(a^n), n_{\alpha_2}(a^n), \cdots, n_{\alpha_{|\mathcal{A}|}}(a^n)). \tag{13}$$

The set of sequences of type $\boldsymbol{t}$ is

$$\mathcal{T}_{\boldsymbol{t}} = \{a^n \in \mathcal{A}^n \,|n_{\alpha_i}(a^n) = t_i \,, \quad i = 1, \cdots, |\mathcal{A}|\} \tag{14}$$

where t_i is the i-th entry of $\boldsymbol{t}$. The cardinality is

$$|\mathcal{T}_{\boldsymbol{t}}| = \frac{n!}{\prod_{i=1}^{|\mathcal{A}|} t_i!}. \tag{15}$$

The CCDM is a function

$$f_{\text{ccdm},\boldsymbol{t}} : \{0, 1\}^k \to \mathcal{C}_{\text{ccdm}} \tag{16}$$

where $\mathcal{C}_{\text{ccdm}}$ is a subset of $\mathcal{T}_{\boldsymbol{t}}$. Thus, all codewords of the CCDM have the same type and therefore the same empirical distribution. The dematcher $f_{\text{ccdm},\boldsymbol{t}}^{-1}$ implements the inverse operation. For large n, the CCDM rate

$$R_{\text{ccdm}} = k/n \tag{17}$$

tends to $\mathbb{H}(P_{\mathsf{A}})$ with $P_{\mathsf{A}}(i) = \frac{t_i}{n}$ [2], where $\mathbb{H}(P_{\mathsf{A}})$ is the entropy of a discrete memoryless source (DMS) with symbol probabilities P_{A}. The difference

$$R_{\text{loss}} = \mathbb{H}(P_{\mathsf{A}}) - R_{\text{ccdm}} \tag{18}$$

is called the rate-loss R_{loss} [3]. In [19, Sec. IV] the CCDM rate-loss is upper and lower bounded by $\mathcal{O}(\log(n)/n)$ where n is the block length. The rate-loss of a CCDM is negligible for large blocks, but for short blocks the CC constraint adds substantial redundancy. Consider a sequence a^n with a type constraint. If we know all symbols except for one, we can recover its value by counting how often each letter appears. This holds for any constraint length n. We want to exploit the redundancy of a CC code at the decoder.

III. Forward-Backward Algorithm for Constant Composition Codes

The BCJR algorithm [20], also known as the forward-backward algorithm, is a SISO algorithm that calculates the a posteriori symbol probabilities

$$P^{\mathsf{APP}}(a_i) = P(a_i|\tilde{\mathbf{L}}) \tag{19}$$

where $\tilde{\mathbf{L}}$ are LLs and a_i is the i-th transmitted symbol. From these probabilities, we can compute the extrinsic LLs $\tilde{\mathbf{L}}^{\mathrm{E}}$ [20]. For binary codes, the input interface may be LLRs, because we can convert easily from LLRs to LLs and vice versa. The constant composition BCJR (CCBCJR) decoder builds the code trellis from the type vector $\boldsymbol{t}$, i.e. it is a function

$$\text{CCBCJR} : \tilde{\boldsymbol{L}} \times \boldsymbol{t} \mapsto \tilde{\boldsymbol{L}}^{\mathsf{E}}. \tag{20}$$

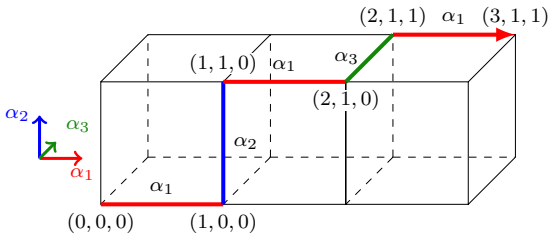


Fig. 1. Constant composition code trellis for type $\boldsymbol{t} = (3, 1, 1)$. This trellis consists of 16 states and 28 branches and represents 20 different CC codewords, thus paths.

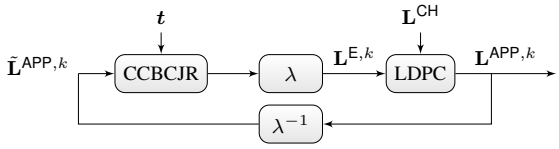


Fig. 2. Symbol-based decoder.

A. CC Code Trellis

The construction of the CC trellis borrows ideas from [21]. The trellis states are tuples

$$\mathcal{S} = \{0, 1, \cdots, n_{\alpha_1}\} \times \{0, 1, \cdots, n_{\alpha_2}\} \times \cdots \{0, 1, \cdots, n_{\alpha_{|\mathcal{A}|}}\}. \tag{21}$$

The number of states in the trellis is

$$|\mathcal{S}| = \prod_{\alpha \in \mathcal{A}} (n_\alpha + 1) \tag{22}$$

and the number of edges is

$$E = \sum_{\alpha \in \mathcal{A}} n_\alpha \prod_{\alpha' \neq \alpha} (n_{\alpha'} + 1). \tag{23}$$

The initial and final states are $(0, \cdots, 0)$ and $(n_{\alpha_1}, \cdots, n_{\alpha_{|\mathcal{A}|}})$, respectively. State $s \in \mathcal{S}$ is connected to an earlier state $s' \in \mathcal{S}$ via symbol α_q if all entries are identical except for the q-th entry of s that is augmented by one.

Example 1. Consider a CC code on the alphabet $\mathcal{A} = \{\alpha_1, \alpha_2, \alpha_3\}$ and with type $\boldsymbol{t} = (3, 1, 1)$. The trellis is depicted in Fig. 1. It consists of $|\{0, 1, 2, 3\}| \cdot |\{0, 1\}| \cdot |\{0, 1\}| = 16$ states. The colored path corresponds to the sequence $(\alpha_1\alpha_2\alpha_1\alpha_3\alpha_1)$. It includes three increment-steps of α_1, one increment-step of α_2, and one increment-step of α_3, and therefore matches the sequence type.

Note that an CCBCJR decoder assumes that that we may use the complete set $\mathcal{T}_{\boldsymbol{t}}$ of sequences of type $\boldsymbol{t}$, however $\mathcal{C}_{\text{ccdm}}$ is usually only a subset [2].

IV. Joint Decoding

We study how the decoder can exploit CC code properties to decrease the error probability.

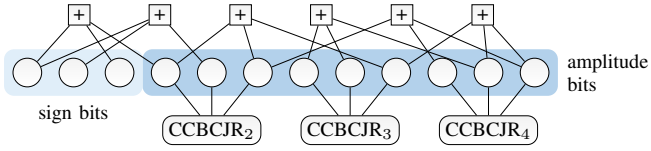


Fig. 3. Tanner graph of a naive bit-based decoder for $m = 4$ and $n = 3$. The amplitude bit variable nodes are connected to the respective BCJR node. The sign bit variable nodes are not connected to a BCJR node.

A. Symbol-Based Decoder

The symbol-based decoder consists of a CCBCJR decoder and a LDPC decoder that exchange messages iteratively, see Fig. 2. The BCJR decoder has a symbol based interface, while the LDPC decoder has a bit based interface. The demodulator provides the LLs of the symbols and bit levels. The symbol-wise LLs are passed to the CCBCJR decoder. The LDPC decoder and the CCBCJR decoder iterate extrinsic information $\mathbf{L}^{\mathsf{E}}$. We use functions λ and λ^{-1} to convert from symbol based to bit based and vice versa. The function λ converts LL into LLRs via

$$L_{i,j} = \ln \left(\frac{\sum\limits_{x:\beta_j^{-1}(x)=0} \exp\left(\tilde{L}_i(x)\right)}{\sum\limits_{x:\beta_j^{-1}(x)=1} \exp\left(\tilde{L}_i(x)\right)} \right). \tag{24}$$

The function λ^{-1} converts from bit-level to symbol-level. For simplicity, we assume for a fixed i that the $\mathsf{B}_{i,j}$, $j = 1, 2, \ldots, m$, are pairwise independent given Y_i. The conversion is then

$$\tilde{L}_i(x) = \log \left(\prod_{j=2}^{m} \frac{\exp(L_{i,j} \cdot (1 - 2\beta_{A,j}^{-1}(x)))}{1 + \exp(L_{i,j} \cdot (1 - 2\beta_{A,j}^{-1}(x)))} \right). \tag{25}$$

Note that $1 - 2\beta_{A,j}^{-1}(x)$ is 1 for the bit 0 and -1 for the bit 1.

B. Bit-Based Decoder

The number of states and edges of the CCBCJR decoder increases exponentially with the alphabet size and polynomially in n. One idea to decrease complexity is to replace one $|\mathcal{A}|$-ary CCBCJR decoder by $\log_2 |\mathcal{A}|$ binary CCBCJR decoders. Additionally, the conversion functions λ, λ^{-1} become obsolete.

Consider a transmission sequence x^n with type constraint $\boldsymbol{t}$ on the amplitudes and its binary representation $\boldsymbol{B} \in \{0, 1\}^{n \times m}$ according to the labeling function β, where the entry $b_{i,j}$ corresponds to the j-th bit of the i-th symbol. Let $\boldsymbol{b}_j^{|} = b_{1,j}, b_{2,j}, \cdots b_{n,j}$ be the j-th column of $\boldsymbol{B}$, i.e., the j-th bit-level of the binary representation of the symbol sequence. Since x^n has a type $\boldsymbol{t}$ constraint on the amplitudes only, the sign bits are unconstrained. All other bit levels j, $2 \leq j \leq m$ are constrained. We derive the type constraint for each bit-level depending on the type $\boldsymbol{t}$ of the sequence x^n and the labeling function β. The number of zeros in bit-level j is equal to the number of amplitudes in the sequence x^n whose binary

representation is zero in the j-th position, i.e., we have

$$n_b(\boldsymbol{b}_j^{|}) = \sum_{\alpha\in\mathcal{A},\ \beta_j^{-1}(\alpha)=b} n_\alpha(\mathrm{amp}(x^n)) \tag{26}$$

where $\mathrm{amp}(x^n)$ is the element-wise absolute value of x^n, $b \in \{0,1\}$, and $2 \le j \le m$. Thus for one amplitude type constraint $\boldsymbol{t}$, we obtain $m-1$ bit constraints $\boldsymbol{t}_2, \cdots, \boldsymbol{t}_m$, where the index denotes the respective bit-level with

$$\boldsymbol{t}_j = \left[n_0(\boldsymbol{b}_j^{|}), n_1(\boldsymbol{b}_j^{|})\right]. \tag{27}$$

Example 2. Consider a sequence x^n with amplitude constraint $\boldsymbol{t} = [37, 20, 6, 1]$, i.e., 37 ones, 20 threes, 6 fives and 1 sevens, and the BRGC labeling β shown below.

x	-7	-5	-3	-1	1	3	5	7
$\beta(x)$	111	110	100	101	001	000	010	011

We find $n_1(\boldsymbol{b}_2^{|}) = 7$ because the second bit of the labeling β is '1' for amplitudes 5 and 7 and they appear 6 times and once, respectively. The corresponding bit types $\boldsymbol{t}_2$ and $\boldsymbol{t}_3$ are

$$\boldsymbol{t}_2 = [57, 7] \tag{28}$$

$$\boldsymbol{t}_3 = [26, 38]. \tag{29}$$

For decoding, we add $m-1$ BCJR nodes into the Tanner graph, as shown in Fig. 3. Note that the bit-based CCBCJR decoders run independently. Their combined trellises allow sequences that do not fulfill the type constraint $\boldsymbol{t}$.

C. Improved Bit-Based Decoder

Each of the $m-1$ BCJR nodes is connected to n/m nodes. This suggests that the girth, i.e., the shortest cycle in the graph, is small. Loopy BP for small-girth was investigated in [22] and leads to oscillations. There are two basic approaches to deal with this issue. Firstly, we may filter the beliefs and thus attenuate oscillations. Second, we could introduce multiple short length CC constraints on a bit-level, i.e., introduce lower degree CCBCJR nodes which increases both the girth and the rate-loss. We consider only the first approach in this paper.

The LDPC decoder outputs the a posteriori LLRs $\boldsymbol{L}_j^{\mathsf{APP}}$. Based on the channel observation, the type vector $\boldsymbol{t}_j$ and a posteriori information, the j-th BCJR decoder CCBCJR_j generates the extrinsic information $\mathbf{L}_j^{\mathsf{E}}$. The outputs of the $m-1$ CCBCJRs are collected in the matrix $\mathbf{L}^{\mathsf{E}}$. $\mathbf{L}$ and $\mathbf{L}^{\mathsf{E}}$ are then processed by the function

$$\mathrm{g}\left(\mathbf{L}^{\mathsf{CH}}, \mathbf{L}^{\mathsf{E}}, \mathbf{L}^{\mathsf{APP}}, k\right) \approx \mathbf{L}^{\mathsf{CH}} + \underbrace{\left(\mu \cdot \mathbf{L}^{\mathsf{E},k-1} + (1-\mu)\cdot \mathbf{L}^{\mathsf{E},k}\right)}_{\text{prior information}} \tag{30}$$

with $k \ge 1$ and $\mu \in [0,1]$. After a number of iterations, the LDPC decoder outputs new a posteriori information, which is sent back to the CCBCJR decoders. The optimal parameter μ is found by grid search.

D. Computational Complexity Comparison

For the computational complexity analysis, we focus on the number of edges E in the code trellises, since the BCJR complexity is $\Theta(E)$ [23]. This analysis depends on the trellis representation of the CC code.

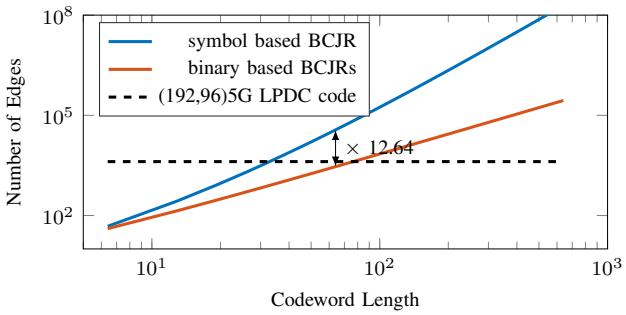


Fig. 4. Number of branches to compute for the bit-based and symbol-based BCJR algorithms. The empirical distribution is $[37, 20, 6, 1]/64$. We interpret (31) and (32) as continuous functions. At output length 64 symbols, the symbol-based BCJR algorithm needs about 12.5 times more states than the binary-based BCJR algorithm. We compare with the number of branches of an iterative LDPC decoder using the BCJR algorithm.

1) Symbol-Based Decoder: For a type $\boldsymbol{t} = [n_1, \cdots, n_{M/2}]$ constraint, we have

$$E_{\text{symb}} = \sum_{i=1}^{M/2} n_i \prod_{j\neq i}(n_j + 1) \tag{31}$$

branches. An increasing alphabet size even for the same block-length may result in a large increase in the number of states and therefore the computational complexity. For a given empirical distribution, the number of states scales with the power of the support of the empirical distribution.

2) Bit-Based Decoder: For the bit-based decoder, we split one amplitude type constraint $\boldsymbol{t}$ into $m-1$ bit constraints $\boldsymbol{t}_2, \cdots, \boldsymbol{t}_m$. The number of edges is then

$$E_{\text{bit}} = \sum_{j=2}^{m} 2n_0(\boldsymbol{b}_j^{|})n_1(\boldsymbol{b}_j^{|}) + n_0(\boldsymbol{b}_j^{|}) + n_1(\boldsymbol{b}_j^{|}). \tag{32}$$

In Fig. 4 we show the number of branches vs. the codeword length for the empirical distribution $[37, 20, 6, 1]/64$. We also add the number of branches that are evaluated during one iteration of LDPC decoding of an (192,96) 5G LDPC code, i.e., we compute the number of branches of all single parity check and repetition nodes. Single parity check and repetition nodes have 4 times and 2 times their degree edges, respectively.

V. Simulation Results

We compare the performance of PAS with the bit-level decoder proposed in [1] with the symbol-based and the heuristically improved bit-based decoder with supplementary CC constrained nodes. We target a spectral efficiency of 1.5 bits per channel use with 8-ASK constellation.

For encoding, we use a DM with type $\boldsymbol{t} = [37, 20, 6, 1]$ from Example 2 and a rate $3/4$ code from the 5G eMBB standard [25] with block length 192. The reference LDPC decoder [1] is biased with the empirical distribution of the FEC input. The symbol-based decoder uses $\boldsymbol{t}$ and the bit-based decoder has two CCBCJRs with $\boldsymbol{t}_2 = [7, 57]$ and $\boldsymbol{t}_3 = [38, 26]$.

Simulation results in Fig. 5 show that the LDPC decoder with a linear combination of $\mathbf{L}^{\mathsf{E,LDPC},k-1}$ and $\mathbf{L}^{\mathsf{E},k}$ outperforms the LDPC decoder with $\mathbf{L}^{\mathsf{E},k}$ as prior information only.

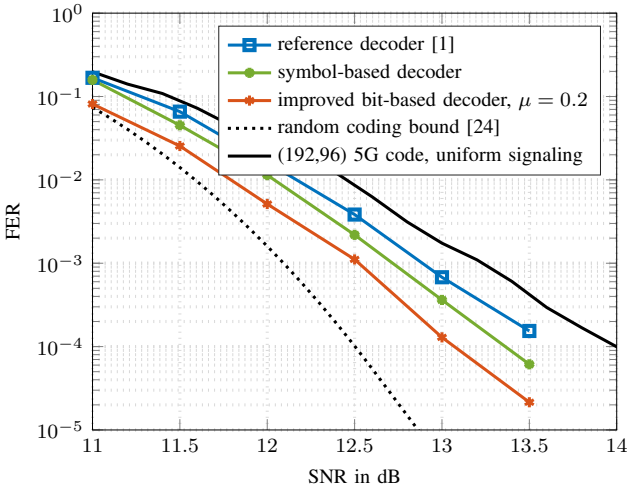


Fig. 5. FER of the different strategies for 24 outer-iterations and 100 inner-iterations. We collected 100 erroneous frames per simulation point. The scheme is implemented by using $8-$ASK with code rate $3/4$ and block-length $n = 192$. The rate-loss R_{loss} is about 0.145 bit/symbol.

We include the performance of a (192,96) 5G LDPC code with an optimized interleaver as a non-shaped baseline with the same spectral efficiency. The bit-based decoding strategy gains 0.5 dB in the simulation setup as compared to the LDPC decoder in [1].

VI. Conclusions and Outlook

A trellis structure for CC codes is introduced. Different decoding strategies based on the combination of BCJR and LDPC decoders are proposed that gain 0.5 dB in the considered short length scenario at a frame error rate of 10^{-4}. In future work, we plan to investigate the design of LDPC codes with CCBCJR nodes. This way long LDPC codes could be combined with short block length DMs that run in parallel during encoding and decoding.

VII. Acknowledgements

We would like to thank Georg Böcherer and Fabian Steiner for continuous support and Gianluigi Liva for the initial idea. Wafa Labidi was supported by the Bundesministerium für Bildung und Forschung (BMBF) through Grant 16KIS1003.

References

[1] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4651–4665, Dec 2015.

[2] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 430–434, Jan 2016.

[3] G. Böcherer, P. Schulte, and F. Steiner, "High throughput probabilistic shaping with product distribution matching," *arXiv preprint arXiv:1702.07510*, 2017.

[4] Y. C. Gültekin, F. M. J. Willems, W. J. van Houtum, and S. Şerbetli, "Approximate enumerative sphere shaping," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2018, pp. 676–680.

[5] Y. C. Gültekin, W. J. van Houtum, S. Şerbetli, and F. M. Willems, "Constellation shaping for IEEE 802.11," in *IEEE Ann. Int. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*. IEEE, 2017, pp. 1–7.

[6] P. Schulte and F. Steiner, "Divergence-optimal fixed-to-fixed length distribution matching with shell mapping," *IEEE Wireless Commun. Lett.*, pp. 1–1, 2019.

[7] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Multiset-partition distribution matching," *IEEE Trans. Commun.*, pp. 1–1, 2018.

[8] M. Pikus and W. Xu, "Bit-level probabilistically shaped coded modulation," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 1929–1932, Sep. 2017.

[9] F. Steiner, F. Da Ros, M. P. Yankov, G. Böcherer, P. Schulte, G. Kramer *et al.*, "Experimental verification of rate flexibility and probabilistic shaping by 4D signaling," in *Proc. Optical Fiber Commun. Conf.* IEEE, 2018, pp. 1–3.

[10] P. Yuan, G. Böcherer, P. Schulte, G. Kramer, R. Böhnke, and W. Xu, "Error detection using symbol distribution in a system with distribution matching and probabilistic amplitude shaping," German WO Application, 10 31, 2016.

[11] W. Bliss, "Circuitry for performing error correction calculations on baseband encoded data to eliminate error propagation," *IBM Tech. Discl. Bul.*, vol. 23, pp. 4633–4634, 1981.

[12] J. L. Fan and J. M. Cioffi, "Constrained coding techniques for soft iterative decoders," in *IEEE Global Telecommun. Conf. (GLOBECOM)*, vol. 1. IEEE, 1999, pp. 723–727.

[13] A. P. Hekstra, "Use of a d-constraint during LDPC decoding in a Bliss scheme," *arXiv preprint arXiv:0707.3925*, 2007.

[14] R. Tanner, "A recursive approach to low complexity codes," *IEEE Trans. Inf. Theory*, vol. 27, no. 5, pp. 533–547, 1981.

[15] R. A. Amjad, "Information rates and error exponents for probabilistic amplitude shaping," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018.

[16] G. Böcherer, "Achievable rates for probabilistic shaping," *arXiv preprint arXiv:1707.01134*, 2017.

[17] F. Gray, "Pulse code communication," *US Patent 2632058*, 1953.

[18] R. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, January 1962.

[19] P. Schulte and B. C. Geiger, "Divergence scaling of fixed-length, binary-output, one-to-one distribution matching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2017, pp. 3075–3079.

[20] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (corresp.)," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 284–287, 1974.

[21] J. Schalkwijk, "An algorithm for source coding," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 395–399, 1972.

[22] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proc. Conf. on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 467–475.

[23] R. J. McEliece, "On the bcjr trellis for linear block codes," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1072–1092, 1996.

[24] G. Liva and F. Steiner, "pretty-good-codes.org: Online library of good channel codes," http://pretty-good-codes.org, Oct. 2017.

[25] T. Richardson and S. Kudekar, "Design of low-density parity check codes for 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 28–34, Mar. 2018.

Maximal Correlation Under Quantization

Dror Drach
Tel Aviv University
dror.drach@gmail.com

Or Ordentlich
Hebrew University of Jerusalem
or.ordentlich@mail.huji.ac.il

Ofer Shayevitz
Tel Aviv University
ofersha@eng.tau.ac.il

Abstract

The maximal correlation between a pair of jointly distributed random variables X and Y is a commonly used measure of dependency, often participating in bounds on the fundamental limits of various problems. One well-known example is Witsenhausen's lower bound on the probability that two Boolean functions of X and Y disagree, given their biases. Witsenhausen's lower bound trivially holds in a stronger form when replacing the maximal correlation between X and Y with the maximal correlation between one-bit functions of X and Y. This provides impetus to study the *quantized maximal correlation*, namely the behavior of maximal correlation under functions of finite bounded cardinality. In this paper, we derive various fundamental properties of the quantized maximal correlation, as well as of a closely related quantity corresponding to the χ^2-mutual information between quantized versions of X and Y.

From Information Inequalities to Computational Lower Bounds in Learning

Emmanuel Abbé
EPFL, Switzerland
Email: emmanuel.abbe@epfl.ch

Abstract

This talk shows how computational lower bounds in learning, which allow to show failure at learning certain function classes due to computational constraints, can be derived using information-theoretic arguments and inequalities. In particular, it is shown that GD-based deep learning cannot learn with polynomial parameters certain function classes that can be learned efficiently with other non-GD based algorithms.

Joint work with C. Sandon (MIT).

Dualizing Le Cam's method with applications to estimating the unseens

Yury Polyanskiy
MIT
Department of EECS
32-D668, MIT, Boston, MA, USA
Email: yp@mit.edu

Yihong Wu
Yale University
Department of Statistics and Data Science
24 Hillhouse Ave, New Haven, CT, USA
Email: yihong.wu@yale.edu

***Abstract*—One of the most commonly used techniques for proving statistical lower bounds, Le Cam's method, has been the method of choice for functional estimation. This papers aims at explaining the effectiveness of Le Cam's method from an optimization perspective. Under a variety of settings it is shown that the maximization problem that searches for the best lower bound provided by Le Cam's method, upon dualizing, becomes a minimization problem that optimizes the bias-variance tradeoff among a family of estimators. While Le Cam's method can be used with arbitrary distance, our duality result applies specifically to the χ^2-divergence, thus singling it out as a natural choice for quadratic risk. For estimating linear functionals of a distribution our work strengthens prior results of Dohono-Liu [DL91] (for quadratic loss) by dropping the Hölderian assumption on the modulus of continuity. For exponential families our results improve those of Juditsky-Nemirovski [JN09] by characterizing the minimax risk for the quadratic loss under weaker assumptions on the exponential family.**

We also provide an extension to the high-dimensional setting for estimating separable functionals. Notably, coupled with tools from complex analysis, this method is particularly effective for characterizing the "elbow effect" – the phase transition from parametric to nonparametric rates. As the main application of our methodology, we consider three problems in the area of "estimating the unseens", recovering the prior result of [PSW17] on population recovery and, in addition, obtaining two new ones:

- ***Distinct elements problem*: Randomly sampling a fraction p of colored balls from an urn containing d balls in total, the optimal normalized estimation error of the number of distinct colors in the urn is within logarithmic factors of $d^{-\frac{1}{2}\min\{\frac{p}{1-p},1\}}$, exhibiting an elbow at $p=\frac{1}{2}$;**
- ***Fisher's species problem*: Given n independent samples drawn from an unknown distribution, the optimal normalized prediction error of the number of unseen symbols in the next (unobserved) $r\cdot n$ samples is within logarithmic factors of $n^{-\min\{\frac{1}{r+1},\frac{1}{2}\}}$, exhibiting an elbow at $r=1$.**

References

[DL91] David L. Donoho and Richard C. Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, 19:668–701, 1991.

[FCW43] Ronald Aylmer Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

[JN09] Anatoli B Juditsky and Arkadi S Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5A):2278–2300, 2009.

[PSW17] Y. Polyanskiy, A. T. Suresh, and Y. Wu. Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory (COLT)*, Amsterdam, Netherland, Jul 2017. arXiv:1702.05574.

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

Information Constrained Optimal Transport: From Talagrand, to Marton, to Cover

Ayfer Ozgur
Stanford University
Palo Alto, CA, USA
Email: aozgur@stanford.edu

Abstract

The optimal transport problem studies how to transport one measure to another in the most cost-effective way and has wide range of applications from economics to machine learning. In this paper, we introduce and study an information constrained variation of this problem. Our study yields a strengthening and generalization of Talagrand's celebrated transportation-cost inequality. Following Marton, we show that our new transportation cost inequality can be used to recover old and new concentration of measure results. Finally, we provide an application of our transportation-cost inequality in network information theory. We show that it can be used to recover our recent solution to Cover's capacity problem of the relay channel.

Joint work with Yikun Bai and Xiugang Wu.

Smooth Wasserstein Distance: Metric Structure and Statistical Efficiency

Ziv Goldfeld
Cornell University
goldfeld@cornell.edu

***Abstract*—The Wasserstein distance has seen a surge of interest and applications in machine learning. Its popularity is driven by many advantageous properties it possesses, such as metric structure (metrization of weak convergence), robustness to support mismatch, compatibility to gradient-based optimization, and rich geometric properties. However, empirical approximation under the Wasserstein distance suffers from a severe curse of dimensionality, rendering it impractical in high dimensions. We propose a novel Gaussian-smoothed Wasserstein distance, that achieves the best of both worlds: preserving the Wasserstein metric structure while alleviating the empirical approximation curse of dimensionality. Furthermore, as the smoothing parameter shrinks to zero, smooth Wasserstein converges towards the classic metric (with convergence of optimizers), thus serving as a natural extension. These theoretic properties establish the smooth Wasserstein distance as favorable alternative to its classic counterpart for high-dimensional analysis and applications.**

I. Extended Abstract

The 1-Wasserstein distance (W_1) between two probability measures P and Q, with finite first moments, is

$$\mathsf{W}_1(P,Q) := \inf_{\pi\in\Pi(P,Q)} \int \|x-y\| \,\mathsf{d}\pi(x,y),$$

where $\Pi(P,Q)$ is the set of couplings of P and Q. This distance has many appealing properties, such as: (i) robustness to mismatched supports of P and Q (crucial for generative modeling applications); (ii) metrization of weak convergence of probability measures; (iii) defining a constant speed geodesic in the space of probability measures (giving rise to a natural interpolation between measures). These advantages, however, come at a price of slow empirical convergence rates, known as the 'curse of dimensionality'.

Suppose $(X_i)_{i=1}^n$ are i.i.d. samples from a Borel probability measure P on $\mathbb{R}^d$. Consider the rate at which the empirical measure $P_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ approaches P in the 1-Wasserstein distance, i.e., the $\mathbb{E}\mathsf{W}_1(P_n,P)$ rate of decay. Since W_1 metrizes narrow convergence, the Glivenko-Cantelli theorem implies $\mathsf{W}_1(P_n,P)\to 0$ as $n\to\infty$. Unfortunately, the convergence *rate* drastically deteriorates with dimension, scaling as $n^{-\frac{1}{d}}$ for any P absolutely continuous w.r.t. the Lebesgue measure [1]. This rate is sharp for all $d>2$. Thus, empirical approximation under W_1 is effectively infeasible in high dimensions – a disappointing shortcoming given the dimensionality of data in modern ML tasks.

To alleviate this impasse, we propose a novel framework, termed Gaussian-smooth Wasserstein distance that inherits the metric structure of W_1 while attaining much stronger statistical guarantees. The smooth Wasserstein distance of parameter $\sigma \geq 0$ between two d-dimensional probability measures P and Q is

$$\mathsf{W}_1^{(\sigma)}(P,Q) \triangleq \mathsf{W}_1(P*\mathcal{N}_\sigma, Q*\mathcal{N}_\sigma),$$

where $*$ stands for convolution and $\mathcal{N}_\sigma \triangleq \mathcal{N}(0,\sigma^2\mathrm{I}_d)$ is the isotropic Gaussian measure of parameter σ. In other words, $\mathsf{W}_1^{(\sigma)}(P,Q)$ is simply the W_1 distance between P and Q after each is smoothed by an isotropic Gaussian kernel.

Theorem 1 of [2] shows that just like W_1, for any $\sigma\in[0,+\infty)$, $\mathsf{W}_1^{(\sigma)}$ is a metric on the space of probability measures that metrizes weak topology. Namely, a sequence of probability measures $(P_k)_{k\in\mathbb{N}}$ converges weakly to P if and only if $\mathsf{W}_1^{(\sigma)}(P_k,P)\to 0$. This further implies that convergence to zero of W_1 and $\mathsf{W}_1^{(\sigma)}$ are equivalent (see [2, Theorem 2]). We next explore properties of $\mathsf{W}_1^{(\sigma)}(P,Q)$ as a function of σ for fixed P and Q. Theorem 3 in [2] establishes continuity and non-increasing monotonicity of $\mathsf{W}_1^{(\sigma)}(P,Q)$ in $\sigma\in[0,+\infty)$. These, in particular, imply that $\lim_{\sigma\to 0}\mathsf{W}_1^{(\sigma)}(P,Q)=\mathsf{W}_1(P,Q)$. Additionally, using the notion of Γ-convergence, Theorem 4 of the aforementioned work establishes convergence of optimal couplings. Namely, if $(\pi_k)_{k\in\mathbb{N}}$ is sequence of optimal couplings for $\mathsf{W}_1^{(\sigma_k)}(P,Q)$, where $\sigma_k\to 0$, then $(\pi_k)_{k\in\mathbb{N}}$ converges weakly to an optimal coupling for $\mathsf{W}_1(P,Q)$.

Lastly, consider empirical approximation under smooth Wasserstein, i.e., the convergence rate of $\mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n,P)$. It was shown in [3, Proposition 1] that Gaussian smoothing alleviates the curse of dimensionality, with $\mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n,P)$ converging as $n^{-\frac{1}{2}}$ in all dimensions. Although $\mathsf{W}_1^{(\sigma)}$ is specialized to Gaussian noise, Theorem 5 of [2] generalizes the empirical approximation result to account for subgaussian noise densities. The expected value analysis is followed by a concentration inequality for $\mathsf{W}_1^{(\sigma)}(P_n,P)$ derived through McDiarmid's inequality [2, Theorem 6].

References

[1] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Stats.*, 40(1):40–50, Feb. 1969.

[2] Z. Goldfeld and K. Greenewald. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *International Conference on Artificial Intelligence and Statistics (AISTATS-2020)*, Palermo, Sicily, Italy, Jun. 2020.

[3] Z. Goldfeld, K. Greenewald, Y. Polyanskiy, and J. Weed. Convergence of smoothed empirical measures with applications to entropy estimation. *arXiv preprint arXiv:1905.13576*, May 2019.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Arbitrarily Varying Broadcast Channel with Uncertain Cooperation

Uzi Pereg and Yossef Steinberg
Dept. of Electrical Engineering
Technion - IIT
Haifa 32000, ISRAEL
Email: {uzipereg@campus.,ysteinbe@}technion.ac.il

***Abstract*—In this work we study the arbitrarily varying physically degraded broadcast channel with cooperating decoders, with high degree of uncertainty in the network model: the channel statistics is arbitrarily varying, and the cooperation link is not reliable, as its existence is not guaranteed a priori. We construct a coding scheme that can cope with the arbitrarily varying nature of the channel, and with the cooperation link uncertainty. Inner and outer bounds are developed on the capacity region of this channel, and conditions are suggested under which the bounds coincide, thus characterising the capacity region of this model.**

Index Terms - Arbitrarily varying channel, broadcast channel, conference, cooperation, random codes, symmetrizabily, unreliable cooperation.

I. Introduction

The broadcast channel (BC) is one of the main building blocks of modern communication networks, and as such has been the subject of extensive research in multiuser communication for the last few decades. The BC with cooperating decoders was introduced in [7], [8], and a closely related model which presents the cooperation link as a relay channel, was suggested in [15], [16]. In this work, we study the BC with cooperating decoders where the model involves high degree of uncertainty: the channel statistics is arbitrarily varying over time, and the cooperation link is unreliable, as its existence is not guaranteed a priori. Our goal is to study network scenarios with the highest degree of uncertainty that can still yield meaningful models and results.

The arbitrarily varying channel (AVC), introduced by Blackwell *et al.* [5], is a channel whose statistics varies over time in an unknown manner, possibly without obeying any specific rule. In practice, such variations can represent physical effects whose statistics is unknown, or irrelevant in short transmission frames, like fading phenomena in wireless communications, defective cells in memory devices, malicious attacks on authentication and identification systems, and more. It is especially relevant to describe a classical communication system where a hostile user, referred to as *jammer*, interferes with the transmitted signals in order to disrupt communication. The arbitrarily varying broadcast channel (AVBC) was examined by Jahn [14], who derived inner bounds on the random code capacity of the AVBC. Jahn further showed that the deterministic code capacity region either coincides with the random code capacity region, or else its interior is empty - a reminiscent of the dichotomy property of single user AVCs, pointed out by Ahlswede [1]. Thus, in order to apply Jahn's inner bounds one first has to verify that the interior of the capacity region is nonempty. Hof and Bross [10] used observations and results by Ericson [9] and Csiszár and Narayan [6] to resolve this dichotomy, and showed that a necessary and sufficient condition for the capacity region to have a non-empty interior is that both marginal channels are non-symmetrizable. In [17], the AVBC with causal side information at the encoder is presented. Inner and outer bounds on the random code capacity region are developed, and sufficient conditions are suggested under which the bounds coincide, thus characterising the random code capacity region. The conditions can be viewed as a set- extension of the minimax theorem in convex optimisation. Similar results are obtained also for the (deterministic code) capacity region. As in [14] and [10], a dichotomy property applies also for the case of causal side information. A symmetrizability condition for channels with causal side information is developed, and it is shown that a sufficient condition for the channel capacity to have a non-empty interior is that both marginals are non symmetrizable.

In the classical approach to AVCs, we seek the maximal communication rate that the channel can support with any sequence of channel statistics which the jammer can choose. It thus pertains to a *worst case design*. In a cooperative setting, if the cooperation link is unreliable, a worst case approach would lead to coding schemes and achievable rates that ignore the cooperation altogether. A less stringent approach, explored in previous works, is to construct coding schemes that are robust in the following sense: the decoders exploit the cooperation link when it is present, but can still operate when it is absent, possibly leading to lower decoding rates. This robust approach to cooperation schemes was suggested in [18] and extensively studied in [11], [12] and [13]. It can be viewed as a compound channel model, where the channel has two possible realisations, one with cooperation link present, and one where it is absent.

In this work we study the physically degraded AVBC with conferencing decoders, where the conference link is unreliable,

This research was supported by the Israel Science Foundation (grant No. 1285/16).

as in [18], [11] and [12]. The main motivation is to provide insights to the role of cooperation in networks with high degree of uncertainty. As observed by Wiese [19] and Wiese and Boche [20] in the context of the multiple access channel, a small amount of cooperation can make a substantial change in the behaviour of an AVC network, as it can be used to apply Ahlswede's Elimination Technique in cases where the channels of part of the users are symmetrizable.

II. Problem formulation

A. Notation and general definitions

Let $\mathcal{X}$,$\mathcal{S}$,$\mathcal{Y}_1$ and $\mathcal{Y}_2$ be finite sets. Denote by $\mathcal{P}(\mathcal{S})$ the collection of all probability mass functions (PMFs) over $\mathcal{S}$, and similarly for $\mathcal{S}^n$, $\mathcal{X}$, $\mathcal{X}^n$, etc. We are given a discrete memoryless state-dependent broadcast channel (BC) $(\mathcal{X}, \mathcal{S}, W_{Y_1,Y_2|X,S}, \mathcal{Y}_1, \mathcal{Y}_2)$, where $\mathcal{X}$ and $\mathcal{S}$ are the input and state alphabets, respectively, $W_{Y_1,Y_2|X,S}$ is the channel transition probability matrix, and $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are the output alphabets of user 1 and user 2, respectively. We will often denote the channel by $W_{Y_1,Y_2|X,S}$. The channel is assumed memoryless and without feedback. Let $q \in \mathcal{P}(\mathcal{S}^n)$ stand for a generic distribution of the state sequence s^n, we will be more specific about the choices of q later. The *arbitrarily varying broadcast channel* (AVBC) is a BC where the distribution of the state sequence is unknown. In particular, it need not be memoryless nor stationary, and can give mass 1 to a specific sequence s^n. We denote the AVBC by $\mathcal{B}$. The *compound broadcast channel* (CBC) $\mathcal{B}^{\mathcal{Q}}$ is a BC with discrete memoryless (iid) state, whose single-letter state distribution q is unknown, but belongs to a given set $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{S})$. With a slight abuse of notation, we use q to denote a member of $\mathcal{P}(\mathcal{S}^n)$ or $\mathcal{P}(\mathcal{S})$, the exact choice will be clear from the context. In our model, there is a (unreliable) conference link, of capacity C_1, from user 1 to user 2. Fix the transmission length n, an integer ν_1, and a conference index set $\mathcal{N}_1 = \{1, 2, \ldots, \nu_1\}$. Let $\mathcal{M}_1 = \{1, 2, \ldots, \mu_1\}$ and $\mathcal{M}_2 = \{1, 2, \ldots, \mu_2\}$ stand for the message sets intended to user 1 and user 2, respectively, and let $\mathcal{M}_2' = \{1, 2, \ldots, \mu_2'\}$ be the set of residual messages, that user 2 can decode if the conference link is present. Throughout, μ_1, μ_2 and μ_2' are integers. The conference rate R_c and transmission rates (R_1, R_2, R_2') are defined as

$$R_c = \frac{1}{n}\log \nu_1, \quad R_k = \frac{1}{n}\log \mu_k, \quad R_2' = \frac{1}{n}\log \mu_2'$$

where $k = 1, 2$.

Definition 1 (A code, achievable rates, and capacity region): A $(2^{nR_c}, 2^{nR_1}, 2^{nR_2}, 2^{nR_2'}, n)$ code for the AVBC $\mathcal{B}$ with unreliable conference link of capacity C_1 is an encoder mapping

$$f: \ \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_2' \to \mathcal{X}^n$$

a conference mapping

$$\phi: \ \mathcal{Y}_1^n \longrightarrow \mathcal{N}_1$$

and three decoder mappings

$$\begin{aligned} g_1: \ \mathcal{Y}_1^n &\longrightarrow \mathcal{M}_1 \\ g_2: \ \mathcal{Y}_2^n &\longrightarrow \mathcal{M}_2 \\ g_2': \ \mathcal{Y}_2^n \times \mathcal{N}_1 &\longrightarrow \mathcal{M}_2' \end{aligned}$$

with the restriction $R_c \le C_1$. We denote the code by $\mathscr{C} = (f, \phi, g_1, g_2, g_2')$. The conditional probabilities of error given a state sequence $s^n \in \mathcal{S}^n$, for the two cases where the link is absent and when it is present, are given by

$$\begin{aligned} P_{e|s^n}(\mathscr{C}) =& \frac{1}{\mu_1\mu_2\mu_2'} \sum_{m_1=1}^{\mu_1} \sum_{m_2=1}^{\mu_2} \sum_{,m_2'=1}^{\mu_2'} \\ & W_{Y_1^n,Y_2^n|X^n,S^n}(\mathcal{D}^c(m_1,m_2)|f(m_1,m_2,m_2'),s^n) \\ P'_{e|s^n}(\mathscr{C}) =& \frac{1}{\mu_1\mu_2\mu_2'} \sum_{m_1=1}^{\mu_1} \sum_{m_2=1}^{\mu_2} \sum_{,m_2'=1}^{\mu_2'} \\ & W_{Y_1^n,Y_2^n|X^n,S^n}(\mathcal{D}^c(m_1,m_2,m_2')|f(m_1,m_2,m_2'),s^n) \end{aligned}$$

where the decoding sets $\mathcal{D}(m_1, m_2)$ and $\mathcal{D}(m_1, m_2, m_2')$ are defined as

$$\mathcal{D}(m_1,m_2) = \{(y_1^n, y_2^n) : g_1(y_1^n) = m_1, g_2(y_2^n) = m_2\} \quad (1a)$$

$$\mathcal{D}(m_1,m_2,m_2') = \mathcal{D}(m_1,m_2) \cap \{y_2^n : g_2'(y_2^n) = m_2'\} \quad (1b)$$

The average probabilities of error of the code $\mathscr{C}$ given a state PMF $q \in \mathcal{P}(\mathcal{S}^n)$ are

$$P_e(q,\mathscr{C}) = \sum_{s^n \in \mathcal{S}^n} q(s^n) P_{e|s^n}(\mathscr{C}) \quad (2)$$

and similarly for $P_e'(q, \mathscr{C})$. We say that $\mathscr{C}$ is $(2^{nR_c}, 2^{nR_1}, 2^{nR_2}, 2^{nR_2'}, n, \epsilon)$ code for the AVBC $\mathcal{B}$ if it further satisfies

$$P_e(q,\mathscr{C}) \le \epsilon \ \text{ and } P_e'(q,\mathscr{C}) \le \epsilon \ \forall q \in \mathcal{P}(\mathcal{S}^n) \quad (3)$$

A rate triplet (R_1, R_2, R_2') is said to be achievable with unreliable conference link of capacity C_1 if for any $\epsilon > 0$ and sufficiently large n there exists a $(2^{nR_c}, 2^{nR_1}, 2^{nR_2}, 2^{nR_2'}, n, \epsilon)$ code for the AVBC $\mathcal{B}$, with $R_c \le C_1$. The capacity region is the closure of the set of all achievable rates, and is denoted by $\mathbb{C}(\mathcal{B})$.

Based on Definition 1, we can define now random codes. A $(2^{nR_c}, 2^{nR_1}, 2^{nR_2}, 2^{nR_2'}, n)$ random code for the channel $\mathcal{B}$ consists of a collection of (deterministic) $(2^{nR_c}, 2^{nR_1}, 2^{nR_2}, 2^{nR_2'}, n)$ codes $\{\mathscr{C}_\gamma\}_{\gamma\in\Gamma}$ with a probability distribution $\mu(\gamma)$ over the set Γ. It is denoted by $\mathscr{C}^\Gamma = (\mu, \Gamma, \{\mathscr{C}_\gamma\}_{\gamma\in\Gamma})$. We say that $\mathscr{C}^\Gamma$ is a $(2^{nR_c}, 2^{nR_1}, 2^{nR_2}, 2^{nR_2'}, n, \epsilon)$ random code for $\mathcal{B}$ if

$$P_e(q,\mathscr{C}^\Gamma) \triangleq \sum_{\gamma\in\Gamma} \mu(\gamma) P_e(q,\mathscr{C}_\gamma) \le \epsilon \ \forall q \in \mathcal{P}(\mathcal{S}^n) \quad (4a)$$

$$P_e'(q,\mathscr{C}^\Gamma) \triangleq \sum_{\gamma\in\Gamma} \mu(\gamma) P_e'(q,\mathscr{C}_\gamma) \le \epsilon \ \forall q \in \mathcal{P}(\mathcal{S}^n) \quad (4b)$$

Achievable rates for $\mathcal{B}$ with random coding are defined as usual. The random code capacity region is the closure of all rates achievable with random codes, and is denoted by $\mathbb{C}^\star(\mathcal{B})$.

The capacity region of the CBC $\mathcal{B}^{\mathcal{Q}}$, denoted by $\mathbb{C}(\mathcal{B}^{\mathcal{Q}})$, is defined similarly with the exception that the state distribution q in (3) is memoryless and restricted to reside in $\mathcal{Q}$.

Note that when $\mathcal{Q} = \{q\}$, i.e. contains a single element, the CBC reduces to the channel studied in [18] and [11]. We denote this channel by $\mathcal{B}^q$.

B. Channel properties

We denote by $W_{Y_1|X,S}$ and $W_{Y_2|X,S}$ the marginal channels. The AVBC $\mathcal{B}$ is said to be *physically degraded* if for any $q \in \mathcal{P}(\mathcal{S})$ we can write:

$$\sum_{s\in\mathcal{S}} q(s)W_{Y_1,Y_2|X,S}(y_1,y_2|x,s) = W_{Y_1|X}(y_1|x)W_{Y_2|Y_1}(y_2|y_1) \tag{5}$$

where the conditional distributions $W_{Y_1|X}$, $W_{Y_2|Y_1}$ may depend on q. This requirement holds if

$$W_{Y_1,Y_2|X,S} = W_{Y_1|X,S}W_{Y_2|Y_1} \tag{6a}$$

or

$$W_{Y_1,Y_2|X,S} = W_{Y_1|X}W_{Y_2|Y_1,S} \tag{6b}$$

In the sequel we will assume that the channel $W_{Y_1,Y_2|X,S}$ is either of the form (6a) or (6b). We turn now to the definition of symmetrizability, which plays a central role in the relations between $\mathbb{C}^{\star}(\mathcal{B})$ and $\mathbb{C}(\mathcal{B})$.

Definition 2: ([9],[6]) A discrete memoryless state dependent channel $W_{Y|X,S}$ is said to be *symmetrizable* if there exists a channel $J(s|x)$ such that

$$\sum_{s\in\mathcal{S}} W_{Y|X,S}(y|x_1,s)J(s|x_2) = \sum_{s\in\mathcal{S}} W_{Y|X,S}(y|x_2,s)J(s|x_1) \qquad \forall x_1,x_2,y.$$

In [9] Ericson stated that if a single user AVC is symmetrizable, then its capacity is zero. Csiszár and Narayan [6] showed that non-symmetrizability is a sufficient condition for the capacity to coincide with the random code capacity, when no constraints are imposed. Hof and Bross [10] showed that for the AVBC with degraded message sets and without conferencing, the interior of the capacity region is non-empty if and only if the marginals $W_{Y_1|X,S}$ and $W_{Y_2|X,S}$ are non-symmetrizable.

Remark 1 (symmetrizability and physical degradedness): If the AVBC $\mathcal{B}$ is physically degraded in the form (6a), then symmetrizability of $W_{Y_1|X,S}$ implies symmetrizability of $W_{Y_1,Y_2|X,S}$ and consequently also that of $W_{Y_2|X,S}$. However, symmetrizability of $W_{Y_2|X,S}$ does not imply that of $W_{Y_1|X,S}$, and therefore neither it implies symmetrizability of $W_{Y_1,Y_2|X,S}$. If the AVBC is physically degraded in the form (6b), $W_{Y_1|X,S}$ is not symmetrizable by definition, except for the case where Y_1 is independent of X. $W_{Y_2|X,S}$ may or may not be symmetrizable.

III. Main results

A. The compound channel

We start by stating the results for the compound channel model. Define the sets

$$\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}}) \triangleq \bigcup_{P_{U,V,X}} \bigcap_{q\in\mathcal{Q}} \left\{ (R_1,R_2,R_2') : \begin{array}{rl} R_2 \le & I_q(U;Y_2) \\ R_2' \le & I_q(V;Y_2|U) + C_1 \\ R_2' \le & I_q(V;Y_1|U) \\ R_1 \le & I_q(X;Y_1|U,V) \end{array} \right\} \tag{7}$$

where U, V are external random variables with alphabets $\mathcal{U}, \mathcal{V}$, respectively, $P_{U,V,X}$ is an arbitrary distribution on $\mathcal{U}\times\mathcal{V}\times\mathcal{X}$, $I_q(U;Y_2)$ stands for the mutual information between U and Y_2 when the state is iid and distributed according to q, and the union is over $\mathcal{P}(\mathcal{U}\times\mathcal{V}\times\mathcal{X})$. Next, define

$$\mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}}) \triangleq \bigcap_{q\in\mathcal{Q}} \bigcup_{P_{U,V,X}} \left\{ (R_1,R_2,R_2') : \begin{array}{rl} R_2 \le & I_q(U;Y_2) \\ R_2' \le & I_q(V;Y_2|U) + C_1 \\ R_2' \le & I_q(V;Y_1|U) \\ R_1 \le & I_q(X;Y_1|U,V) \end{array} \right\} \tag{8}$$

Since in (7) the intersection is the inner operation, it can be expressed as

$$\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}}) \triangleq \bigcup_{P_{U,V,X}} \left\{ (R_1,R_2,R_2') : \begin{array}{rl} R_2 \le & \inf\limits_{q\in\mathcal{Q}} I_q(U;Y_2) \\ R_2' \le & \inf\limits_{q\in\mathcal{Q}} I_q(V;Y_2|U) + C_1 \\ R_2' \le & \inf\limits_{q\in\mathcal{Q}} I_q(V;Y_1|U) \\ R_1 \le & \inf\limits_{q\in\mathcal{Q}} I_q(X;Y_1|U,V) \end{array} \right\} \tag{9}$$

The next lemma states that these are inner and outer bounds on the capacity of the CBC.

Lemma 1: For any physically degraded CBC $\mathcal{B}^{\mathcal{Q}}$ with unreliable conference link of capacity C_1,

$$\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}}) \subseteq \mathbb{C}(\mathcal{B}^{\mathcal{Q}}) \subseteq \mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})$$

Moreover, if $(R_1,R_2,R_2') \in \mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}})$, then for some $a > 0$ and sufficiently large n, there exists a $(2^{nR_c}, 2^{nR_1}, 2^{nR_2}, 2^{nR_2'}, n, e^{-an})$ code for $\mathcal{B}^{\mathcal{Q}}$, with $R_c \le C_1$.

The exponential error estimate is needed in the proofs of Theorem 2 and Theorem 3. *Proof of Lemma 1*: The proof of the inner bound in Lemma 1 uses a coding scheme similar to the scheme used in [18], [11] for the classical (non-AVC) physically degraded BC with unreliable conference. The decoder performs classical joint typicality technique with a search over a fine (but finite) quantization of the set $\mathcal{Q}$, as in [17, Lemma 5]. The proof of the outer bound resembles that of [18], [11]. Due to lack of space, the details are omitted. □

When $\mathcal{Q} = \{q\}$ the bounds coincide with the capacity region derived in [18] and [11]. Observe that the difference between the inner and outer bounds is the order of union and intersection. The next definitions provide conditions, in the spirit of [17], under which the order can be interchanged. We say that $\mathcal{D} \subseteq \mathcal{P}(\mathcal{U} \times \mathcal{V} \times \mathcal{X})$ achieves both $\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}})$ and $\mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})$ if the union operations in (7) and (8) can be taken over $\mathcal{D}$ instead of the whole collection $\mathcal{P}(\mathcal{U}\times\mathcal{V}\times\mathcal{X})$. Clearly, if $\mathcal{D} \subseteq \mathcal{P}(\mathcal{U}\times\mathcal{V}\times\mathcal{X})$ achieves $\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}})$ and $\mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})$, then so does any $\mathcal{D}'$ that contains $\mathcal{D}$. Thus using $\mathcal{D}$ can be beneficial only if we can minimise it in some sense; in particular, the following property may hold for $\mathcal{D}$ but not for $\mathcal{D}'$.

Definition 3: Let $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{S})$ be a compact set of state distributions, and let $\mathcal{D}$ be a set that achieves $\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}})$ and $\mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})$. We say that Condition $\mathscr{I}^{\mathcal{Q}}$ holds if there exists $q^* \in \mathcal{Q}$ that simultaneously minimises the functions $I_q(U;Y_2)$, $\min\{I_q(V;Y_2|U)+C_1, I_q(V;Y_1|U)\}$ and $I_q(X;Y_2|U,V)$ for all $P_{U,V,X} \in \mathcal{D}$.

The operational meaning of Condition $\mathscr{I}^{\mathcal{Q}}$ is that there exists a state strategy q^* that is the worst strategy for both users, under both conditions of the link - present or absent. We have the following result.

Theorem 1: Let $\mathcal{B}^{\mathcal{Q}}$ be a physically degraded CBC with unreliable cooperation link of capacity C_1. If Condition $\mathscr{I}^{\mathcal{Q}}$ holds, then

$$\mathbb{C}(\mathcal{B}^{\mathcal{Q}}) = \mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}}) = \mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})$$

Proof of Theorem 1: We only have to show that under Condition $\mathscr{I}^{\mathcal{Q}}$, $\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}}) \supseteq \mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})$. Since $\mathcal{D}$ achieves $\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}})$ and $\mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})$, we can write

$$\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}}) = \bigcup_{P_{U,V,X}\in\mathcal{D}} \left\{ (R_1,R_2,R_2') : \begin{array}{rl} R_2 \le & \inf\limits_{q\in\mathcal{Q}} I_q(U;Y_2) \\ R_2' \le & \inf\limits_{q\in\mathcal{Q}} I_q(V;Y_2|U)+C_1 \\ R_2' \le & \inf\limits_{q\in\mathcal{Q}} I_q(V;Y_1|U) \\ R_1 \le & \inf\limits_{q\in\mathcal{Q}} I_q(X;Y_1|U,V) \end{array} \right\} \tag{10}$$

and

$$\mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}}) = \bigcap_{q\in\mathcal{Q}} \bigcup_{P_{U,V,X}\in\mathcal{D}} \left\{ (R_1,R_2,R_2') : \begin{array}{rl} R_2 \le & I_q(U;Y_2) \\ R_2' \le & I_q(V;Y_2|U)+C_1 \\ R_2' \le & I_q(V;Y_1|U) \\ R_1 \le & I_q(X;Y_1|U,V) \end{array} \right\} \tag{11}$$

Since Condition $\mathscr{I}^{\mathcal{Q}}$ holds, there exists $q^* \in \mathcal{Q}$ such that

$$\begin{aligned}\mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}}) = & \bigcup_{P_{U,V,X}\in\mathcal{D}} \left\{ (R_1,R_2,R_2') : \begin{array}{rl} R_2 \le & I_{q^*}(U;Y_2) \\ R_2' \le & I_{q^*}(V;Y_2|U)+C_1 \\ R_2' \le & I_{q^*}(V;Y_1|U) \\ R_1 \le & I_{q^*}(X;Y_1|U,V) \end{array} \right\} \\ \supseteq & \mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}}) \end{aligned} \tag{12}$$

where the last inclusion is due to the intersection in (11). □

B. The arbitrarily varying channel

We proceed to state the results for the physically degraded AVBC $\mathcal{B}$. Define the sets

$$\mathcal{R}_{in}^{\star}(\mathcal{B}) = \mathcal{R}_{in}(\mathcal{B}^{\mathcal{Q}})\big|_{\mathcal{Q}=\mathcal{P}(\mathcal{S})} \tag{13a}$$

$$\mathcal{R}_{out}^{\star}(\mathcal{B}) = \mathcal{R}_{out}(\mathcal{B}^{\mathcal{Q}})\big|_{\mathcal{Q}=\mathcal{P}(\mathcal{S})} \tag{13b}$$

$$\mathscr{I} = \mathscr{I}^{\mathcal{Q}}\big|_{\mathcal{Q}=\mathcal{P}(\mathcal{S})} \tag{13c}$$

We will need also the projections of $\mathcal{R}_{in}^{\star}(\mathcal{B})$ and $\mathcal{R}_{out}^{\star}(\mathcal{B})$ on the hyperplane $R_2 = 0$. Specifically, define

$$\widetilde{\mathcal{R}}_{in}(\mathcal{B}) = \mathcal{R}_{in}^{\star}(\mathcal{B})\big|_{R_2=0} \tag{14a}$$

$$\widetilde{\mathcal{R}}_{out}(\mathcal{B}) = \mathcal{R}_{out}^{\star}(\mathcal{B})\big|_{R_2=0} \tag{14b}$$

The regions $\widetilde{\mathcal{R}}_{in}(\mathcal{B})$ and $\widetilde{\mathcal{R}}_{out}(\mathcal{B})$ correspond to the case where we wish to transmit to user 2 only messages that can be decoded without the cooperation link. Thus the focus is entirely on the scenario where the conference link of capacity C_1 is active. Note that we can write

$$\widetilde{\mathcal{R}}_{in}(\mathcal{B}) = \bigcup_{P_{V,X}} \bigcap_{q\in\mathcal{P}(\mathcal{S})} \left\{ (R_1,0,R_2') : \begin{array}{rl} R_2' \le & I_q(V;Y_2)+C_1 \\ R_2' \le & I_q(V;Y_1) \\ R_1 \le & I_q(X;Y_1|V) \end{array} \right\} \tag{15}$$

and similarly for $\widetilde{\mathcal{R}}_{out}(\mathcal{B})$

$$\widetilde{\mathcal{R}}_{out}(\mathcal{B}) = \bigcap_{q\in\mathcal{P}(\mathcal{S})} \bigcup_{P_{V,X}} \left\{ (R_1,0,R_2') : \begin{array}{rl} R_2' \le & I_q(V;Y_2)+C_1 \\ R_2' \le & I_q(V;Y_1) \\ R_1 \le & I_q(X;Y_1|V) \end{array} \right\} \tag{16}$$

As in Definition 3, we can provide conditions under which the bounds coincide with $R_2 = 0$. Thus, we say that $\widetilde{\mathcal{D}} \subseteq \mathcal{P}(\mathcal{V}\times\mathcal{X})$ achieves both $\widetilde{\mathcal{R}}_{in}(\mathcal{B})$ and $\widetilde{\mathcal{R}}_{out}(\mathcal{B})$ if the union operations in (15) and (16) can be taken over $\widetilde{\mathcal{D}}$ instead of $\mathcal{P}(\mathcal{V}\times\mathcal{X})$.

Definition 4: Let $\widetilde{\mathcal{D}}$ be a set that achieves both $\widetilde{\mathcal{R}}_{in}(\mathcal{B})$ and $\widetilde{\mathcal{R}}_{out}(\mathcal{B})$. We say that Condition $\tilde{\mathscr{I}}$ holds if there exists $q^* \in \mathcal{P}(\mathcal{S})$ that simultaneously minimises $\min\{I_q(V;Y_2)+C_1, I_q(V;Y_1)\}$ and $I_q(X;Y_2|V)$ for all $P_{V,X} \in \mathcal{P}(\mathcal{V}\times\mathcal{X})$. Note that Condition $\tilde{\mathscr{I}}$ is milder than Condition $\mathscr{I}$ as it involves less constraints.

We proceed to state our results on the capacity region under random coding.

1) Random codes:

Theorem 2: Let $\mathcal{B}$ be a physically degraded AVBC with unreliable conference link of capacity C_1. Then

1) $\mathcal{R}_{in}^{\star}(\mathcal{B}) \subseteq \mathbb{C}^{\star}(\mathcal{B}) \subseteq \mathcal{R}_{out}^{\star}(\mathcal{B})$
2) If Condition $\mathscr{I}$ holds, then

$$\mathcal{R}_{in}^{\star}(\mathcal{B}) = \mathbb{C}^{\star}(\mathcal{B}) = \mathcal{R}_{out}^{\star}(\mathcal{B})$$

Proof of Theorem 2: Part 1: The proof of the inner bound uses Ahlswede's Robustification Technique (RT) [2], [3] (see also [4]). It makes use of the exponential error result in Lemma 1. The outer bound follows quite closely the proof

of the converse in [18] and [11], taking into account also the common randomness. The details are omitted. The proof of Part 2 is similar to the proof of Theorem 1. □

The symmetrizability conditions, the sets $\widetilde{\mathcal{R}}_{in}(\mathcal{B})$ and $\widetilde{\mathcal{R}}_{out}(\mathcal{B})$, and Condition $\tilde{\mathscr{I}}$ play a role in the characterisation of the capacity region with deterministic codes, stated next.

2) Deterministic codes:

Theorem 3: For any physically degraded AVBC with unreliable cooperation link, the following hold

1) If $W_{Y_1|X,S}$ is symmetrizable, then
$$\mathbb{C}(\mathcal{B}) = (0, 0, 0)$$
2) If $W_{Y_1|X,S}$ and $W_{Y_2|X,S}$ are non-symmetrizable, then
$$\mathbb{C}(\mathcal{B}) = \mathbb{C}^{\star}(\mathcal{B})$$
3) If $W_{Y_1|X,S}$ is non-symmetrizable and $W_{Y_2|X,S}$ is symmetrizable, then
$$\widetilde{\mathcal{R}}_{in}(\mathcal{B}) \subseteq \mathbb{C}(\mathcal{B}) \subseteq \widetilde{\mathcal{R}}_{out}(\mathcal{B})$$
4) If $W_{Y_1|X,S}$ is non-symmetrizable, $W_{Y_2|X,S}$ is symmetrizable and Condition $\tilde{\mathscr{I}}$ holds, then
$$\mathbb{C}(\mathcal{B}) = \widetilde{\mathcal{R}}_{in}(\mathcal{B}) = \widetilde{\mathcal{R}}_{out}(\mathcal{B})$$

Note that if the marginal channels $W_{Y_1|X,S}$ and $W_{Y_2|X,S}$ are non-symmetrizable, Part 2 of Theorem 3 characterises the capacity region for the case that Condition $\mathscr{I}$ holds, by Part 2 of Theorem 2.

Proof of Theorem3: For the proof of Part 1, observe that if $W_{Y_1|X,S}$ is symmetrizable, then the channel is in the form of (6a), $W_{Y_1,Y_2|X,S}$ is symmetrizable, and Part 1 follows from previous results. Part 2 follows classical arguments, using Ahlswede's Elimination Technique and transferring the residual common randomness to both users. For the proof of Part 3, note that if $W_{Y_2|X,S}$ is symmetrizable, there is no point in transmitting to user 2 messages that cannot be decoded without the conference link. Thus all the transmission to user 2 is delegated to the case where the cooperation link is active. In that case, the residual common randomness (after applying the Elimination Technique) can be transferred to user 2 via the conference link, from user 1 (whose channel is not symmetrizable). With this approach, applying classical arguments (Robustification and Elimination) yield the inner bound. As for the outer bound, note that random coding outer bounds apply also to deterministic codes. The restriction to $R_2 = 0$ in the definition of $\widetilde{\mathcal{R}}_{out}(\mathcal{B})$ follows from the same argument as in the inner bound. The proof of Part 4 follows the same lines as that of Theorem 1. □

REFERENCES

[1] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 44.2, pp. 159-175, June 1978.

[2] R. Ahlswede, "Coloring hypergraphs: A new approach to multi-user source coding, Part 1," *J. Combinatorics* 4.1 (1979), pp. 76-115.

[3] R. Ahlswede, "Coloring hypergraphs: A new approach to multi-user source coding, Part 2," *J. Combinatorics* 5.3 (1980), pp. 220-268.

[4] R. Ahlswede, "Arbitrarily varying channels with state sequence known to the sender," *IEEE Trans. Inform. Theory*, vol. 32, no. 5, pp. 621-629, Sept. 1986.

[5] D. Blackwell, L. Breiman and A. J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math.Statist.* 31.3 (Sept. 1960), pp. 558-567.

[6] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: positivity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 181-193, March 1988.

[7] R. Dabora and S. D. Servetto, "Broadcast channels with cooperating receivers: A downlink for the sensor reachback problem," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, July 2004, p. 176.

[8] R. Dabora and S. D. Servetto, "Broadcast channels with cooperating decoders," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5438-5454, 2006.

[9] T. Ericson, "Exponential error bounds for random codes in the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 31, no. 1, pp. 42-48, Jan. 1985.

[10] E. Hof and S. I. Bross, "On the deterministic-code capacity of the two-user discrete memoryless arbitrarily varying general broadcast channel with degraded message sets," *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 5023-5044, Nov. 2006.

[11] W. Huleihel and Y. Steinberg, "Channels with cooperation links that may be absent," *IEEE Trans. Inform. Theory*,vol. 63, no. 9, pp 5886-5906, Sept. 2017.

[12] D. Itzhak and Y. Steinberg, "The broadcast channel with degraded message sets and unreliable conference," [Online]. Available: https://arxiv.org/abs/1701.05780

[13] D. Itzhak and Y. Steinberg, "The broadcast channel with degraded message sets and unreliable conference," in *Proc. IEEE Symp. Information Theory*, Aachen, Germany, June 2017.

[14] J. H. Jahn, "Coding for arbitrarily varying multiuser channels," *IEEE Trans. Inform. Theory*,vol. 27, no. 2, pp 212-226, March 1981

[15] Y. Liang and V. V. Veeravalli, "The impact of relaying on the capacity of broadcast channels," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, July 2004, p. 403.

[16] Y. Liang and V. V. Veeravalli, "Cooperative relay broadcast channels," *IEEE Trans. Inform. Theory*,vol. 53, no. 3, pp 900-928, March 2007.

[17] U. Pereg and Y. Steinberg, "The arbitrarily varying broadcast channel with causal side information at the encoder," *IEEE Trans. Inform. Theory*, accepted.

[18] Y. Steinberg, "Channels with cooperation links that may be absent," in *Proc. IEEE Int. Symp. Information Theory*, Honolulu, HI, June 29-July 4, 2014.

[19] M. Wiese, *Multiple access channels with cooperating encoders*, Ph.D. Thesis, Technische Universität Münich, 2013.

[20] M. Wiese and H. Boche, "The arbitrarily varying multiple-access channel with conferencing encoders," *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp 1405-1416, Mar. 2013.

The Duality Upper Bound for Unifilar Finite-State Channels with Feedback

Oron Sabag
California Institute of Technology
Pasadena, CA 91125, USA
Email: oron@caltech.edu

Haim H. Permuter
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel
Email: haimp@bgu.ac.il

***Abstract*—The capacity of finite-state channels (FSCs) with feedback is considered. We derive a simple upper bound on the feedback capacity. The upper bound is based on an extension of the known duality upper bound on mutual information to the case of directed information. The upper bound is a function of test distributions on the channel outputs ensemble. We show that if the test distribution is structured on a Q-graph, and the FSC is unifilar, the upper bound can be formulated as a Markov decision process (MDP). The resulted MDP has finite states, actions and disturbances and, therefore, can be solved analytically with standard MDP tools. We illustrate the simplicity of computing the bounds for the dicode erasure and the Ising channels. The resulted upper bounds are tight and their derivation serves as an alternative and simple converse proof. The developed methodology is capable of solving channels with large alphabets of the channel states, inputs and outputs.**

I. Introduction

[1] Finite-state channels (FSCs) are commonly used to model scenarios in which the channel or the system have memory. In this paper, we focus on unifilar FSCs with feedback (Fig. 1). A useful approach for computing the feedback capacity of such channels is via a Markov decision processes (MDPs) formulation of the capacity expression [1]–[3]. In many cases, applying dynamic programming methods or reinforcement learning algorithms lead to a solution that is conjectured to be optimal. However, to conclude its optimality, a solution for the involved Bellman equation is needed.

In a recent paper, we proposed a simple upper bound on the feedback capacity of unifilar FSCs [4]. The upper bound is a single-letter expression and holds for any choice of a *Q-graph*, an auxiliary structure to map channel outputs sequences. It was further shown that the bound is a standard convex optimization problem [5], [6]. The bound led to new capacity results, but is still challenging to analytically compute when the channel parameters have large alphabets. The recent development of reinforcement learning algorithms to (numerically) compute and conjecture optimal solution on feedback capacity for large alphabets [7] motivated the current paper.

In this paper, we derive a new bound that fits a framework of channels with large alphabets. Its derivation is based on the dual upper bounding technique [8]–[10], but for directed information. The resulted duality bound is a multi-letter formula that holds for any choice of a test distribution on the channel outputs ensemble. The main contribution is a formulation of the upper bound as an MDP. Specifically, we present an MDP formulation that for any unifilar FSC and test distributions that are chosen on a Q-graph. Due to the finite alphabets in the MDP formulation, simple numerical and analytical MDP tools can be applied.

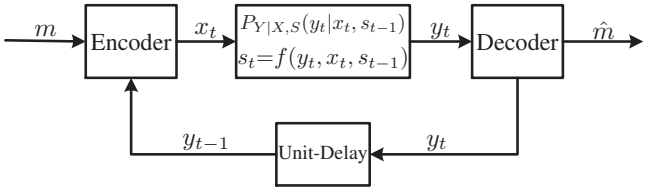


Fig. 1. Unifilar FSC with feedback. The new channel state, s_t, is a function of (y_t, x_t, s_{t-1}).

The remainder of this paper is organized as follows. Section II introduces the notation and the model definition. Section III introduces the dual capacity upper bound, Q-graphs and our main result. Section IV contains the MDP formulation of the upper bounds. In Section V, we illustrate the simplicity of computing the upper bound for two examples via solutions to the Bellman equation. Finally, proofs are given in Section VI.

II. The setting

In this section, we present the notation and the setting.

A. Notation

Random variables, realizations and alphabets are denoted by upper-case letters (e.g., X), lower-case letters (e.g., x) and calligraphic letters (e.g., $\mathcal{X}$), respectively. All vectors follow the above notation with superscripts, e,g., $x^n = (x_1, x_2, \ldots, x_n)$. The probability mass function (pmf) of a random variable X is denoted by P_X, and conditional and joint pmfs are denoted by $P_{Y|X}$ and $P_{X,Y}$, respectively; when the random variables are clear from the context we omit the random variable, i.e., $P(x)$, $P(y|x)$ and $P(x, y)$. We use the standard notation of directed information, as in [11],

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1})$$

[1]The work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) via the Deutsch-Israelische Projektkooperation (DIP), in part by the Israel Science Foundation, and in part by the Cyber Center and at Ben-Gurion University of the Negev. The work of O. sabag was supported by the ISEF international fellowship.

and causal conditioning

$$P(X^n||Y^n) = \prod_{i=1}^{n} P(X_i|X^{i-1}, Y^{i-1}).$$

When the causal conditioning is particularized for deterministic functions we write $f(x^n||y^n)$.

B. Finite-state channels

A *finite state channel* is defined by the triplet $(\mathcal{X} \times \mathcal{S}, P(s, y|x, s'), \mathcal{Y} \times \mathcal{S})$ where X is the channel input, Y is the channel output, S' is the channel state at the beginning of the transmission and S is the channel state at the end of the transmission. The cardinalities of X, Y, S are assumed to be finite. At each time t, the channel has a time-invariant probabilistic characterization

$$P(s_t, y_t|x^t, s^{t-1}, y^{t-1}) = P(s_t, y_t|x_t, s_{t-1}).$$

The FSC is called unifilar is the state evolution is given by a time-invariant and deterministic function. That is, $P(s_t|y_t, x_t, s_{t-1}) = \mathbb{1}\{s_t = f(y_t, x_t, s_{t-1})\}$. The capacity of a strongly connected FSC with feedback is given by the following theorem:

Theorem 1 ([1, Th. 3]). *The feedback capacity of a strongly connected FSC is*

$$C_{\text{fb}} = \lim_{n\to\infty} \frac{1}{n} \max_{p(x^n||y^{n-1})} I(X^n \to Y^n).$$

III. Main Result - Computable Upper Bound

In this section, we present the duality upper bound and its extension to the directed information. Then, we simplify the general bound for the case of unifilar FSCs and graph-based test distributions. Finally, its computability will be shown via a novel MDP formulation.

A. The duality upper bound

For a memoryless channel, the following upper bound is well known:

Theorem 2 ([12, Th. 8.4]). *For any test distribution T_Y and a memoryless channel, the capacity is bounded by*

$$C \le \max_x D(P_{Y|X=x}||T_Y).$$

The bound follows from the non-negativity of the KL divergence $D(P_Y||T_Y)$. It is called the *duality upper bound* since it can be simply deduced from the dual capacity expression [13]. If the test distribution is equal to the optimal outputs distribution, the bound is tight. Therefore, one should aim at choosing T_Y as close as possible to the optimal P_Y^*.

For the directed information, the same methodology gives the following result.

Theorem 3 (Duality UB for Directed Information). *For a fixed channel $P(y^n||x^n)$ and any test distribution $T(y^n)$,*

$$C_n \triangleq \frac{1}{n} I(X^n \to Y^n)$$

$$\le \frac{1}{n} \max D(P(Y^n||X^n = x^n)||T(Y^n)),$$

where the maximum is over $f(X^n||Y^{n-1})$.

Note that the theorem holds for any channel. The optimal output distribution for channels with memory is not i.i.d.. Therefore, in order to minimize the upper bound, one needs to choose a test distribution with memory. Markov test distributions are standard choice in the literature, but it can be shown that the optimal outputs distribution for certain channels do not admit a Markovian structure (of any finite order). In this paper, we choose the test distribution as an extension of the Markov model to be a variable-order Markov model on a *Q-graph*.

B. The Q-graph

The Q-graph is defined as a directed graph with edges that are labelled with symbols from the channel outputs alphabet $\mathcal{Y}$. It also has the property that the outgoing edges from each node are disjointly labelled with all possible labels from $\mathcal{Y}$ (See Fig. 2). Thus, the Q-graph can be used as a mapping of (any-length) output sequences into the graph nodes by walking along the labelled edges. For a fixed graph, we denote the induced mapping with

$$\Phi : \mathcal{Q} \times \mathcal{Y} \to \mathcal{Q},$$

where $\mathcal{Q}$ denotes the set of graph nodes.

Remark 1 (Variable-order Markov model). *A special case of the variable-order Markov model on a Q-graph is a Markov model of order k. This can be seen by choosing a graph with $\mathcal{Y}^k$ nodes, where each node represents a tuple of k channel outputs and the edges are connected accordingly.*

C. Main result

The following is the duality upper bound in Theorem 3 when simplified to Q-graph test distributions and unifilar FSCs.

Theorem 4 (Duality UB for FSCs using Q-graphs). *For any Q-graph test distribution, the feedback capacity of a strongly connected unifilar FSC is bounded by*

$$C_{\text{fb}} \le \lim_{n\to\infty} \max_{f(x^n||y^{n-1})} \min_{s_0, q_0} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Y^{i-1}||X^{i-1}}[D(P(\cdot|x_i, S_{i-1})||T(\cdot|Q_{i-1}))]. \quad (1)$$

The notation $f(x^n||y^{n-1})$ stands for causal conditioning of deterministic functions:

$$f(x^n||y^{n-1}) = \prod_i \mathbb{1}\{x_i = f_i(x^{i-1}, y^{i-1})\}.$$

The following is our main result that shows the simple computation required in order to evaluate the RHS of (1).

Theorem 5 (MDP formulation). *The upper bound* (1) *is equal to the optimal average reward of an MDP with finite states, actions and disturbances.*

Theorem 5 is a computational result. MDPs are well-investigated in the optimization and control literatures (e.g.,

[14]–[16]). When the MDP states, actions and disturbances have finite spaces, the MDP can be solved using standard dynamic programming methods such as the value and policy iterations. Moreover, the Bellman equation can be used to simplify the involved upper bound in (1) to very simple expressions. It will be shown in Section V that the Bellman equation provides tight and simple upper bounds if the test distribution is chosen correctly.

The next section concerns with the MDP formulation.

IV. MDP FORMULATION

Markov decision processes are common models for sequential decisions making. In this paper, we focus on the class of average-reward infinite horizon MDPs. It will be shown that their optimal average rewards are equal to the upper bound in Theorem 3.

The MDP state at time t is defined as the pair $z_{t-1} \triangleq (q_{t-1}, s_{t-1})$. The action is x_t, the disturbance is y_t, and the reward is

$$R(z_{t-1}, x_t) = D\left(P_{Y|X,S}(\cdot|x_t, s_{t-1})||T_{Y|Q}(\cdot|q_{t-1})\right) \quad (2)$$

Before we proceed to the connection between the optimal average reward of the defined MDP and the upper bound on the capacity, we technically show that this is a valid MDP.

Lemma 1 (MDP formulation)**.** *For the MDP in Table I:*

1) *The MDP state, z_t, is a function of z_{t-1}, y_t, x_t.*
2) *The MDP reward is time-invariant function of z_{t-1}, x_t.*
3) *The MDP disturbance, y_t, is conditionally independent of the past, given z_{t-1}, x_t.*

Proof. Straightforward computations using the Markov chain $Y_t - (X_t, S_{t-1}) - \Psi(X^{t-1}, Y^{t-1}, S^{t-1})$ for any function $\Psi(\cdot)$. □

Following Lemma 1, one can define the MDP average reward in the infinite horizon regime as

$$J^* = \sup \liminf_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[D\left(P_{Y|X,S}(\cdot|x_t, S_{t-1})||T_{Y|Q}(\cdot|Q_{t-1})\right)\right], \quad (3)$$

where the supremum is over all deterministic functions $\{f_i : \mathcal{X}^{i-1} \times \mathcal{Y}^{i-1} \to \mathcal{X}\}_{i\geq 1}$. Indeed, the MDP formulation in Lemma 1 provides a powerful conclusion on the maximization domain:

Corollary 1. *It is sufficient to evaluate the supremum in* (3) *over $\{f_i : \mathcal{S} \times \mathcal{Q} \to \mathcal{X}\}_{i\geq 1}$.*

The following theorem relates the upper bound in (1) to the average reward of the defined MDP.

Theorem 6. *The optimal average reward of the MDP is an upper bound to the capacity. That is, $C_{\text{fb}} \leq J^*$.*

TABLE I
SUMMARY OF THE MDP FORMULATION

	MDP
State	(q_{t-1}, s_{t-1})
Action	x_t
Disturbance	y_t
Reward	$D(P_{Y\|X,S}(\cdot\|x_t, s_{t-1})\|\|T_{Y\|Q}(\cdot\|q_{t-1}))$

A. The Bellman equation

An alternative characterization for the optimal average reward in MDPs is offered by the Bellman equation. The following theorem is a simplification of the Bellman equation for our formulation.

Theorem 7 (Bellman equation, [14])**.** *If $\rho \in \mathbb{R}$ and a bounded function $h : \mathcal{X} \times \mathcal{Q} \to \mathbb{R}$ satisfies*

$$\rho + h(s, q) = \max_x R(z, x) + \mathbb{E}[h(S^+, \Phi(Y, q)|S = s, X = x],$$

for all (s, q), then $\rho = \rho^*$.

V. EXAMPLES

In this section, we study two known examples, the DEC and the Ising channels. The main objective here is to provide simple proofs for tight upper bounds. For both channels, we establish their tight upper bounds via an explicit solution of the Bellman equation.

The MDP formulation: For both channels, the channel state is the previous channel input x_{t-1}. The MDP formulation in this case is the following: the MDP state at time t is (x_{t-1}, q_{t-1}), the MDP action is x_t, the reward is $D(P_{Y|X,X^-}(\cdot|x_t, x_{t-1})||T_{Y|Q}(\cdot|q_{t-1})$ and the disturbance is y_t.

For convenience, we define the MDP operator on a function $h : \mathcal{X} \times \mathcal{Q} \to \mathbb{R}$ as:

$$(Th)(x^-, q) = \max_x D(P_{Y|X,X^-}(\cdot|x, x^-)||T_{Y|Q}(\cdot|q)) + \mathbb{E}[h(x, \Phi(Y, q))|X = x, X^- = x^-]. \quad (4)$$

A. The DEC

The dicode erasure channel (DEC) is a binary-input channel whose characterization is given by:

$$Y_i = \begin{cases} X_i - X_{i-1} & \text{, w.p.1} - \epsilon \\ ? & \text{, w.p. } \epsilon. \end{cases}$$

The channel outputs can take values from $\mathcal{Y} = \{0, \pm 1, ?\}$. The channel was introduced as a simplified version of the known dicode channel (with Gaussian noise). Its feedback capacity was investigated in [17], [18], but only a lower bound could be deduced. In [4], we developed a new framework that is called the Q-graph upper bounds and showed that the lower bound is indeed tight:

Theorem 8 (DEC capacity, [4])**.** *The feedback capacity of the DEC is:*

$$C_{\text{DEC}} = \max_{0\leq p\leq 1} (1 - \epsilon)\frac{p + \epsilon H_2(p)}{\epsilon + (1 - \epsilon)p}. \quad (5)$$

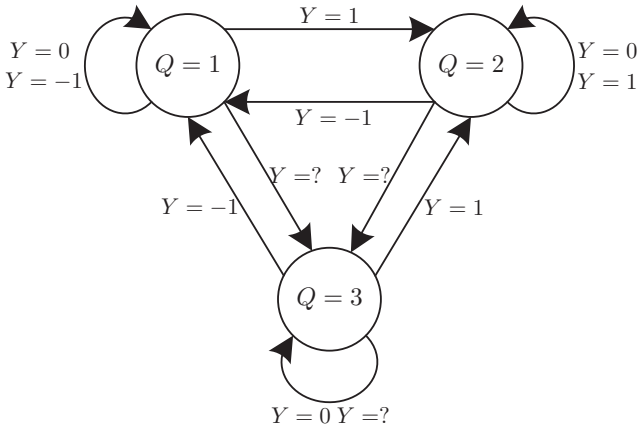


Fig. 2. The optimal Q-graph for the DEC.

In the following, we provide our new result, an alternative and simple converse proof for the above capacity result.

Theorem 9 (Upper bound for the DEC)**.** *The feedback capacity of the DEC satisfies*

$$C_{\text{DEC}} \leq \bar{\epsilon}(1-\epsilon\log p), \tag{6}$$

where p is the unique solution for $x^{\epsilon}=2(1-x)$.

It can be shown that (6) is equal to the feedback capacity in Theorem 8, so that the upper bound is tight.

Solution for the MDP: Recall that in order to compute the upper bound, one needs a graph-based test distribution on a Q-graph. Consider the Q-graph in Fig. 2 and the test distribution $T_{Y|Q}$ that is parameterized with $p\in[0,1]$:

	$Y=?$	$Y=-1$	$Y=0$	$Y=1$
$Q=1$	ϵ	0	$0.5\bar{\epsilon}$	$0.5\bar{\epsilon}$
$Q=2$	ϵ	$0.5\bar{\epsilon}$	$0.5\bar{\epsilon}$	0
$Q=3$	ϵ	$0.5p\bar{\epsilon}$	$(1-p)\bar{\epsilon}$	$0.5p\bar{\epsilon}$

By iterating the value iteration algorithm, one can deduce a conjectured solution[2]: define the constant

$$\rho^* = \bar{\epsilon}(1-\epsilon\log p), \tag{7}$$

where p is the unique solution for $x^{\epsilon}=2(1-x)$. Also, define the function $h:\mathcal{X}\times\mathcal{Q}\to\mathbb{R}$:

$$\begin{aligned} h(0,1)=h(1,2)&=\bar{\epsilon}\bar{\epsilon}\log p\\ h(0,3)=h(1,3)&=-\bar{\epsilon}\bar{\epsilon}\log p. \end{aligned} \tag{8}$$

The value function at $h(0,2)$ and $h(1,1)$ are not defined since they have zero probability for any choice of actions when choosing the initial state to be $Q=3$.

The following technical result concludes Theorem 9.

Lemma 2. *The constant ρ^* and the function h solve the Bellman equation. Consequently, $C_{DEC}\leq\rho^*$.*

Proof of Lemma 2. We compute explicitly the MDP operator in (4) with (8).

[2]Further reading on the numerical evaluation can be found in [1], [19]

For the MDP state $(x^-=0,q=1)$, the MDP operator is a maximum over

$$\begin{aligned} x=0:&D([\epsilon,0,\bar{\epsilon},0]|||[\epsilon,0,0.5\bar{\epsilon},0.5\bar{\epsilon}])+0.5h(1,1)+0.5h(0,3)\\ x=1:&D([\epsilon,0,0,\bar{\epsilon}]|||[\epsilon,0,0.5\bar{\epsilon},0.5\bar{\epsilon}])+0.5h(1,2)+0.5h(1,3). \end{aligned}$$

Both equations simplify to $\bar{\epsilon}$, so that $\rho^*+h(0,1)=\bar{\epsilon}$.

For the state $x^-=0$ and $q=3$, the MDP operator is the maximum on

$$\begin{aligned} x=0:&D([\epsilon,0,\bar{\epsilon},0]|||[\epsilon,0.5p\bar{\epsilon},(1-p)\bar{\epsilon},0.5p\bar{\epsilon}])+h(0,3)\\ x=1:&D([\epsilon,0,0,\bar{\epsilon}]|||[\epsilon,0.5p\bar{\epsilon},(1-p)\bar{\epsilon},0.5p\bar{\epsilon}])\\ &+0.5h(1,2)+0.5h(1,3) \end{aligned}$$

For $x=0$, we have $-\bar{\epsilon}\log(1-p)-\bar{\epsilon}\bar{\epsilon}\log p$ and for $x=1$ we have $\bar{\epsilon}(1-\log p)$. When we choose the optimal parameter these equations are equal. The computations for the other MDP states are similar. □

B. The Ising channel

The Ising channel, introduced in [20], is given by:

$$Y_i=\begin{cases} X_i & \text{, w.p. } 0.5\\ X_{i-1} & \text{, w.p. } 0.5. \end{cases}$$

Its feedback capacity for the binary-input case is:

Theorem 10 (Ising capacity, [21])**.** *The feedback capacity of the Ising channel is:*

$$C_{\text{ISING}}=\max_{0\leq p\leq 1}\frac{H_2(p)}{2+p}. \tag{9}$$

In the following, we provide an alternative and simple converse proof for the Ising channel with the binary alphabet.

Theorem 11 (Upper bound for the Ising channel)**.** *The feedback capacity of the Ising channel satisfies*

$$C_{ISING}\leq -0.5\log p, \tag{10}$$

where p is the unique solution for $(1-x)^4=x^3$.

It can be shown that the feedback capacity in Theorem 10 is equal to the upper bound in Theorem 11.

Remark 2 (Ising channel with larger alphabet)**.** *In a recent advancement to the RL algorithms in [7], the capacity of the Ising channel with alphabet size $|\mathcal{X}|\leq 8$ has been established. The duality bound, derived in the current paper, used to prove the converse, and its computation is still very simple despite the large alphabets.*

Solution for the MDP: The Q-graph consists of four nodes, and its evolution function is given by the vectors representation $\underline{\Phi}(y=0,\underline{q})=[2,1,1,1]$ and $\underline{\Phi}(y=1,\underline{q})=[3,3,4,3]$.

For some p, define the test distribution

$$T(y=0|\underline{q})=\left[\frac{1+p}{2},\frac{2p}{1+p},\frac{1-p}{2},\frac{1-p}{1+p}\right].$$

Define the constant:

$$\rho^*=\frac{D(1||\frac{2p}{1+p})+D(1||\frac{1+p}{2})}{2}$$

$$= -0.5 \log p, \tag{11}$$

where p solves $(1-x)^4 = x^3$.

Define the value function:

$$\begin{aligned} h(0,1) &= h(1,3) = \rho^* \\ h(0,2) &= h(1,4) = \log(1+p) + 2\rho^* - 1 \\ h(0,3) &= h(1,1) = 1 - \log(1-p) \\ h(0,4) &= h(1,2) = \log \frac{1+p}{1-p}. \end{aligned} \tag{12}$$

The following technical lemma concludes the proof of Theorem 11.

Lemma 3. *The constant ρ^* in* (11) *and the value function in* (12) *solve the Bellman equation.*

The proof of the Lemma 3 is omitted due to space limitations and follows from explicit computations as in Lemma 2.

VI. Proofs

In this section, we provide the proofs of Theorems 3 and 4.

Proof of Theorem 3. Consider the following chain of inequalities

$$\begin{aligned} & I(X^n \to Y^n) \\ &= \sum_{x^n, y^n} P(x^n||y^{n-1}) P(y^n||x^n) \log \frac{P(y^n||x^n)}{P(y^n)} \\ &\le \sum_{x^n, y^n} P(x^n||y^{n-1}) P(y^n||x^n) \log \frac{P(y^n||x^n)}{T(y^n)} \\ &\le \max D(P_{Y^n||X^n=x^n} || T(Y^n)), \end{aligned} \tag{13}$$

where the maximum is taken over sequences of deterministic functions $f_i : \mathcal{X}^{i-1} \times \mathcal{Y}^{i-1} \to \mathcal{X}$ for $i = 1, \dots, n$. □

Proof of Theorem 4. Here, we simplify (13) when the channel is unifilar and the test distribution is defined on a Q-graph. Consider the following chain of inequalities,

$$\begin{aligned} & D(P_{Y^n||X^n=x^n} || T(Y^n)) \\ &= \sum_{y^n} P(y^n||x^n) \log \frac{P(y^n||x^n)}{T(y^n)} \\ &= \sum_{y^n} P(y^n||x^n) \log \prod_{i=1}^{n} \left(\frac{p(y_i|y^{i-1}, x^i)}{T(y_i|y^{i-1})} \right) \\ &\overset{(a)}{=} \sum_{i} \sum_{y^i} P(y^i||x^i) \log \frac{P_{Y|X,S}(y_i|x_i, s_{i-1})}{T_{Y|Q}(y_i|q_{i-1})} \\ &= \sum_{i} \sum_{y^{i-1}} P(y^{i-1}||x^{i-1}) D(P_{Y|X,S}(\cdot|x_i, s_{i-1}) || T(\cdot|q_{i-1})) \\ &\le \max_{f(x^n||y^{n-1})} \sum_{i} \mathbb{E}[D(P_{Y|X,S}(\cdot|x_i, S_{i-1}) || T(\cdot|Q_{i-1})] \end{aligned}$$

where (a) follows by summing over y_{i+1}^n and the fact that q_{i-1} is a function of y^{i-1}. The limits existence follows from the super-additivity of the sequence (proof is omitted):

$$\underline{C}_n \triangleq \max_{f(x^n||y^{n-1})} \min_{s_0, q_0} \sum_{i} \mathbb{E}[D(P_{Y|X,S}(\cdot|x_i, S_{i-1}) || T(\cdot|Q_{i-1})].$$

□

Proof of Theorem 6. The exchange of the limit and the maximization follows from standard arguments (e.g., [19, Theorem 3]) that are based on the super-additivity property of $\underline{C}_n$. □

References

[1] H. H. Permuter, P. Cuff, B. V. Roy, and T. Weissman, "Capacity of the Trapdoor channel with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3150–3165, Jul. 2009.

[2] S. Yang, A. Kavčić, and S. Tatikonda, "Feedback capacity of finite-state machine channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 799–810, Mar. 2005.

[3] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.

[4] O. Sabag, H. H. Permuter, and H. D. Pfister, "A single-letter upper bound on the feedback capacity of unifilar finite-state channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1392–1409, Mar. 2017.

[5] O. Sabag, B. Huleihel, and H. H. Permuter, "Graph-based encoders and their achievable rates for channels with feedback," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 1121–1125.

[6] ——, "Graph-based encoders and their performance for finite-state channels with feedback," accepted to *IEEE Trans. Comm.*, 2020. Available at arxiv.org/abs/1907.08063.

[7] Z. Aharoni, O. Sabag, and H. Permuter, "Computing the feedback capacity of finite state channels using reinforcement," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019.

[8] A. Mohanan and A. Thangaraj, "Dual capacity upper bounds for binary-input single-tap isi channels," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[9] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. on Inf. Theory*, vol. 49, no. 10, pp. 2426–2467, Oct. 2003.

[10] A. Thangaraj, G. Kramer, and G. Bcherer, "Capacity bounds for discrete-time, amplitude-constrained, additive white gaussian noise channels," *IEEE Trans. on Inf. Theory*, vol. 63, no. 7, pp. 4172–4182, July 2017.

[11] G. Kramer, "Directed information for channels with feedback," Ph.D. Dissertation, Swiss Federal Institute of Technology (ETH) Zurich, 1998.

[12] I. Csiszr and J. Krner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[13] I. Csiszár, "Informationstheoretische Konvergenzbegriffe im Raum der Wahrscheinlichkeitsverteilungen," *Publ. Math. Inst. Hungar. Acad.*, vol. 7, pp. 137–158, 1962.

[14] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. Marcus, "Discrete time controlled Markov processes with average cost criterion - a survey," *SIAM Journal of Control and Optimization*, vol. 31, no. 2, pp. 282–344, 1993.

[15] D. P. Bertsekas, *Dynamic Programming and Optimal Control: Vols 1 and 2*. Belmont, MA.: Athena Scientific, 2000.

[16] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley and Sons, 1994.

[17] H. Pfister and P. Siegel, "Joint iterative decoding of LDPC codes for channels with memory and erasure noise," *Selected Areas in Communications, IEEE Journal on*, vol. 26, no. 2, pp. 320–337, February 2008.

[18] H. D. Pfister, "On the capacity of finite state channels and the analysis of convolutional accumulate-m codes," Ph.D. Dissertation, University of California, San Diego, La Jolla, 2003.

[19] O. Sabag, H. Permuter, and N. Kashyap, "The feedback capacity of the binary erasure channel with a no-consecutive-ones input constraint," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 8–22, Jan 2016.

[20] T. Berger and F. Bonomi, "Capacity and zero-error capacity of Ising channels," *IEEE Trans. Inf. Theory*, vol. 36, pp. 173–180, 1990.

[21] O. Elishco and H. Permuter, "Capacity and coding for the Ising channel with feedback," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5138–5149, Sep. 2014.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

On the Information Bottleneck Problems: An Information Theoretic Perspective

Abdellatif Zaidi[†] [‡] Shlomo Shamai[†]

[‡] Université Paris-Est, Champs-sur-Marne, 77454, France
[†] Paris Research Center, Huawei Technologies, Boulogne-Billancourt, 92100, France
[†] Technion Institute of Technology, Technion City, Haifa 32000, Israel
{abdellatif.zaidi@u-pem.fr, sshlomo@ee.technion.ac.il}

***Abstract*—This paper focuses on variants of the bottleneck problem taking an information theoretic perspective. The intimate connections of this setting to: remote source-coding, information combining, common reconstruction, the Wyner-Ahlswede-Korner problem, the efficiency of investment information, CEO source coding under logarithmic-loss distortion measure and others are highlighted. We discuss the distributed information bottleneck problem with emphasis on the Gaussian model. For this model, the optimal tradeoffs between relevance (i.e., information) and complexity (i.e., rates) in the discrete and vector Gaussian frameworks is determined.**

I. Statistical Inference

Let a measurable variable $X \in \mathcal{X}$ and a target variable $Y \in \mathcal{Y}$ with unknown joint distribution $P_{X,Y}$ be given. In the classic problem of statistical learning, one wishes to infer an accurate predictor of the target variable $Y \in \mathcal{Y}$ based on observed realizations of $X \in \mathcal{X}$. That is, for a given class $\mathcal{F}$ of admissible predictors $\psi : \mathcal{X} \to \hat{\mathcal{Y}}$ and a loss function $\ell : \mathcal{Y} \to \hat{\mathcal{Y}}$ that measures discrepancies between true values and their estimated fits, one aims at finding the mapping $\psi \in \mathcal{F}$ that minimizes the expected (population) risk

$$\mathcal{C}_{P_{X,Y}}(\psi, \ell) = \mathbb{E}_{P_{X,Y}}[\ell(Y, \psi(X))]. \tag{1}$$

An abstract inference model is shown in Figure 1.

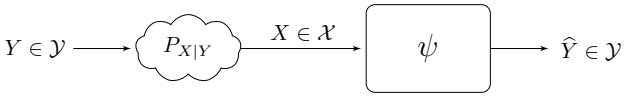


Fig. 1. An abstract inference model for learning.

The choice of a "good" loss function $\ell(\cdot)$ is often controversial in statistical learning theory. There is however numerical evidence that models that are trained to minimize the error's entropy often outperform ones that are trained using other criteria such as mean-square error (MSE) and higher-order statistics [1], [2]. This corresponds to choosing the loss function given by the logarithmic loss, which is defined as

$$\ell_{\log}(y, \hat{y}) := \log \frac{1}{\hat{y}(y)} \tag{2}$$

for $y \in \mathcal{Y}$ and $\hat{y} \in \mathcal{P}(\mathcal{Y})$ designates here a probability distribution on $\mathcal{Y}$ and $\hat{y}(y)$ is the value of that distribution evaluated at the outcome $y \in \mathcal{Y}$. Although a complete and rigorous justification of the usage of the logarithmic loss as distortion measure in learning is still awaited, recently a partial explanation appeared in [3] where Painsky and Wornell show that, for binary classification problems, by minimizing the logarithmic-loss one actually minimizes an upper bound to any choice of loss function that is smooth, proper (i.e., unbiased and Fisher consistent) and convex. Along the same line of work, the authors of [4] show that under some natural data processing property Shannon's mutual information uniquely quantifies the reduction of prediction risk due to side information. Perhaps, this justifies partially why the logarithmic-loss fidelity measure is widely used in learning theory and has already been adopted in many algorithms in practice such as the *infomax* criterion [5]. The logarithmic loss measure also plays a central role in the theory of prediction [6, Ch. 09], where it is often referred to as the *self-information* loss function, as well as in Bayesian modeling [7] where priors are usually designed so as to maximize the mutual information between the parameter to be estimated and the observations.

Let for every $x \in \mathcal{X}$, $\psi(x) = Q(\cdot|x) \in \mathcal{P}(\mathcal{Y})$. It is easy to see that

$$\mathbb{E}_{P_{X,Y}}[\ell_{\log}(Y, Q)] = \sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} P_{X,Y}(x, y) \log\Big(\frac{1}{Q(y|x)}\Big) \tag{3a}$$

$$= H(Y|X) + D\big(P_{Y|X} \| Q\big) \tag{3b}$$

$$\geq H(Y|X) \tag{3c}$$

with equality iff $\psi(X) = P_{Y|X}$. That is,

$$\min_{\psi} \mathcal{C}_{P_{X,Y}}(\psi, \ell_{\log}) = H(Y|X). \tag{4}$$

If the joint distribution $P_{X,Y}$ is unknown, which is most often the case in practice, the population risk as given by (1) cannot be computed directly; and, in the standard approach, one usually resorts to choosing the predictor with minimal risk on a training dataset consisting of n labeled samples $\{(x_i, y_i)\}_{i=1}^n$ that are drawn independently from the unknown joint distribution $P_{X,Y}$. In this case, it is important to restrict the set $\mathcal{F}$ of admissible predictors to a low-complexity class to prevent overfitting. One way to reduce the model's complexity is by restricting the range of the prediction function as shown in Figure 2. Here, the stochastic mapping $\phi : \mathcal{X} \longrightarrow \mathcal{U}$ is a compressor with

$$\|\phi\| \leq R \tag{5}$$

for some prescribed 'input-complexity' value R.

Let $U = \phi(X)$. The expected logarithmic loss is now given by

$$\mathcal{C}_{P_{X,Y}}(\phi, \psi; \ell_{\log}) = \mathbb{E}_{P_{X,Y}}[\ell_{\log}(Y, \psi(U))] \tag{6}$$

and takes its minimum value with the choice $\psi(U) = P_{Y|U}$,

$$\min_{\psi} \mathcal{C}_{P_{X,Y}}(\phi, \psi; \ell_{\log}) = H(Y|U) \tag{7}$$

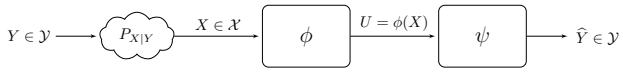


Fig. 2. Inference problem with constrained model's complexity.

where the choice of U is subjected to the input constraint (5). Noting that the right-hand-side (RHS) of (7) is larger for small values of R, it is clear that a good predictor ϕ should strike a right balance between reducing the model's complexity and reducing the error's entropy, or, equivalently, maximizing the mutual information $I(U;Y)$ about the target variable Y.

A. Remote Source Coding under Logarithmic Loss

The aforementioned inference problem is a one-shot coding problem, in the sense that the prediction and estimation operations are performed letter-wise. Consider now the (asymptotic) remote source coding problem shown in Figure 3 in which the coding operations are performed over blocks of size n, with n assumed to be large. Here, Y designates a memoryless remote source; and X a noisy version of it that is observed at the encoder. The range of the encoder map is allowed to grow with the size of the input sequence as

$$\|\phi^{(n)}\| \leq nR. \tag{8}$$

That is, the encoder uses at most R bits per sample to describe its observation to a decoder which is interested in reconstructing the remote source Y^n to within an average distortion level D, i.e.,

$$\mathbb{E}[\ell_{\log}^{(n)}(\mathbf{y},\hat{\mathbf{y}})] \leq D \tag{9}$$

where the incurred distortion between two vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$ is given by

$$\ell_{\log}^{(n)}(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\log}(y_i,\hat{y}_i) \tag{10}$$

with the per-letter distortion defined as specified by (2).

Fig. 3. A remote source coding problem.

The rate distortion region of this model is given by the union of all pairs (R,D) that satisfy [8], [9]

$$R \geq I(U;X) \tag{11a}$$
$$D \geq H(Y|U) \tag{11b}$$

where the union is over all auxiliary random variables U that satisfy that $U \multimap X \multimap Y$ forms a Markov Chain in this order. Invoking the support lemma [10, p. 310], it is easy to see that this region is not altered if one restricts U to satisfy $|\mathcal{U}| \leq |\mathcal{X}| + 1$. Also, using the substitution $\Delta := H(Y) - D$, the region can be written equivalently as the union of all pairs $(R, H(Y) - \Delta)$ that satisfy

$$R \geq I(U;X) \tag{12a}$$
$$\Delta \leq I(U;Y) \tag{12b}$$

where the union is over all U's that satisfy $U \multimap X \multimap Y$, with $|\mathcal{U}| \leq |\mathcal{X}| + 1$.

B. Information Bottleneck

The Information Bottleneck (IB) method has been introduced by Tishby *et al.* in [11] as a method for extracting the information that some variable $X \in \mathcal{X}$ provides about another one $Y \in \mathcal{Y}$ that is of interest. Specifically, IB finds a representation U that is maximally informative about Y, i.e., large mutual information $I(U;Y)$, while being minimally informative about X, i.e., small mutual information $I(U;X)$ [1]. The auxiliary random variable U satisfies that $U \multimap X \multimap Y$ is a Markov chain in this order; and is chosen so a to strike a suitable balance between the degree of *relevance* of the representation as measured by the mutual information $I(U;Y)$ and its degree of *complexity* as measured by the mutual information $I(U;X)$. For example, U can be determined so as to minimize the IB-Lagrangian

$$\mathcal{L} : I(U;X) - \beta I(U;Y) \tag{13}$$

over all mappings that satisfy $U \multimap X \multimap Y$. The tradeoff parameter β is a positive Lagrange multiplier associated with the constraint on $I(U;Y)$. The solution of this constrained optimization problem is determined by the following self-consistent equations, for all $(u,x,y) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$,

$$P_{U|X}(u|x) = \frac{P_U(u)}{Z(\beta,x)} \exp\Big(-\beta D\Big(P_{Y|X}(\cdot|x)\|P_{Y|U}(\cdot|u)\Big)\Big) \tag{14a}$$
$$P_U(u) = \sum_{x\in\mathcal{X}} P_X(x) P_{U|X}(u|x) \tag{14b}$$
$$P_{Y|U}(y|u) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_{X|U}(x|u) \tag{14c}$$

where $P_{X|U}(x|u) = P_{U|X}(u|x)P_X(x)/P_U(u)$ and $Z(\beta,x)$ is a normalization term. It is shown in [11] that alternating iterations of these equations converges to a solution of the problem for any initial $P_{U|X}$. However, by opposition to the standard Blahut-Arimoto algorithm [13], [14] which is classically used in the computation of rate-distortion functions of discrete memoryless sources for which convergence to the optimal solution is guaranteed, convergence here may be to a local optimum only. If $\beta = 0$ the optimization is non-constrained and one can set $U = \varnothing$, which yields minimal relevance and complexity levels. Increasing the value of β steers towards more accurate and more complex representations, until $U = X$ in the limit of very large (infinite) values of β for which the relevance reaches its maximal value $I(X;Y)$.

C. Variational Inference

Recall the IB goal of finding a representation U of X that is maximally informative about Y while being concise enough (i.e., bounded $I(U;X)$). This corresponds to the Lagrangian formulation

$$\mathcal{L} : \max I(U;Y) - \beta I(U;X) \tag{15}$$

where the maximization is over all stochastic mappings $P_{U|X}$ such that $U \multimap X \multimap Y$ and $|\mathcal{U}| \leq |\mathcal{X}| + 1$. The main drawback of the IB principle is that in the exception of small-sized discrete (X,Y) for which iterating (14) converges to an (at least local) solution and jointly Gaussian (X,Y) for which an explicit analytic solution was found, solving (15) is generally computationally costly

[1] As such, the usage of Shannon's mutual information seems to be motivated by the intuition that such a measure provides a natural quantitative approach to the questions of meaning, relevance and common-information, rather than the solution of a well-posed information-theoretic problem – a connection with source coding under logarithmic loss measure appeared later on in [12].

especially for high-dimensional data since it requires computation of mutual information terms. Another important barrier in solving (15) directly is that IB necessitates knowledge of the joint distribution $P_{X,Y}$. A major step ahead, which widened up the range of applications of IB inference for various learning problems, appeared in [15] where the authors use variational inference to derive a lower bound on (15) and show that its optimization can be done through the classic and widely used stochastic gradient descendent (SGD). This has allowed to use deep neural networks to parametrize the involved distributions (including the test channel $P_{U|X}$); and, thus, to handle high-dimensional, possibly continuous, data.

II. Connections

A. Common Reconstruction

Consider the problem of source coding with side information at the decoder, i.e., the well known Wyner-Ziv setting [16], with the distortion measured under logarithmic-loss. Specifically, a memoryless source X is to be conveyed lossily to a decoder that observes a statistically correlated side information Y. The encoder uses R bits per sample to describe its observation to the decoder which wants to reconstruct an estimate of X to within an average distortion level D, where the distortion is evaluated under the log-loss distortion measure. The rate distortion region of this problem is given by the set of all pairs (R, D) that satisfy

$$R + D \geq H(X|Y). \tag{16}$$

The optimal coding scheme utilizes standard Wyner-Ziv compression at the encoder and the decoder map $\psi : \mathcal{U} \times \mathcal{Y} \to \hat{\mathcal{X}}$ is given by

$$\psi(U, Y) = \Pr[X = x|U, Y] \tag{17}$$

for which it is easy to see that

$$\mathbb{E}[\ell_{\log}(X, \psi(U, Y))] = H(X|U, Y). \tag{18}$$

Now, assume that we constrain the coding in a manner that the encoder be able to produce an exact copy of the compressed source constructed by the decoder. This requirement, termed *common reconstruction* constraint (CR), was introduced and studied by Steinberg in [17] for various source coding models, including the Wyner-Ziv setup, in the context of a "general distortion measure. For the Wyner-Ziv problem under log-loss measure that is considered in this section, such a CR constraint causes some rate loss because the reproduction rule (17) is no longer possible. In fact, it is not difficult to see that under the CR constraint the above region reduces to the set of pairs (R, D) that satisfy

$$R \leq I(U; X|Y) \tag{19a}$$
$$D \geq H(X|U) \tag{19b}$$

for some auxiliary random variable for which $U \multimap X \multimap Y$ holds. Observe that (19b) is equivalent to $I(U; X) \geq H(X) - D$ and that, for a given prescribed fidelity level D, the minimum rate is obtained for a description U that achieves the inequality (19b) with equality, i.e.,

$$R(D) = \min_{P_{U|X} \,:\, I(U;X)=H(X)-D} I(U; X|Y). \tag{20}$$

Because $U \multimap X \multimap Y$, we have

$$I(U; Y) = I(U; X) - I(U; X|Y). \tag{21}$$

Under the constraint $I(U; X) = H(X) - D$ it is easy to see that minimizing $I(U; X|Y)$ amounts to maximizing $I(U; Y)$, an aspect which bridges the problem at hand with the IB problem.
In the above, the side information Y is used for binning but not for the estimation at the decoder. If the encoder ignores whether Y is present or not at the decoder side, the benefit of binning is reduced – see the Heegard-Berger model with common reconstruction studied in [18], [19].

B. Information Combining

Consider again the IB problem. Say one wishes the find the representation U that maximizes the relevance $I(U; Y)$ for a given prescribed complexity level, e.g., $I(U; X) = R$. For this setup,

$$I(X; U, Y) = I(U; X) + I(Y; X) - I(U; Y) \tag{22}$$
$$= R + I(Y; X) - I(U; Y) \tag{23}$$

where the first equality holds since $U \multimap X \multimap Y$ is a Markov chain. Maximizing $I(U; Y)$ is then equivalent to minimizing $I(X; U, Y)$. This is reminiscent of the problem of *information combining* [20], where X can be interpreted as a source information that is conveyed through two channels: the channel $P_{Y|X}$ and the channel $P_{U|X}$. The outputs of these two channels are conditionally independent given X; and they should be processed in a manner such that, when combined, they preserve as much information as possible about X.

C. Wyner-Ahlswede-Korner Problem

Here, the two memoryless sources X and Y are encoded separately at rates R_X and R_Y respectively. A decoder gets the two compressed streams and aims at recovering Y losslessly. This problem was studied and solved separately by Wyner [21] and Ahlswede and Körner [22]. For given $R_X = R$, the minimum rate R_Y that is needed to recover Y losslessly is

$$R_Y^\star(R) = \min_{P_{U|X} \,:\, I(U;X) \leq R} H(Y|U). \tag{24}$$

So, we get

$$\max_{P_{U|X} \,:\, I(U;X) \leq R} I(U; Y) = H(Y) - R_Y^\star(R).$$

D. The Privacy Funnel

Consider again the setting of Figure 3; and let us assume that the pair (Y, X) models data that a user possesses and which has the following properties: the data Y is some sensitive (private) data that is not meant to be revealed at all, or else not beyond some level Δ; and the data X is non-private and is meant to be shared with another user (analyst). Because X and Y are correlated, sharing the non-private data X with the analyst possibly reveals information about Y. For this reason, there is a tradeoff between the amount of information that the user shares about X and the information that he keeps private about Y. The data X is passed through a randomized mapping ϕ whose purpose is to make $U = \phi(X)$ maximally informative about X while being minimally informative about Y.
The analyst performs an inference attack on the private data Y based on the disclosed information U. Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \longrightarrow \bar{\mathbb{R}}$ be an arbitrary loss function with reconstruction alphabet $\hat{\mathcal{Y}}$ that measures the cost of inferring Y after observing U. Given $(X, Y) \sim P_{X,Y}$ and under the given loss function ℓ, it is natural to quantify the difference between the prediction losses in predicting $Y \in \mathcal{Y}$ prior and after observing $U = \phi(X)$. Let

$$C(\ell, P) = \inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_P[\ell(Y, \hat{y})] - \inf_{\hat{Y}(\phi(X))} \mathbb{E}_P[\ell(Y, \hat{Y})] \tag{25}$$

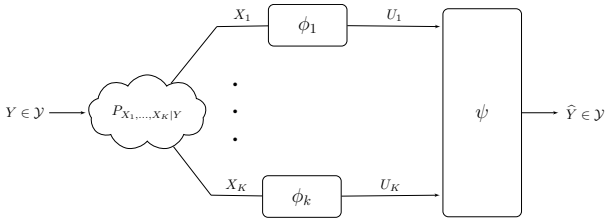


Fig. 4. A model for distributed, e.g., multi-view, learning.

where $\hat{y} \in \hat{\mathcal{Y}}$ is deterministic and $\hat{Y}(\phi(X))$ is any measurable function of $U = \phi(X)$. The quantity $C(\ell, P)$ quantifies the reduction in the prediction loss under the loss function ℓ that is due to observing $U = \phi(X)$, i.e., the inference cost gain. In [23] (see also [24]) it is shown that that under some mild conditions the inference cost gain $C(\ell, P)$ as defined by (25) is upper-bounded as

$$C(\ell, P) \leq 2\sqrt{2}L\sqrt{I(U;Y)} \tag{26}$$

where L is a constant. The inequality (26) holds irrespective to the choice of the loss function ℓ; and this justifies the usage of the logarithmic loss function as given by (2) in the context of finding a suitable tradeoff between utility and privacy, since

$$I(U;Y) = H(Y) - \inf_{\hat{Y}(U)} \mathbb{E}_P[\ell_{\log}(Y, \hat{Y})]. \tag{27}$$

Under the logarithmic loss function, the design of the mapping $U = \phi(X)$ should strike a right balance between the utility for inferring the non-private data X as measured by the mutual information $I(U;X)$ and the privacy metric about the private date Y as measured by the mutual information $I(U;Y)$.

E. Efficiency of Investment Information

Let Y model a stock market data and X some correlated information. In [25], Erkip and Cover investigated how the description of the correlated information X improves the investment in the stock market Y. Specifically, let $\Delta(C)$ denote the maximum increase in growth rate when X is described to the investor at rate C. Erkip and Cover found a single-letter characterization of the incremental growth rate $\Delta(C)$. When specialized to the horse race market, this problem is related to the aforementioned source coding with side information of Wyner [21] and Ahlswede-Körner [22]; and, so, also to the IB problem. The work [25] provides explicit analytic solutions for two horce race examples, jointly binary and jointly Gaussian horce races.

III. Distributed Learning

Consider now a generalization of the IB problem in which the prediction is to be performed in a distributed manner. The model is shown in Figure 4. Here, the prediction of the target variable $Y \in \mathcal{Y}$ is to be performed on the basis of samples of statistically correlated random variables $(X_1, \cdots, X_K)$ that are observed each at a distinct predictor. Throughout, we assume that the following Markov chain holds for all $k \in \mathcal{K} := \{1, \cdots, K\}$,

$$X_k \multimap Y \multimap X_{\mathcal{K}/k}. \tag{28}$$

The variable Y is a target variable and we seek to characterize how accurate it can be predicted from a measurable random vector $(X_1, \cdots, X_K)$ when the components of this vector are processed separately, each by a distinct encoder.

A. Optimal relevance-complexity tradeoff region

The distributed IB problem of Figure 4 is studied in [26], [27] from information-theoretic grounds. For both discrete memoryless (DM) and memoryless vector Gaussian models, the authors establish fundamental limits of learning in terms of optimal tradeoffs between relevance and complexity. The following theorem [26], [27] states the result for the case of discrete memoryless sources.

Theorem 1. *The relevance-complexity region $\mathcal{IR}_{\mathrm{DIB}}$ of the distributed learning problem is given by the union of all non-negative tuples $(\Delta, R_1, \ldots, R_K) \in \mathbb{R}_+^{K+1}$ that satisfy*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k - I(X_k; U_k|Y,T)] + I(Y; U_{\mathcal{S}^c}|T), \quad \forall \mathcal{S} \subseteq \mathcal{K} \tag{29}$$

for some joint distribution of the form $P_T P_Y \prod_{k=1}^K P_{X_k|Y} \prod_{k=1}^K P_{U_k|X_k,T}$.

B. A Variational Bound

Let us consider the problem of maximizing the relevance under a sum-complexity constraint. Let $R_{\mathrm{sum}} = \sum_{k=1}^K R_k$ and

$$\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}} := \Big\{(\Delta, R_{\mathrm{sum}}) \in \mathbb{R}_+^2 : \exists (R_1, \ldots, R_K) \in \mathbb{R}_+^K \text{ s.t. } \sum_{k=1}^K R_k = R_{\mathrm{sum}} \text{ and } (\Delta, R_1, \ldots, R_K) \in \mathcal{RI}_{\mathrm{DIB}}\Big\}. \tag{30}$$

It is easy to see that the region $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$ is composed of all the pairs $(\Delta, R_{\mathrm{sum}}) \in \mathbb{R}_+^2$ for which $\Delta \leq \Delta(R_{\mathrm{sum}}, P_{X_{\mathcal{K}},Y})$, with

$$\Delta(R_{\mathrm{sum}}, P_{X_{\mathcal{K}},Y}) = \max_{\mathbf{P}} \min \left\{ I(Y; U_{\mathcal{K}}), R_{\mathrm{sum}} - \sum_{k=1}^K I(X_k; U_k|Y) \right\}, \tag{31}$$

where the maximization is over joint distributions that factorize as $P_Y \prod_{k=1}^K P_{X_k|Y} \prod_{k=1}^K P_{U_k|X_k}$. The pairs $(\Delta, R_{\mathrm{sum}})$ that lie on the boundary of $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$ can be characterized as given in the following proposition [27, Section 7.3].

Proposition 1. *For every pair $(\Delta, R_{\mathrm{sum}}) \in \mathbb{R}_+^2$ that lies on the boundary of the region $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$ there exists a parameter $s \geq 0$ such that $(\Delta, R_{\mathrm{sum}}) = (\Delta_s, R_s)$, with*

$$\Delta_s = \frac{1}{(1+s)} \Big[(1+sK)H(Y) + sR_s + \max_{\mathbf{P}} \mathcal{L}_s(\mathbf{P}) \Big], \tag{32}$$

$$R_s = I(Y; U_{\mathcal{K}}^*) + \sum_{k=1}^K [I(X_k; U_k^*) - I(Y; U_k^*)], \tag{33}$$

where $\mathbf{P}^*$ is the set of conditional pmfs $\mathbf{P} = \{P_{U_1|X_1}, \cdots, P_{U_K|X_K}\}$ that maximize the cost function

$$\mathcal{L}_s(\mathbf{P}) := -H(Y|U_{\mathcal{K}}) - s \sum_{k=1}^K [H(Y|U_k) + I(X_k; U_k)]. \tag{34}$$

The optimization of (34) generally requires to compute marginal distributions that involve the descriptions $U_1, \cdots, U_K$, which might not be possible in practice. In what follows, we derive a variational lower bound on $\mathcal{L}_s(\mathbf{P})$ on the DIB cost function in terms of families of stochastic mappings $Q_{Y|U_1,\ldots,U_K}$ (a decoder), $\{Q_{Y|U_k}\}_{k=1}^K$ and priors $\{Q_{U_k}\}_{k=1}^K$. For the simplicity of the notation, we let

$$\mathbf{Q} := \{Q_{Y|U_1,\ldots,U_K}, Q_{Y|U_1}, \ldots, Q_{Y|U_K}, Q_{U_1}, \ldots, Q_{U_K}\}. \tag{35}$$

Let

$$\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) := \underbrace{\mathbb{E}[\log Q_{Y|U_{\mathcal{K}}}(Y|U_{\mathcal{K}})]}_{\text{av. logarithmic-loss}}$$

$$+ s \underbrace{\sum_{k=1}^{K} \Big(\mathbb{E}[\log Q_{Y|U_k}(Y|U_k)] - D_{\mathrm{KL}}(P_{U_k|X_k} \| Q_{U_k}) \Big)}_{\text{regularizer}}. \quad (36)$$

Lemma 1. *([27, Section 7.4]) For fixed* $\mathbf{P}$*, we have*

$$\mathcal{L}_s(\mathbf{P}) \geq \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}), \qquad \text{for all pmfs } \mathbf{Q}. \quad (37)$$

In addition, there exists a unique $\mathbf{Q}$ *that achieves the maximum* $\max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) = \mathcal{L}_s(\mathbf{P})$*, and is given by,* $\forall k \in \mathcal{K}$*,*

$$Q^*_{U_k} = P_{U_k} \quad (38a)$$
$$Q^*_{Y|U_k} = P_{Y|U_k} \quad (38b)$$
$$Q^*_{Y|U_1,\ldots,U_k} = P_{Y|U_1,\ldots,U_K}, \quad (38c)$$

where the marginals P_{U_k} *and the conditional marginals* $P_{Y|U_k}$ *and* $P_{Y|U_1,\ldots,U_K}$ *are computed from* $\mathbf{P}$*.*

C. Vector Gaussian Model

In this section, we show that for the jointly vector Gaussian data model it is enough to restrict to Gaussian auxiliaries $(\mathbf{U}_1, \cdots, \mathbf{U}_K)$ in order to exhaust the entire relevance-complexity region. Also, we provide an explicit analytical expression of this region. Let $(\mathbf{X}_1, \ldots, \mathbf{X}_K, \mathbf{Y})$ be a jointly vector Gaussian vector that satisfies the Markov chain (28). Without loss of generality, let the target variable be a complex-valued, zero-mean multivariate Gaussian $\mathbf{Y} \in \mathbb{C}^{n_y}$ with covariance matrix $\mathbf{\Sigma_y}$, i.e., $\mathbf{Y} \sim \mathcal{CN}(\mathbf{y}; \mathbf{0}, \mathbf{\Sigma_y})$, and $\mathbf{X}_k \in \mathbb{C}^{n_k}$ given by

$$\mathbf{X}_k = \mathbf{H}_k \mathbf{Y} + \mathbf{N}_k, \quad (39)$$

where $\mathbf{H}_k \in \mathbb{C}^{n_k \times n_y}$ models the linear model connecting $\mathbf{Y}$ to the observation at encoder k, and $\mathbf{N}_k \in \mathbb{C}^{n_k}$ is the noise vector at encoder k, assumed to be Gaussian with zero-mean and covariance matrix $\mathbf{\Sigma}_k$, and independent from all other noises and $\mathbf{Y}$.

The following theorem [27, Section 7.5] characterizes the relevance-complexity region of the model (39), which we denote hereafter as $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{G}}$. The theorem also shows that in order to exhaust this region it is enough to restrict to no time sharing, i.e., $T = \varnothing$ and multivariate Gaussian test channels

$$\mathbf{U}_k = \mathbf{A}_k \mathbf{X}_k + \mathbf{Z}_k \sim \mathcal{CN}(\mathbf{u}_k; \mathbf{A}_k \mathbf{X}_k, \mathbf{\Sigma}_{z,k}), \quad (40)$$

where $\mathbf{A}_k \in \mathbb{C}^{n_k \times n_k}$ projects the observation $\mathbf{X}_k$ and $\mathbf{Z}_k$ is a zero-mean Gaussian noise with covariance $\mathbf{\Sigma}_{z,k}$.

Theorem 2. *For the model* (39) *the region* $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{G}}$ *is given by the union of all tuples* $(\Delta, R_1, \ldots, R_L)$ *that satisfy* $\forall \mathcal{S} \subseteq \mathcal{K}$

$$\Delta \leq \sum_{k \in \mathcal{S}} \Big(R_k + \log \Big| \mathbf{I} - \mathbf{\Sigma}_k^{1/2} \mathbf{\Omega}_k \mathbf{\Sigma}_k^{1/2} \Big| \Big) + \log \Big| \mathbf{I} + \sum_{k \in \mathcal{S}^c} \mathbf{\Sigma}_{\mathbf{y}}^{1/2} \mathbf{H}_k^\dagger \mathbf{\Omega}_k \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{y}}^{1/2} \Big|$$

for some matrices $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$*.*

Acknowledgment: The work of S. Shamai has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630.

References

[1] D. Erdogmus, "Information theoretic learning: Renyi's entropy and its applications to adaptive system training," Ph.D. dissertation, University of Florida Gainesville, Florida, 2002.

[2] J. C. Principe, N. R. Euliano, and W. C. Lefebvre, *Neural and adaptive systems: fundamentals through simulations*. Wiley New York, 2000, vol. 672.

[3] A. Painsky and G. W. Wornell, "On the universality of the logistic loss function," *arXiv preprint arXiv:1805.03804*, 2018.

[4] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5357–5365, 2015.

[5] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[6] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning and games*. New York, USA: Cambridge, Univ. Press, 2006.

[7] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[8] R.-L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IRE Trans. on Info. Theory*, vol. 85, pp. 293–304, 1962.

[9] H.-S. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. on Info. Theory*, vol. IT-26, pp. 518–521, Sep. 1980.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. London, U. K.: Academic Press, 1981.

[11] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999, pp. 368–377.

[12] P. Harremoes and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proc. IEEE Int. Symp. Information Theory*, Jun. 2007, pp. 566–570.

[13] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul 1972.

[14] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 12 – 20, Jan. 1972.

[15] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational information bottleneck," in *ICLR*, 2017. [Online]. Available: https://arxiv.org/abs/1612.00410

[16] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, pp. 1–10, Jan. 1976.

[17] Y. Steinberg, "Coding and common reconstruction," *IEEE Trans. Inf. Theory*, vol. IT-11, pp. 4995–5010, Nov. 2009.

[18] M. Benammar and A. Zaidi, "Rate-distortion of a heegard-berger problem with common reconstruction constraint," in *Proc. of International Zurich Seminar on Information and Communication*. IEEE, Mar. 2016.

[19] ——, "Rate-distortion function for a heegard-berger problem with two sources and degraded reconstruction sets," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5080–5092, 2016.

[20] I. Sutskover, S. Shamai, and J. Ziv, "Extremes of information combining," *IEEE Trans. Inform. Theory*, vol. 51, no. 04, pp. 1313–1325, 2005.

[21] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-21, pp. 294–300, May 1975.

[22] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, November 1975.

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

[24] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *IEEE Trans. Info. Theory*, vol. 65, no. 03, pp. 1512–1534, Mar. 2019.

[25] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Trans. Info. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[26] I. E. Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources," in *Proc. of Int. Zurich Seminar on Information and Communication, IZS*, Zurich, Switzerland, 2018.

[27] ——, "Distributed variational representation learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence. To appear. Available at https://arxiv.org/abs/1807.04193*, 2018.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Communication Subject to State Obfuscation

Ligong Wang* and Gregory W. Wornell†
*ETIS—Université Paris Seine, Université de Cergy-Pontoise, ENSEA, CNRS, Cergy-Pontoise, France
†Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

***Abstract*—We consider communication over a state-dependent discrete memoryless channel subject to a constraint that requires that the output sequence be nearly independent of the state. We consider three scenarios for the transmitter: where it knows the state, where it does not know the state and can use a stochastic encoder, and where it does not know the state and must use a deterministic encoder. For the state, we assume it to be either independent and identically distributed across channel uses or randomly generated but constant over all channel uses. We present single-letter capacity formulas for all except one combination of the above scenarios, and also solve some illustrative examples.**

I. Introduction

State-dependent channels have been extensively studied in Information Theory [1]–[3]. The current work considers communication over a state-dependent channel, with an additional requirement that the channel state should remain unknown to the receiver. A potential application for such a model is a scenario where the transmitter wishes to conceal its physical location: its location may affect the statistics of the channel to the receiver, hence can be modeled as a channel state.

The problem we study is closely related to "state masking" and, to a lesser extent, "state amplification" [4]–[8]. Consider a state-dependent discrete memoryless channel (DMC) where, given input $X = x$ and state $S = s$, the probability for the output Y to equal y is given by $W(y|x, s)$. Assume that the state is independent and identically distributed (IID) across channel uses according to a known distribution. The state-masking constraint considered in [4] is

$$\lim_{n\to\infty} \frac{1}{n} I(S^n; Y^n) \leq E \tag{1}$$

for some parameter E, where n denotes the number of times the channel is used. When channel-state information (CSI) is available noncausally to the transmitter (meaning the transmitter knows the realization of S^n before sending any input to the channel), a communication rate R is achievable under the above constraint if, and only if [4, Theorem 2]

$$R \leq I(U; Y) - I(U; S) \tag{2}$$

for some auxiliary random variable U such that $U \multimap (X, S) \multimap Y$ form a Markov chain, and that

$$I(S; U, Y) \leq E. \tag{3}$$

Note that (2) is the Gel'fand-Pinsker rate expression [2], while the condition (3) concerns $I(S; U, Y)$ and not $I(S; Y)$.

In the current paper we are interested in problems where the states must be almost completely concealed from the receiver, namely, where the limit in (1) must equal zero. Our result when CSI is available to the transmitter then follows almost immediately from [4]. We also consider situations where CSI is not available and derive similar capacity formulas. Interestingly, capacity differs between the cases where the transmitter must use a deterministic encoder and where it may use a stochastic encoder (that is not known to the receiver). Furthermore, keeping in mind that the state may be used to model the transmitter's physical location, we study models where the state remains constant during the entire transmission, instead of being IID. When CSI is available to the transmitter, or when CSI is not available and the transmitter must use a deterministic encoder, the capacity turns out to be the same as in the IID-state case. When CSI is not available and transmitter may use a stochastic encoder, however, capacity is different.

We consider IID states in Section II and constant states in Section III, and then conclude with some remarks.

II. IID States

Consider a DMC with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ that is affected by a random state S which takes value in the set $\mathcal{S}$. The sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{S}$ are all assumed to be finite. The channel law is, given the input $x \in \mathcal{X}$ and state $s \in \mathcal{S}$, the probability of the output being $y \in \mathcal{Y}$ is $W(y|x, s)$.

In this section, we assume that the states are drawn IID across channel uses according to a probability mass function P_S.

The message to be communicated is drawn from the set $\{1, \ldots, \lfloor 2^{nR} \rfloor\}$, where n denotes the total number of channel uses, and R the rate of communication in bits per channel use. The message is fed to an encoder, which in turn produces the channel input sequence x^n. We consider both cases where the state realizations are known and unknown to the transmitter, respectively. When the states are unknown to the transmitter, we further distinguish between deterministic and stochastic encoders; details are provided below. In all cases, the receiver tries to guess the message based on its observations y^n.

The *state-obfuscation* constraint we impose is

$$\lim_{n\to\infty} \frac{1}{n} I(S^n; Y^n) = 0, \tag{4}$$

where the mutual information is computed for the joint distribution induced by the encoder and a uniformly drawn message. As will become clear via our achievability proofs, all results in this section will continue to hold when we replace (4) by the stronger condition

$$I(S^n; Y^n) = 0 \quad \text{for every } n. \tag{5}$$

In each of the following cases, we define capacity as the supremum over all rates R for which a sequence of encoder-decoder pairs can be constructed such that the probability of a guessing error by the decoder tends to zero as n grows to infinity, and such that (4) is satisfied.

A. With CSI

Assume that the state realizations are available to the encoder. In the case of *noncausal* CSI, the encoder is a (possibly random) mapping from the message m and the state sequence s^n to the input sequence x^n. In the case of *causal* CSI, the encoder is a sequence of (possibly random) mappings from m and s^i to x_i, with $i \in \{1, \ldots, n\}$.

Theorem 1: When the transmitter has either noncausal or causal CSI, the capacity is

$$C_{\text{CSI}}^{\text{IID}} = \sup I(U;Y), \tag{6}$$

where the supremum is taken over joint probability distributions of the form

$$P_S(s)P_U(u)P_{X|US}(x|u,s)W(y|x,s) \tag{7}$$

subject to

$$I(S;U,Y) = 0. \tag{8}$$

Proof: The noncausal case follows from [4, Theorem 2] by noting that (8) requires that U be independent of S. It thus remains only to prove the achievability part for the causal case. To this end, fix any joint distribution of the form (7). For each message $m \in \{1, \ldots, \lfloor 2^{nR} \rfloor\}$, randomly generate a vector $u^n(m)$ by choosing each entry IID according to P_U. To send m, the encoder randomly picks its input at time i to be x_i with probability $P_{X|US}(x_i|u_i(m), s_i)$. Each vector $u^n(m)$ is reveaed to the receiver, but the transmitter's choice of x_i is *not* revealed to the receiver. A standard argument shows that the probability of decoding error can be made arbitrarily close to zero as n grows large provided that $R < I(U;Y)$.

We next examine the constraint (4). Note that (8) implies

$$P_{Y|US}(y|u,s) = P_{Y|U}(y|u) \quad \text{for all } s,u,y. \tag{9}$$

When the code is used to transmit a uniformly chosen message, the probability of $Y^n = y^n$ and $S^n = s^n$, for any y^n and s^n, can be written as

$$\begin{aligned} P_{S^nY^n}(s^n,y^n) &= \sum_{m=1}^{\lfloor 2^{nR} \rfloor} \frac{1}{\lfloor 2^{nR} \rfloor} \prod_{i=1}^{n} P_S(s_i) P_{Y|US}(y|u_i(m),s_i) \\ &= \prod_{i=1}^{n} P_S(s_i) \sum_{m=1}^{\lfloor 2^{nR} \rfloor} \frac{P_{Y|U}(y_i|u_i(m))}{\lfloor 2^{nR} \rfloor}. \end{aligned} \tag{10}$$

Clearly, we have $I(S^n;Y^n) = 0$ for every n. ∎

B. No CSI, Deterministic Encoder

We next consider the case where no CSI is available to the encoder, and where the encoder must be deterministic. Thus, the transmitted sequence x^n is a deterministic function of the message m.

Theorem 2: When the transmitter has no CSI and cannot use a stochastic encoder, the capacity is

$$C_{\text{det}}^{\text{IID}} = \sup I(X;Y), \tag{11}$$

where the supremum is taken over joint distributions of the form

$$P_S(s)P_X(x)W(y|x,s) \tag{12}$$

subject to

$$I(S;X,Y) = 0. \tag{13}$$

Proof: For achievability, we generate each codeword IID according to P_X. The analysis is essentially identical to that in the proof of Theorem 1 and hence omitted.

For converse, by the fact that X^n is a deterministic function of the message M, and by Fano's inequality, we have

$$H(X^n|Y^n) \le H(M|Y^n) \le n\epsilon_n, \tag{14}$$

for some $\epsilon_n \downarrow 0$ as $n \to \infty$. We thus have

$$\begin{aligned} I(S^n;X^n,Y^n) &= I(S^n;X^n|Y^n) + I(S^n;Y^n) \\ &\le H(X^n|Y^n) + I(S^n;Y^n) \\ &\le 2n\epsilon_n, \end{aligned} \tag{15}$$

where the last step follows by the constraint (4). We also have

$$\begin{aligned} I(S^n;X^n,Y^n) &= H(S^n) - H(S^n|X^n,Y^n) \\ &= \sum_{i=1}^{n} H(S_i) - H(S_i|X^n,Y^n,S^{i-1}) \\ &\ge \sum_{i=1}^{n} I(S_i;X_i,Y_i) \\ &\ge nI(S;\bar{X},\bar{Y}), \end{aligned} \tag{16}$$

where $\bar{X}$ denotes a random variable whose distribution is the average of the marginal distributions for every X_i, $i = 1, \ldots, n$, and $\bar{Y}$ is the output corresponding to $\bar{X}$. Here, the last step follows because the distributions for S_i are identical, and by the convexity of mutual information in the conditional distribution of (X_i, Y_i) given S_i. Combining (15) and (16) we obtain

$$I(S;\bar{X},\bar{Y}) \le 2\epsilon_n. \tag{17}$$

On the other hand, by the standard converse proof procedure (see, e.g., [9]),

$$R \le I(\bar{X},\bar{Y}) + \epsilon_n. \tag{18}$$

Combining (17) and (18) we obtain that

$$C_{\text{det}}^{\text{IID}} \le \liminf_{n\to\infty} \sup_{P_n} I(X;Y) \tag{19}$$

where the mutual information is computed according to a distribution of the form

$$P_S(s)P_X(x)W(y|x,s) \tag{20}$$

subject to

$$\lim_{n\to\infty} I(S;X,Y) = 0. \tag{21}$$

The converse to the theorem follows by invoking continuity properties of mutual information. ∎

Remark 1: Theorem 2 is equivalent to saying that the transmitter can only use those input symbols that are not affected by S, namely, it can only use x if $W(\cdot|x,s_1) = W(\cdot|x,s_2)$ for all $s_1, s_2 \in \mathcal{S}$.

C. No CSI, Stochastic Encoder

Next we consider the case where the transmitter has no CSI, but is allowed to use a stochastic encoder. The receiver knows the distribution according to which the codebook is chosen, but not the actual choice by the transmitter. Thus, the encoder is a *random* mapping from message m to input sequence x^n, while the decoder is, as before, a mapping from y^n to its guess of m.

Theorem 3: When the transmitter has no CSI but can use a stochastic encoder, the capacity is

$$C_{\text{sto}}^{\text{IID}} = \sup I(U;Y), \tag{22}$$

where the supremum is taken over joint distributions of the form

$$P_S(s)P_U(u)P_{X|U}(x|u)W(y|x,s) \tag{23}$$

subject to

$$I(S;U,Y) = 0. \tag{24}$$

Proof: The achievability part is similar to the previous cases and is omitted. To prove the converse part, we first use Fano's inequality to obtain

$$\begin{aligned} n(R-\epsilon_n) &\le I(M;Y^n) \\ &\le \sum_{i=1}^{n} I(M,Y^{i-1};Y_i). \end{aligned} \tag{25}$$

We also have

$$\begin{aligned} I(S^n;M,Y^n) &= I(S^n;M|Y^n) + I(S^n;Y^n) \\ &\le H(M|Y^n) + I(S^n;Y^n) \\ &\le 2n\epsilon_n, \end{aligned} \tag{26}$$

where the last step follows by Fano's inequality and the constraint (4). On the other hand,

$$\begin{aligned} I(S^n;M,Y^n) &= \sum_{i=1}^{n} I(S_i;M,Y^n,S^{i-1}) \\ &\ge \sum_{i=1}^{n} I(S_i;M,Y^{i-1},Y_i). \end{aligned} \tag{27}$$

Let $U_i \triangleq (M,Y^{i-1})$, $i = 1,\dots,n$. We have shown

$$\sum_{i=1}^{n} I(U_i;Y_i) \ge n(R-\epsilon_n) \tag{28}$$

$$\sum_{i=1}^{n} I(S;U_i,Y_i) \le 2n\epsilon_n, \tag{29}$$

where $\epsilon_n \downarrow 0$ as $n \to \infty$. Note that U_i is independent of S_i because S^n is IID. The rest of the proof is similar to that for Theorem 2. ∎

The next example shows that $C_{\text{sto}}^{\text{IID}}$ can be larger than $C_{\text{det}}^{\text{IID}}$.

Example 1: Consider a channel where $\mathcal{X} = \mathcal{Y} = \{0,1,2\}$ and $\mathcal{S} = \{0,1\}$. The channel law is, when $S = 0$, $Y = X$ with probability one; when $S = 1$, $Y = 0$ if $X = 0$, but the other two symbols are reversed: $Y = 2$ if $X = 1$ and $Y = 1$ if $X = 2$ (all with probability one). A deterministic encoder can only use the input symbol 0, hence it cannot send any information. A stochastic encoder can choose $U \in \{0,1\}$ uniformly, $X = 0$ if $U = 0$, and $X = 1$ or 2 equally likely if $U = 1$. This achieves one bit per channel use. One can verify that this is in fact optimal.

D. A Consequence

A simple consequence to the above results is that the capacity in every case is upper-bounded by the worst-state capacity over all $s \in \mathcal{S}$.

Corollary 4: In all settings above, capacity is upper-bounded by

$$\min_s \sup_{P_X} I(X;Y|S=s). \tag{30}$$

Proof: It suffices to consider the CSI case, since clearly

$$C_{\text{CSI}}^{\text{IID}} \ge C_{\text{sto}}^{\text{IID}} \ge C_{\text{det}}^{\text{IID}}. \tag{31}$$

Recall that, in the formula (6), S must be independent of the pair (U,Y). It follows that

$$I(U;Y) = I(U;Y|S=s) \tag{32}$$

for every $s \in \mathcal{S}$. Hence

$$\begin{aligned} C_{\text{CSI}}^{\text{IID}} &\le \sup_{P_U,P_{X|US}} \min_s I(U;Y|S=s) \\ &\le \min_s \sup_{P_U,P_X|U} I(U;Y|S=s) \\ &\le \min_s \sup_{P_X} I(X;Y|S=s), \end{aligned} \tag{33}$$

where the last step follows because $U \multimap\!\!- (X,S) \multimap\!\!- Y$ form a Markov chain. ∎

Example 2: Consider a channel where $\mathcal{X} = \mathcal{Y} = \mathcal{S} = \{0,1\}$. Assume that P_S is uniform. When $S = 0$, the channel is a perfect bit pipe: $Y = X$ with probability one; when $S = 1$, it is a Z-channel with $1 \to 0$ cross-over probability $p \in (0,1)$ (see [9]). Corollary 4 implies that $C_{\text{CSI}}^{\text{IID}}$ cannot exceed the capacity of the Z-channel. We show that they are equal. Let U be a binary random variable with the capacity-achieving input distribution of the Z-channel. Let $P_{X|US}$ be such that

$$P_{X|US}(1|0,s) = 0, \qquad s = 1,2 \tag{34a}$$

$$P_{X|US}(1|1,0) = 1-p \tag{34b}$$

$$P_{X|US}(1|1,1) = 1, \tag{34c}$$

namely, when $S = 1$, we choose $X = U$ with probability one; when $S = 0$, X is produced by passing U through the above Z-channel. By this choice, we have the same Z-channel from U to Y irrespectively of the value of S, hence $I(S;U,Y) = 0$, whereas $I(U;Y)$ equals the capacity of the Z-channel.

One can show that $C_{\text{sto}}^{\text{IID}} = C_{\text{det}}^{\text{IID}} = 0$. We delay the proof to the end of the next section, when we return to this example.

III. Constant States

Consider the same DMC as described in the first paragraph of Section II. We now assume the state to be constant instead of IID. This means the state is generated randomly according to P_S before communication starts, and remains the same throughout the n channel uses when transmission takes place. The decoder is, as in Section II, a mapping from y^n to a guess of the message. For state obfuscation, we now require

$$\lim_{n\to\infty} I(S;Y^n) = 0. \tag{35}$$

All our claims in this section will continue to hold under the stronger condition that removes the limit in (35). In all cases below, capacity is defined as the supremum over all rates for which one can find a sequence of encoder-decoder pairs such that (35) is satisfied while the decoding error probability will approach zero when n grows large.[1]

A. With CSI

When CSI is available to the transmitter, the encoder is a possibly random mapping from (s,m) to x^n, where m denotes the message and x^n the input sequence. The capacity in this case is the same for constant and IID states.

Theorem 5: For any DMC described by transition law $W(\cdot|\cdot,\cdot)$ and state distribution P_S, the capacity when S is constant and when CSI is available to the transmitter is

$$C_{\text{CSI}}^{\text{const}} = C_{\text{CSI}}^{\text{IID}}, \tag{36}$$

where $C_{\text{CSI}}^{\text{IID}}$ is given by Theorem 1.

Proof: The achievability proof is essentially the same as that for Theorem 1. We note that, since by the choice of joint distribution, the pair (U,Y) is independent of S, we can use typicality to treat (u^n,y^n), even though the state is constant and not ergodic.

To prove the converse, we define auxiliary random variables

$$U_i \triangleq M, Y^{i-1}, \quad i = 1,\dots,n. \tag{37}$$

Using Fano's inequality and the chain rule, we have

$$\begin{aligned} n(R-\epsilon_n) &\le I(M;Y^n) \\ &\le \sum_{i=1}^{n} I(M,Y^{i-1};Y_i) \\ &= \sum_{i=1}^{n} I(U_i;Y_i). \end{aligned} \tag{38}$$

We next show that $I(S;U_i,Y_i)$ must be close to zero for every i. Clearly, it is enough to show that $I(S;M,Y^n)$ must approach zero as n grows large. To this end, define a binary random variable F that equals 0 when decoding is correct and equals 1 when decoding is incorrect. Then we have

$$\begin{aligned} I(S;M,Y^n) &= I(S;Y^n) + I(S;M|Y^n) \\ &\le I(S;Y^n) + I(S;M,F|Y^n) \\ &= I(S;Y^n) + I(S;F|Y^n) + I(S;M|Y^n,F) \\ &\le I(S;Y^n) + H(F) + I(S;M|Y^n,F). \end{aligned} \tag{39}$$

[1]Since the state remains constant during communication, our definition requires that the error probability be small for every possible realization of S.

The first two terms on the right-hand side of (39) both tend to zero as n grows large, the first by (35), and the second because the probability of a decoding error must tend to zero. For the last term, let ϵ denote the probability of a decoding error, then we have

$$\begin{aligned} &I(S;M|Y^n,F) \\ &= (1-\epsilon)\sum_{y^n} \Pr(Y^n=y^n|F=0) I(S;M|Y^n=y^n,F=0) \\ &\quad + \epsilon \sum_{y^n} \Pr(Y^n=y^n|F=1) I(S;M|Y^n=y^n,F=1) \\ &\le (1-\epsilon)\cdot 0 + \epsilon \cdot \log|\mathcal{S}| \\ &= \epsilon \cdot \log|\mathcal{S}|, \end{aligned} \tag{40}$$

which also must tend to zero as n grows large. Hence we have shown that, as n grows large, the right-hand side of (39) must tend to zero, and consequently $I(S;U_i,Y_i)$ must tend to zero for every i. This, together with (38) and a continuity argument, completes the converse proof. ∎

B. No CSI, Deterministic Encoder

Assume that the encoder must be a deterministic mapping that maps the message m to an input sequence x^n. The capacity is again the same as in the IID-state case.

Theorem 6: For any $W(\cdot|\cdot,\cdot)$ and P_S, the capacity in the current setting is

$$C_{\text{det}}^{\text{const}} = C_{\text{det}}^{\text{IID}}. \tag{41}$$

Proof: The achievability is essentially the same as before. For converse, we have, for every $i \in \{1,\dots,n\}$,

$$\begin{aligned} I(S;X_i,Y_i) &\le I(S;X_i,Y^n) \\ &= I(S;Y^n) + I(S;X_i|Y^n) \\ &\le I(S;Y^n) + H(X_i|Y^n). \end{aligned} \tag{42}$$

Since the encoder is deterministic, the decoder should be able to correctly guess every X_i from Y^n (by first guessing M). By Fano's inequality, $H(X_i|Y^n)$ must vanish together with the error probability. Hence, for every i,

$$\lim_{n\to\infty} I(S;X_i,Y_i) = 0. \tag{43}$$

Next consider the communication rate R. For some vanishing ϵ_n,

$$\begin{aligned} n(R-\epsilon_n) &\le I(X^n;Y^n) \\ &\le I(X^n,S;Y^n) \\ &\le \sum_{i=1}^{n} I(X_i,S;Y_i) \\ &\le \sum_{i=1}^{n} I(X_i;Y_i) + I(S;X_i,Y_i). \end{aligned} \tag{44}$$

Combining (43) and (44) completes the converse. ∎

C. No CSI, Stochastic Encoder

When the transmitter has no CSI, a stochastic encoder is a random mapping that maps m to x^n. The decoder knows the distribution used by the stochastic encoder, but not which codebook is chosen. Denote the capacity in this case subject to (35) by $C_{\text{sto}}^{\text{const}}$. We have not been able to find a single-letter expression for $C_{\text{sto}}^{\text{const}}$. One can verify that the achievability part of Theorem 3 is still valid. We can thus order the capacities in various cases as

$$C_{\text{det}}^{\text{IID}} = C_{\text{det}}^{\text{const}} \leq C_{\text{sto}}^{\text{IID}} \leq C_{\text{sto}}^{\text{const}} \leq C_{\text{CSI}}^{\text{IID}} = C_{\text{CSI}}^{\text{const}}. \tag{45}$$

That the first inequality above can be strict was demonstrated by Example 1. The other two inequalities can also be strict, as we show via the next two examples.

Example 3: Let $\mathcal{X} = \mathcal{Y} = \mathcal{S} = \{0, 1\}$. When $S = 0$ the channel is a noiseless bit pipe; when $S = 1$ the bit is flipped at the output with probability one.

We have $C_{\text{sto}}^{\text{IID}} = 0$ because, without CSI and when the states are IID, it is impossible for the transmitter to send any information, even without the constraint (4). We show that

$$C_{\text{sto}}^{\text{const}} = 1 \text{ bit}. \tag{46}$$

Consider the following simple scheme. The transmitter generates a random variable B uniformly over $\{0, 1\}$. To send $(n-1)$ information bits over n channel uses, it sends B followed by the XOR of each information bit and B. The output string is then IID and uniform irrespectively of the value of S. To decode, the receiver obtains $B \oplus S$ from the first bit, and computes its XOR with the next $(n-1)$ received bits to recover the information bits.

Example 4: Consider the same channel as in Example 2, except now the state remains the same for all n channel uses. Recall that $C_{\text{CSI}}^{\text{IID}}$ equals the capacity of the Z-channel; by Theorem 5, so does $C_{\text{CSI}}^{\text{const}}$. We shall show that

$$C_{\text{sto}}^{\text{const}} = 0. \tag{47}$$

Together with (45), this will imply $C_{\text{det}}^{\text{IID}} = C_{\text{det}}^{\text{const}} = C_{\text{sto}}^{\text{IID}} = 0$. To show (47), consider any sequence of encoder-decoder pairs, and define

$$A_n \triangleq \sum_{i=1}^{n} X_i \tag{48}$$

$$B_n \triangleq \sum_{i=1}^{n} Y_i. \tag{49}$$

Further define

$$\alpha \triangleq P\text{-}\limsup_{n\to\infty} \frac{A_n}{n}, \tag{50}$$

where $P\text{-}\limsup$ denotes the limit-supremum in probability: α is the smallest real number for which the probability that $\frac{A_n}{n} > \alpha$ tends to zero as $n \to \infty$. Assume that $\alpha > 0$. Note that, when $S = 0$, $B_n = A_n$ with probability one. Thus we have

$$\limsup_{n\to\infty} \Pr\left(\frac{B_n}{n} \geq \left(1 - \frac{p}{2}\right)\alpha \middle| S = 0 \right) > 0. \tag{51}$$

When $S = 1$, B_n is conditionally a binomial distribution with parameters A_n and p, so its limit-supremum in probability given $S = 1$ must equal $(1-p)\alpha$, therefore

$$\lim_{n\to\infty} \Pr\left(\frac{B_n}{n} \geq \left(1 - \frac{p}{2}\right)\alpha \middle| S = 1 \right) = 0. \tag{52}$$

It follows from (51) and (52) that the total variation distance between the conditional distributions of B_n conditional on $S = 0$ and $S = 1$, respectively, cannot approach zero as n grows large. By Pinsker's Inequality [9], this further implies that $I(S; B_n)$ cannot approach zero, and therefore $I(S; Y^n)$ cannot approach zero either. Thus the assumption that $\alpha > 0$ is incompatible with the requirement (35). But having $\alpha = 0$ clearly does not permit communication at a positive rate. We have thus proven (47).

IV. Concluding Remarks

We have presented information-theoretic capacity expressions for several instances of communication subject to state obfuscation. The case where the state remains constant during transmission time and is unknown to the transmitter, and where the transmitter can use a stochastic encoder, is yet unsolved. We have demonstrated via examples that the capacity in this case differs from both the IID-state no-CSI stochastic-encoder case and the constant-state with-CSI case.

To analyze real-life scenarios where the transmitter wishes to guarantee a low probability of geolocation by the receiver, one may replace the abstract models considered in the current paper by specific channel models. For example, in line-of-sight multiple-antenna wireless communication, the state S may correspond to the phase difference between observation at receive antennas. For free-space optical communication, S may correspond to attenuation of the transmitted signal. Examples 2 and 4 may be considered a first step along the latter direction.

References

[1] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Research and Development*, vol. 2, pp. 289–293, 1958.

[2] S. I. Gel'fand and M. S. Pinsker, "Coding for channels with random parameters," *Prob. Contr. and Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

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

[4] N. Merhav and S. Shamai, "Information rates subject to state masking," *IEEE Trans. Inform. Theory*, vol. 53, pp. 2254–2261, June 2007.

[5] Y.-H. Kim, A. Sutivong, and T. M. Cover, "State amplification," *IEEE Trans. Inform. Theory*, vol. 54, pp. 1850–1859, May 2008.

[6] O. O. Koyluoglu, R. Soundararajan, and S. Vishwanath, "State amplification under masking constraints," in *Proc. 49th Allerton Conf. Comm., Contr. and Comp.*, (Monticello, IL), Sept. 28–30, 2011.

[7] T. Courtade, "Information masking and amplification: The source coding setting," in *Proc. IEEE Int. Symp. Inform. Theory*, (Cambridge, MA, USA), July 1–6 2012.

[8] M. Dikshtein and S. Shamai, "Broadcasting information subject to state masking." 2018, arXiv:1810.11781.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, second ed., 2006.

ITENE: Intrinsic Transfer Entropy Neural Estimator

Jingjing Zhang, Osvaldo Simeone, Zoran Cvetkovic, Eugenio Abela, and Mark Richardson

***Abstract*—Quantifying the directionality of information flow is instrumental in understanding, and possibly controlling, the operation of many complex systems, such as transportation, social, neural, or gene-regulatory networks. The standard Transfer Entropy (TE) metric follows Granger's causality principle by measuring the Mutual Information (MI) between the past states of a source signal X and the future state of a target signal Y while conditioning on past states of Y. Hence, the TE quantifies the improvement, as measured by the log-loss, in the prediction of the target sequence Y that can be accrued when, in addition to the past of Y, one also has available past samples from X. However, by conditioning on the past of Y, the TE also measures information that can be synergistically extracted by observing both the past of X and Y, and not solely the past of X. Building on a private key agreement formulation, the Intrinsic TE (ITE) aims to discount such synergistic information to quantify the degree to which X is *individually* predictive of Y, independent of Y's past. In this paper, an estimator of the ITE is proposed that is inspired by the recently proposed Mutual Information Neural Estimation (MINE). The estimator is based on variational bound on the KL divergence, two-sample neural network classifiers, and the pathwise estimator of Monte Carlo gradients.**

***Index Terms*—Transfer entropy, neural networks, machine learning, intrinsic transfer entropy.**

I. Introduction

A. Context and Key Definitions

Quantifying the causal flow of information between different components of a system is an important task for many natural and engineered systems, such as neural, genetic, transportation and social networks. A well-established metric that has been widely applied to this problem is the information-theoretic measure of Transfer Entropy (TE) [1], [2]. To define it mathematically, consider two jointly stationary random processes $\{X_t, Y_t\}$ with $t = 1, 2, \dots$ The TE from process $\{X_t\}$ to process $\{Y_t\}$ with memory parameters (m, n) is defined as the conditional Mutual Information (MI) [1], [3]

$$\text{TE}_{X\to Y}(m,n) \triangleq I(X_{t-m}^{t-1}; Y_t | Y_{t-n}^{t-1}), \tag{1}$$

where $X_{t-m}^{t-1} = (X_{t-m}, \dots, X_{t-1})$ and $Y_{t-n}^{t-1} = (Y_{t-n}, \dots, Y_{t-1})$ denote the past m and n samples of time sequences $\{X_t\}$ and $\{Y_t\}$. By definition (1), the TE measures the MI between the past m samples of process $\{X_t\}$ and the current sample Y_t of process $\{Y_t\}$ when conditioning on the past n samples Y_{t-n}^{t-1} of the same process. Therefore, the TE quantifies the amount by which the prediction of the sample Y_t can be improved, in terms of average log-loss in bits, through the knowledge of m samples of process $\{X_t\}$ when the past n samples of the same process $\{Y_t\}$ are also available. While not further considered in this paper, we note for reference that a related information-theoretic measure that originates from the analysis of communication channels with feedback [4], [5] is the Directed Information (DI). The DI is defined as

$$\text{DI}_{X\to Y} \triangleq \frac{1}{T}\sum_{t=1}^{T} I(X_1^{t-1}; Y_t | Y_1^{t-1}), \tag{2}$$

where we have normalized by the number T of samples to facilitate comparison with TE. For jointly Markov processes[1] $\{X_t\}$, $\{Y_t\}$ with memory parameters m and n, the TE (1) is an upper bound on the DI (2) [6].

The TE, and the DI, have limitations as measures of *intrinsic, or exclusive*, information flow from $\{X_t\}$ to $\{Y_t\}$. This is due to the fact that conditioning on past samples of $\{Y_t\}$ does not discount the information that the past samples of $\{Y_t\}$ contain about its current sample Y_t: Conditioning also captures the information that can be *synergistically* obtained by observing both past samples X_{t-m}^{t-1} and Y_{t-n}^{t-1}. In fact, there may be information about Y_t that can be extracted from X_{t-m}^{t-1} only if this is observed jointly with Y_{t-n}^{t-1}. This may not be considered as part of the intrinsic information flow from $\{X_t\}$ to $\{Y_t\}$.

Example [7]: Assume that the variables are binary, and that the joint distribution of the variables (X_{t-1}, Y_{t-1}, Y_t) is given as $p(0,0,0) = p(0,1,1) = p(1,0,1) = p(1,1,0) = 1/4$. It can be seen that observing both X_{t-1} and Y_{t-1} allows the future state Y_t to be determined with certainty, while X_{t-1} alone is not predictive of Y_t, since X_{t-1} and Y_t are statistically independent. The TE with memory parameter $m = n = 1$ is given as $\text{TE}_{X\to Y}(1,1) = I(X_{t-1}; Y_t | Y_{t-1}) = 1$ bit, although there is no *intrinsic* information flow between the two sequences but only a synergistic mechanism relating both Y_{t-1} and X_{t-1} to Y_t. □

In order to distinguish intrinsic and synergistic information flows, reference [7] proposed to decompose the TE into Intrinsic Transfer Entropy (ITE) and Synergistic Transfer Entropy (STE). The ITE aims to capture the amount of information on Y_t that is contained in the past of $\{X_t\}$ in addition to that already present in the past of $\{Y_t\}$; while the STE measures the information about Y_t that is obtained only when combining the past of both $\{X_t\}$ and $\{Y_t\}$. Formally, the ITE from

J. Zhang, O. Simeone, and Z. Cvetkovic are with the Department of Engineering at King's College London, UK (emails: jingjing.1.zhang@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk, zoran.cvetkovic@kcl.ac.uk). E. Abela and M. Richardson are with the Department of Basic and Clinical Neuroscience at King's College London, UK (emails: eugenio.abela@kcl.ac.uk, mark.richardson@kcl.ac.uk). J. Zhang and O. Simeone have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731). J. Zhang has also been supported by a King's Together award. Code can be found at https://github.com/kclip/ITENE.

[1]This implies the Markov chain $Y_t - (X_{t-m}^{t-1}, Y_{t-n}^{t-1}) - (X_1^{t-m-1}, Y_1^{t-n-1})$.

process $\{X_t\}$ to process $\{Y_t\}$ with memory parameters (m, n) is defined as [7]

$$\mathrm{ITE}_{X\to Y}(m,n) \triangleq \inf_{p(\bar{y}_{t-n}^{t-1}|y_{t-n}^{t-1})} I(X_{t-m}^{t-1}; Y_t|\bar{Y}_{t-n}^{t-1}). \quad (3)$$

In definition (3), auxiliary variables $\bar{Y}_{t-n}^{t-1}$ can take values without loss of generality in the same alphabet as the corresponding variables Y_{t-n}^{t-1} [8], and are obtained by optimising the conditional distribution $p(\bar{y}_{t-n}^{t-1}|y_{t-n}^{t-1})$. The quantity (3) can be shown to be an upper bound on the size (in bits) of a secret key that can be generated by two parties, one holding X_{t-m}^{t-1} and the other Y_t, via public communication when the adversary has Y_{t-n}^{t-1} [9]. This intuitively justifies its use as a measure of intrinsic information flow. The STE is then defined as the residual

$$\mathrm{STE}_{X\to Y}(m,n) \triangleq \mathrm{TE}_{X\to Y}(m,n) - \mathrm{ITE}_{X\to Y}(m,n). \quad (4)$$

B. TE and DI Estimation

The TE can be estimated using tools akin to the estimation of MI, including plug-in methods [10], non-parametric techniques based on kernel [1] or k-nearest-neighbor (k-NN) methods [11], [12], and parametric techniques, such as Maximum Likelihood [13] or Bayesian estimators [14]. Popular implementations of some of these standard methods can be found in the Java Information Dynamics Toolkit (JIDT) [15] and TRENTOOL toolbox [16]. For the DI, estimators have been designed that rely on parametric and non-parametric techniques, making use also of universal compressors [17]–[19]. In order to enable scaling over large data sets and/or data dimensions, MI estimators that leverage neural networks have been recently the subject of numerous studies. Notably, reference [20] introduced the Mutual Information Neural Estimator (MINE), which reduces the problem of estimating MI to that of classifying dependent vs. independent pairs of samples via the Donsker-Varadhan (DV) variational equality. Specifically, reference [20] proposes to train a neural network to approximate the solution of the optimization problem defined by the DV equality. The follow-up paper [21] proposes to train a two-sample neural network classifier, which is then used as an approximation of the likelihood ratio in the DV equality. Theoretical limitations of general variational MI estimators were derived in [22], which also proposes a variational MI estimator with reduced variance. We note that reference [21] also considers the estimation of the conditional MI, which applies directly to the estimate of the TE as discussed in Section II.

C. Main Contributions, Paper Organization, and Notation

This work proposes an estimator, referred to as ITE Neural Estimator (ITENE), of the ITE that is based on two-sample classifier and on the pathwise estimator of Monte Carlo gradients, also known as reparameterization trick [23]. We also present numerical results to illustrate the performance of the proposed estimator. The paper is organized as follows. In Section II, we review the classifier-based MINE approach proposed in reference [21]. Based on this approach, we introduce the proposed ITENE method in Section III. Section IV presents experimental results. Throughout this paper, we use uppercase letters to denote random variables and corresponding lowercase letters to denote their realizations. $\log$ represents the natural logarithm. $\nabla_x f(x)$ represents the gradient of scalar function $f(x)$ and $\mathrm{J}_x f(x)$ the Jacobian matrix of vector function $f(x)$.

II. Background: Classifier-based Mutual Information Neural Estimator (MINE)

In this section, we review the classifier-based MINE for the estimation of the MI $I(U;V)$ between jointly distributed continuous random variables U and V. The MI satisfies the DV variational representation [24]

$$I(U;V) = \sup_{f(u,v)} \mathrm{E}_{p(u,v)}[f(U,V)] - \log(\mathrm{E}_{p(u)p(v)}[e^{f(U,V)}]) \quad (5a)$$

$$= \sup_{r(u,v)} \mathrm{E}_{p(u,v)}\left[\log\left(\frac{r(U,V)}{\mathrm{E}_{p(u)p(v)}[r(U,V)]}\right)\right], \quad (5b)$$

where the supremum is taken over all functions $f(U,V)$ in (5a) and $r(U,V) = e^{f(U,V)}$ in (5b) such that the two expectations in (5a) are finite. Note that (5) contains expectations both over the joint distribution $p(u,v)$ of U and V and over the product of the marginals $p(u)$ and $p(v)$. Intuitively, the functions $f(u,v)$ and $r(u,v)$ act as classifiers of a sample (u,v) being either generated by the joint distribution $p(u,v)$ or by the product distribution $p(u)p(v)$. This is done by functions $f(u,v)$ and $r(u,v)$ ideally outputing a larger value in the former case than in the latter [25, Chapter 6]. More precisely, following [22], we can interpret function $r(u,v)$ as an unnormalized estimate of the likelihood ratio $p(u,v)/\big(p(u)p(v)\big)$, with $\tilde{r}(U,V) = r(U,V)/\mathrm{E}_{p(u)p(v)}[r(U,V)]$ being its normalized version. This normalization ensures the condition $\mathrm{E}_{p(u)p(v)}[\tilde{r}(U,V)] = 1$, which is satisfied by the true likelihood ratio $p(u,v)/(p(u)p(v))$ [22]. Mathematically, the supremum in (5b) is achieved when $r(u,v)$ is equal to the likelihood ratio [22, Theorem 1], i.e.,

$$r^*(u,v) = \frac{p(u,v)}{p(u)p(v)}. \quad (6)$$

This observation motivates the classifier-based estimator introduced in [21]. To elaborate, given a data set $\mathcal{D} = \{(u_t, v_t)\}_{t=1}^T$ of T data points from the joint distribution $p(u,v)$, we label the samples with a target value $a = 1$. Furthermore, we construct a data set $\mathcal{D}_0$ approximately distributed according to the product distribution $p(u)p(v)$ by randomly resampling the values of v_t (see line 3 in Algorithm 1). These samples are labeled as $a = 0$. We use notation $p(a = 1|u,v)$ to represent the posterior probability that a sample is generated from the distribution $p(u,v)$ when the hypotheses $a = 1$ and $a = 0$ are a priori equally likely. An estimate of the probability $p(a = 1|u,v)$ can be obtained by training a function $p_\theta(a = 1|u,v)$ parametrized as a neural network with input u and v, target output a, and weight vector θ. This is done via the minimization of the empirical cross-entropy loss evaluated on the described data sets (see lines

Algorithm 1 Classifier Based MINE [20], [21]

1: **Input:**
$\mathcal{D}_1 = \{(u_t, v_t)\}_{t=1}^T$: observed data samples
2: **Output:**
$\hat{I}(U;V)$: mutual information estimate
3: obtain data set $\mathcal{D}_0 = \{(u_n, v_{\pi(n)})\}_{n=1}^T$, where $\pi(n)$ is sampled i.i.d. from set $\{1, \dots, T\}$
4: label samples $i \in \mathcal{D}_1$ as $a = 1$ and $j \in \mathcal{D}_0$ as $a = 0$ to create labeled data sets $\bar{\mathcal{D}}_1$ and $\bar{\mathcal{D}}_0$
5: $\theta \leftarrow$ initialize neural network parameters
6: $\alpha \leftarrow$ set learning rate
7: $\tau \leftarrow$ set hyperparameter
8: split $\bar{\mathcal{D}}_1$ into two subsets $\bar{\mathcal{D}}_{1,t}$ (training) and $\bar{\mathcal{D}}_{1,e}$ (estimation)
9: split $\bar{\mathcal{D}}_0$ into two subsets $\bar{\mathcal{D}}_{0,t}$ (training) and $\bar{\mathcal{D}}_{0,e}$ (estimation)
10: train binary classifier using training set $\{\bar{\mathcal{D}}_{1,t}, \bar{\mathcal{D}}_{0,t}\}$
11: output: $\hat{I}(U;V) = \frac{1}{|\bar{\mathcal{D}}_{1,e}|}\sum_{i\in\bar{\mathcal{D}}_{1,e}} \log \frac{p_\theta(a=1|i)}{1-p_\theta(a=1|i)} - \log\big(\frac{1}{|\bar{\mathcal{D}}_{0,e}|}\sum_{j\in\bar{\mathcal{D}}_{0,e}} \mathrm{clip}_\tau(\frac{p_\theta(a=1|j)}{1-p_\theta(a=1|j)})\big)$

8-10 in Algorithm 1) via Stochastic Gradient Descent (SGD) (see, e.g., [25, Chapter 6]). Having completed training, the likelihood ratio can be estimated as

$$\hat{r}_\theta(u,v) = \frac{p_\theta(a=1|u,v)}{1-p_\theta(a=1|u,v)}. \tag{7}$$

This follows since, at convergence, if training is successful, the following equality holds approximately

$$\begin{aligned} &p_\theta(a=1|u,v) \\ &= \frac{p(a=1)p(u,v|a=1)}{p(a=1)p(u,v|a=1)+p(a=0)p(u,v|a=0)} \\ &= \frac{p(u,v)}{p(u,v)+p(u)p(v)}. \end{aligned} \tag{8}$$

Finally, the estimate (7) can be plugged into an empirical approximation of (5b) as

$$\hat{I}(U;V) = \mathrm{E}_{\hat{p}(u,v)}\Big[\log\Big(\frac{\hat{r}_\theta(U,V)}{\mathrm{E}_{\hat{p}(u)\hat{p}(v)}[\mathrm{clip}_\tau(\hat{r}_\theta(U,V))]}\Big)\Big], \tag{9}$$

where $\hat{p}(u,v)$ represents the empirical distribution of the observed data sample pairs in an held-out part of data set $\mathcal{D}_1$, while $\hat{p}(u)$ and $\hat{p}(v)$ are the corresponding empirical marginal distributions for U and V (see line 11 in Algorithm 1); and the clip function is defined as $\mathrm{clip}_\tau(v) = \max\{\min\{v, e^\tau\}, e^{-\tau}\}$ with some constant $\tau \geq 0$ [22]. Clipping was suggested in [22] in order to reduce variance of the estimate (9), and a similar approach is also used in [21]. The estimator (9) is known to be consistent but biased [20], and an analysis of the variance can be found in [22] (see also Lemma 1 below). Details are presented in Algorithm 1.

III. Intrinsic Transfer Entropy Neural Estimator ITENE

In this section, inspired by the classifier-based MINE, we introduce an estimator for the ITE, which we refer to as ITENE. Throughout this section, we assume the availability of data in the form of time series $\mathcal{D} = \{(x_t, y_t) : t = 1, 2, \dots, T\}$ generated as a realization of jointly stationary random processes $\{X_t, Y_t\}_{t\geq 1}$. We use the notations $X_t^- \triangleq X_{t-m}^{t-1}$, $Y_t^- \triangleq Y_{t-n}^{t-1}$ and $Y_t^0 \triangleq Y_t$ and we also drop the subscript t when no confusion may arise.

A. TENE

We start by noting that, using the chain rule [26], the TE in (1) can be written as the difference

$$\mathrm{TE}_{X\to Y}(m,n) = I(X^-; Y^0, Y^-) - I(X^-; Y^-). \tag{10}$$

Therefore, the TE can be estimated by applying the classifier-based MINE in Algorithm 1 to both terms in (10) separately. This approach was proposed in [21] and found empirically to outperform other estimates of the conditional MI. Accordingly, we have the estimate

$$\widehat{\mathrm{TE}}_{X\to Y}(m,n) = \hat{I}(X^-; Y^0, Y^-) - \hat{I}(X^-; Y^-), \tag{11}$$

where the MINE estimates in (9) are obtained by applying Algorithm 1 to the data sets $\mathcal{D}_1^A = \{u_t = x_t^-, v_t = (y_t^0, y_t^-)\}_{t=1}^T$ and $\mathcal{D}_1^B = \{u_t = x_t^-, v_t = y_t^-\}_{t=1}^T$, respectively (zero padding is used for out-of-range indices). We refer to the resulting estimator (11) as TENE. Following [21], TENE is consistent but biased. Furthermore, without using clipping, i.e., when $\tau \to \infty$, we have that the following lemma holds.

Lemma 1: Assume that the estimates $\hat{r}_\theta(x^-, y^0, y^-)$ and $\hat{r}_\theta(x^-, y^-)$ equal their respective true likelihood ratios, i.e., $\hat{r}_\theta(x^-, y^0, y^-) = p(x^-, y^0, y^-)/(p(x^-)p(y^0, y^-))$ and $\hat{r}_\theta(x^-, y^-) = p(x^-, y^-)/(p(y^-)p(y^-))$. Then, under the randomness of the sampling procedure generating the data set $\mathcal{D}$, we have

$$\lim_{T\to\infty} T\mathrm{Var}[\widehat{\mathrm{TE}}_{X\to Y}(m,n)] \geq e^{I(X^-;Y^0,Y^-)} + e^{I(X^-;Y^-)} - 2. \tag{12}$$

The proof follows directly from [22, Theorem 1]. Lemma 1 demonstrates that, without clipping, the variance of TENE in (11) can grow exponentially with the maximum of the true values of $I(X^-; Y^0, Y^-)$ and $I(X^-; Y^-)$. Note that a similar result applies to MINE [22]. Setting a suitable value for τ is hence important in order to obtain reliable estimates.

B. ITENE

We now move on to the estimator of the ITE (3). To this end, we first parameterize the distribution $p_\phi(\bar{y}^-|y^-)$ under optimization as

$$\bar{y}_\phi^- = \mu_\phi(y^-) + \sigma_\phi(y^-) \odot \epsilon, \tag{13}$$

where $\mu_\phi(y^-)$ and $\log\sigma_\phi(y^-)$ are disjoint sets of outputs of a neural network with weights ϕ; $\odot$ is the element-wise product; and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is a Gaussian vector independent of all other variables. Parameterization (13) follows the so-called reparameterization trick popularized by the variational auto-encoder [27]. An estimator of the ITE (3) can be defined by optimizing over ϕ the ITE (10) as

$$\widehat{\mathrm{ITE}}_{X\to Y}(m,n) = \inf_\phi\big(\hat{I}_\phi(X^-; Y^0, \bar{Y}^-) - \hat{I}_\phi(X^-; \bar{Y}^-)\big), \tag{14}$$

Algorithm 2 ITENE

Input:
$\mathcal{D} = \{(x_t, y_t)\}_{t=1}^T$: observed data samples from the random process $\{X_t, Y_t\}$
Output:
$\widehat{\mathrm{ITE}}_{X\to Y}(m,n)$: ITE estimate
$(\phi, \theta, \theta') \leftarrow$ initialize network parameters
$\alpha \leftarrow$ set learning rate
$\tau \leftarrow$ set hyperparameter
repeat
  randomly generate T samples $\{\epsilon_t\}_{t=1}^T$ from distribution $\mathcal{N}(0, \mathbf{I})$
  for each $t = 1, \ldots, T$:
    compute $\bar{y}_{\phi,t}^- = \mu_\phi(y_t^-) + \sigma_\phi(y_t^-) \odot \epsilon_t$
  define data set $\mathcal{D}^A = \{u_t^A, v_t^A\}_{t=1}^T$, with $u_t^A = x_t^-, v_t^A = \{y_t^0, \bar{y}_{\phi,t}^-\}$
  apply Algorithm 1 to output $\hat{I}_\phi(X^-; Y^0, Y^-) = \hat{I}(U^A; V^A)$
  define data set $\mathcal{D}^B = \{u_t^B, v_t^B\}_{t=1}^T$, with $u_t^B = x_t^-, v_t^B = \bar{y}_{\phi,t}^-$
  apply Algorithm 1 to output $\hat{I}_\phi(X^-; Y^-) = \hat{I}(U^B; V^B)$
  update the network parameters using the pathwise gradient estimators (17)-(19)
    $\phi \leftarrow \phi - \alpha \nabla_\phi \big(\hat{I}_\phi(X^-; Y^0, Y^-) - \hat{I}_\phi(X^-; Y^-)\big)$
until convergence
output:
$\widehat{\mathrm{ITE}}_{X\to Y}(m,n) = \hat{I}_\phi(X^-; Y^0, Y^-) - \hat{I}_\phi(X^-; Y^-)$

where we have made explicit the dependence of estimates $\hat{I}_\phi(X^-; Y^0, \bar{Y}^-)$ and $\hat{I}_\phi(X^-; \bar{Y}^-)$ on ϕ. In particular, using (10), the first MINE estimate in (11) can be written as a function of ϕ as

$$\hat{I}_\phi(X^-; Y^0, \bar{Y}^-) = \mathrm{E}_{\hat{p}(x^-, y^0, y^-)}\big[\mathrm{E}_{p(\epsilon)}[\log(\hat{r}_\theta(X^-, Y^0, \bar{Y}_\phi^-))]\big] - \log\big(\mathrm{E}_{\hat{p}(x^-)\hat{p}(y^0, y^-)}\big[\mathrm{E}_{p(\epsilon)}[\mathrm{clip}_\tau(\hat{r}_\theta(X^-, Y^0, \bar{Y}_\phi^-))]\big], \quad (15)$$

where parameter θ is obtained from Algorithm 1 by considering as input the data set $\mathcal{D}_{\phi,1}^A = \{u_t = x_t^-, v_t = (y_t^0, \bar{y}_{\phi,t}^-)\}_{t=1}^T$, where samples $\bar{y}_{\phi,t}^-$ are generated using (13) as $\bar{y}_{\phi,t}^- = \mu_\phi(\bar{y}_t) + \sigma_\phi(\bar{y}_t) \odot \epsilon_t$ for i.i.d. samples $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. Furthermore, the empirical distributions $\hat{p}(\cdot)$ in (15) are obtained from the held-out (estimation) data set in Algorithm 1. In a similar manner, the second MINE estimate in (14) is given as

$$\hat{I}_\phi(X^-; \bar{Y}^-) = \mathrm{E}_{\hat{p}(x^-, y^-)}\big[\mathrm{E}_{p(\epsilon)}[\log(\hat{r}_{\theta'}(X^-, \bar{Y}_\phi^-))]\big] - \log\big(\mathrm{E}_{\hat{p}(x^-)\hat{p}(y^-)}\big[\mathrm{E}_{p(\epsilon)}[\mathrm{clip}_\tau(\hat{r}_{\theta'}(X^-, \bar{Y}_\phi^-))]\big], \quad (16)$$

where parameter θ' is obtained from Algorithm 1 by considering as input the data set $\mathcal{D}_{\phi,1}^B = \{u_t = x_t^-, v_t = \bar{y}_{\phi,t}^-)\}_{t=1}^T$.

We propose to tackle problem (14) in a block coordinate fashion by iterating between SGD steps with respect to ϕ and updates of parameters (θ, θ') using Algorithm 1. To this end, when fixing (θ, θ'), the optimization over parameter ϕ requires the gradient

$$\nabla_\phi \hat{I}_\phi(X^-; Y^0, \bar{Y}^-) = \mathrm{E}_{\hat{p}(x^-, y^0, y^-)}\left[\mathrm{E}_{p(\epsilon)}\left[\frac{\nabla_{\bar{y}_\phi^-}\hat{r}_\theta}{\hat{r}_\theta} \times \mathrm{J}_\phi \bar{y}_\phi^-\right]\right] - \frac{\mathrm{E}_{\hat{p}(x^-)\hat{p}(y^0, y^-)}\big[\mathrm{E}_{p(\epsilon)}[\nabla_{\bar{y}_\phi^-}\hat{r}_\theta \times \mathrm{J}_\phi \bar{y}_\phi^-]\big]}{\mathrm{E}_{\hat{p}(x^-)\hat{p}(y^0, y^-)}\big[\mathrm{E}_{p(\epsilon)}[\hat{r}_\theta]\big]}, \quad (17)$$

where, from (7), we have the gradient

$$\nabla_{\bar{y}_\phi^-}\hat{r}_\theta = \frac{\nabla_{\bar{y}_\phi^-} p_\theta(a=1|x^0, y^-, \bar{y}_\phi^-)}{(1 - p_\theta(a=1|x^0, y^-, \bar{y}_\phi^-))^2}; \quad (18)$$

and, from (13), we have the Jacobian $\mathrm{J}_\phi \bar{y}_\phi^- = \mathrm{J}_\phi \mu_\phi(Y^-) + \big(\mathrm{J}_\phi(\sigma_\phi(Y^-)\big) \odot \epsilon$. It also requires the gradient

$$\nabla_\phi \hat{I}_\phi(X^-; \bar{Y}^-) = \mathrm{E}_{\hat{p}(x^-, y^-)}\left[\mathrm{E}_{p(\epsilon)}\left[\frac{\nabla_{\bar{y}_\phi^-}\hat{r}_{\theta'}}{\hat{r}_{\theta'}} \times \mathrm{J}_\phi \bar{y}_\phi^-\right]\right] - \frac{\mathrm{E}_{\hat{p}(x^-)\hat{p}(y^-)}\big[\mathrm{E}_{p(\epsilon)}[\nabla_{\bar{y}_\phi^-}\hat{r}_{\theta'} \times \mathrm{J}_\phi \bar{y}_\phi^-]\big]}{\mathrm{E}_{\hat{p}(x^-)\hat{p}(y^-)}\big[\mathrm{E}_{p(\epsilon)}[\hat{r}_{\theta'}]\big]}, \quad (19)$$

where we have

$$\nabla_{\bar{y}_\phi^-}\hat{r}_{\theta'} = \frac{\nabla_{\bar{y}_\phi^-} p_{\theta'}(a=1|x^0, \bar{y}_\phi^-)}{(1 - p_{\theta'}(a=1|x^0, \bar{y}_\phi^-))^2}. \quad (20)$$

We note that the gradients (17)-(19) are instances of pathwise gradient estimators [23]. The resulting ITENE is summarized in Algorithm 2. Due to the consistency of TENE, ITENE is also consistent if the capacity of the model p_ϕ is large enough.

IV. Experiments

In this section, we provide some results to illustrate the type of insights that can be obtained by decomposing the TE into ITE and STE as in (4). To this end, consider first the following simple example. The joint processes $\{X_t, Y_t\}_{t\geq 1}$ are generated according to

$$Y_t = \begin{cases} Z_t, & \text{if } Y_{t-1} < \lambda \\ \rho X_{t-1} + \sqrt{1-\rho^2} Z_t, & \text{if } Y_{t-1} \geq \lambda, \end{cases} \quad (21)$$

for some threshold λ, where variables $\{X_t, Y_t\}$ are independent and identically distributed as $\mathcal{N}(0,1)$. Intuitively, for large values of the threshold λ, there is no information flow between $\{X_t\}$ and $\{Y_t\}$, while for small values, there is a purely intrinsic flow of information. For intermediate values of λ, the information flow is partly synergistic, since knowing both Y_{t-1} and X_{t-1} is instrumental in obtaining information about Y_t. To quantify the intuition above, we apply the discussed estimators with $m = n = 1$. To this end, for all two-sample neural network classifiers, we consider two hidden layers with 100 hidden neurons with ELU activation functions, while for the probability $p_\phi(\bar{y}^-|y^-)$, we adopt a neural network with hidden layer of 200 neurons with ELU activation functions and outputs $\mu_\phi(y^-)$ and $\log(\sigma_\phi(y^-))$. The data set size T is split into a 75%-fraction for classifier training and a 25%-fraction for estimation. We set learning rate $\alpha = 0.001$ and clipping parameter $\tau = 0.9$.

The computed estimates $\widehat{\mathrm{TE}}_{X\to Y}(1,1)$, $\widehat{\mathrm{ITE}}_{X\to Y}(1,1)$, $\widehat{\mathrm{STE}}_{X\to Y}(1,1)$ are plotted in Fig. 1 as a function of the threshold λ, along with the true TE. The latter can be computed in closed form as $\mathrm{TE}_{X\to Y}(m,n) = \mathrm{TE}_{X\to Y}(1,1) = -0.5Q(\lambda)\log(1-\rho^2)$ (nats), where $Q(\cdot)$ is the standard complementary cumulative distribution function of a standard Gaussian variable. In a manner consistent with the intuition

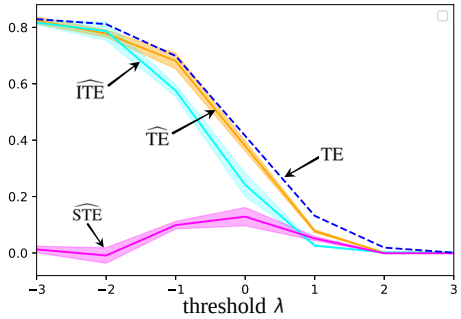


Figure 1: TENE, ITENE, STENE (obtained as the difference (4)) and true TE versus threshold λ with $\rho = 0.9$ for the example (21). Dashed areas represent the range of observed estimates within 10 trials.

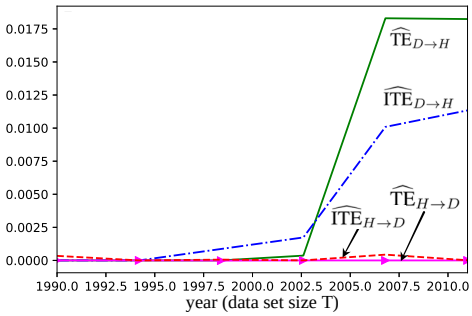


Figure 2: TENE and ITENE between the DJIA, denoted as "D", and the HSI, denoted as "H".

provided above, when λ is either small, i.e., $\lambda \leq -2$, or large, i.e., $\lambda \geq 2$, the ITE is seen in Fig. 1 to be close to the TE, yielding nearly zero STE. This is not the case for intermediate values of λ, in which regime a non-negligible STE is observed.

For a real-world example, we apply the estimators at hand to historic data of the values of the Hang Seng Index (HSI) and of the Dow Jones Index (DJIA) between 1990 and 2011. As done in [17], for each stock, we classify its values into three levels, namely 1, 0, and -1, where 1 indicates an increase in the stock price by more than 0.8% in one day, -1 indicates a drop by more than -0.8%, and 0 indicates all other cases. As illustrated in Fig. 2, and in line with the results in [17], both the TE and ITE from the DJIA to the HSI are much larger than in the reverse direction, implying that the DJIA influenced the HSI more significantly than the other way around for the given time range. Furthermore, we observe that not all the information flow is estimated to be intrinsic, and hence the joint observation of the history of the DJIA and of the HSI is partly responsible for the predictability of the HSI from the DJIA.

References

[1] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul. 2000.

[2] R. Vicente, M. Wibral, and G. Lindner, Michaeland Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 45–67, Feb. 2011.

[3] M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente, "Measuring information-transfer delays," *PLOS ONE*, vol. 8, pp. 1–19, Feb. 2013.

[4] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Information Theory Applications (ISITA)*, Waikiki, Hawaii, Nov. 1990.

[5] H. H. Permuter, Y. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3248–3259, Jun. 2011.

[6] Y. Liu and S. Aviyente, "The relationship between transfer entropy and directed information," in *Proc. of Statistical Signal Process. Workshop (SSP)*, Michigan, USA, Aug. 2012, pp. 73–76.

[7] R. G. James, B. D. M. Ayala, B. Zakirov, and J. P. Crutchfield, "Modes of information flow." [Online]. Available: https://arxiv.org/abs/1808.06723

[8] J. P. Crutchfield and D. P. Feldman, "Regularities unseen, randomness observed: levels of entropy convergence," *Chaos*, vol. 13, no. 1, p. 25–54, 2003.

[9] U. M. Maurer and S. Wolf, "Unconditionally secure key agreement and the intrinsic conditional information," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 499–514, Mar. 1999.

[10] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L2 theory," *Probability Theory and Related Fields*, vol. 57, no. 4, pp. 453–476, Dec. 1981.

[11] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, Jun. 2004.

[12] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Phys. Rev. Lett.*, vol. 99, p. 204101(4), Nov. 2007.

[13] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori, "Approximating mutual information by maximum likelihood density ratio estimation," in *Proc. of the Int. Conf. on New Challenges for Feature Selection in Data Min. and Knowledge Discovery*, 2008, pp. 5–20.

[14] D. H. Wolpert and D. R. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Phys. Rev. E*, vol. 52, pp. 6841–6854, Dec. 1995.

[15] J. T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems," *Frontiers in Robotics and AI*, vol. 1, p. 11, Dec. 2014.

[16] M. Lindner, R. Vicente, V. Priesemann, and M. Wibral, "TRENTOOL: A matlab open source toolbox to analyse information flow in time series data with transfer entropy," *BMC Neuroscience 12, 119*, Nov. 2011.

[17] J. Jiao, H. H. Permuter, L. Zhao, Y. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6220–6242, Oct. 2013.

[18] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 17–44, Feb. 2011.

[19] R. Malladi, G. Kalamangalam, N. Tandon, and B. Aazhang, "Identifying seizure onset zone from the causal connectivity inferred using directed information," *IEEE Journal of Selected Topics in Signal Process.*, vol. 10, no. 7, pp. 1267–1283, Oct. 2016.

[20] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mutual information neural estimation," in *Proc. Int. Conf. on Machine Learning*, Stockholm, Sweden, Jul. 2018.

[21] S. Mukherjee, H. Asnani, and S. Kannan, "CCMI: Classifier based conditional mutual information estimation," in *Proc. the Conference on Uncertainty in Artificial Intelligence (UAI)*, Tel Aviv, Israel, Jul. 2019.

[22] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," 2019. [Online]. Available: https://arxiv.org/abs/1910.06222

[23] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, "Monte carlo gradient estimation in machine learning." [Online]. Available: https://arxiv.org/abs/1906.10652

[24] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time," *Communications on Pure and Applied Mathematics*, vol. 36, pp. 183–212, 1983.

[25] O. Simeone, *A Brief Introduction to Machine Learning for Engineers*. Foundations and Trends in Signal Processing, 2018. [Online]. Available: http://arxiv.org/abs/1709.02840

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, USA: Wiley-Interscience, 1991.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. on Learning Representations (ICLR)*, Scottsdale, USA, May 2013.

Sampling for Faster Neural Estimation

Chung Chan
City University of Hong Kong
Hong Kong
Email: chung.chan@cityu.edu.hk

Abstract

In training a neural network by gradient descent, the loss is estimated based only on a limited number of sample outputs of the neural network. Therefore, if more samples can be placed in regions where the neural network is different from the optimal solution, the neural network should converge faster to the optimal solution. We demonstrate that a better sampling distribution could significantly speed up the convergence rate of a recent promising neural estimation of mutual information proposed by Belgahzi *et al.* The method, called the mutual information neural estimation (MINE), trains the neural network to maximize a tractable lower bound of the divergence in terms of its Fenchel–Legendre transform. In particular, we discover a limitation of MINE where the network has slow staircase convergence when estimating the mutual information of a simple mixed Gaussian distribution with overlapping modes. To solve this problem, we propose a faster method called the mutual information neural entropic estimation (MI-NEE). Our solution first generalizes MINE to estimate the entropy using a custom reference distribution. The entropy estimate can then be used to estimate the mutual information. The seemingly unnecessary intermediate step of entropy estimation allows one to improve the convergence by an appropriate reference distribution that samples the neural network around regions of interest. This idea may be further generalized to adaptive sampling and cross-training with different loss functions at different training phases. It can also be applied to the problems of classification and clustering where a discrete target variable is involved.

Reinforcement Learning for Channel Coding

(Extended Abstract)

Mengke Lian*, Fabrizio Carpi†, Christian Häger‡, and Henry D. Pfister*
*Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina
†Department of Electrical and Computer Engineering, New York University, Brooklyn, New York, USA
‡Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

***Abstract*—We discuss the application of reinforcement learning (RL) to problems associated with decoding binary linear codes. In particular, we consider two different iterative decoding algorithms that involve sequential decisions and apply RL to optimize those decisions. For the first example, we consider bit-flipping (BF) decoders and observe that learned BF decoders can offer a range of performance–complexity trade-offs and achieve near-optimal performance in some cases. For the second example, we consider protograph low-density parity-check (LDPC) codes and use RL to optimize the decoding schedule. Our results show that, the comparison with the flooding schedule, a fixed error rate can be achieved with fewer update operations.**

I. Overview

The decoding of error-correcting codes can be seen as a classification problem and solved using methods introduced for supervised machine learning. The general idea is to treat the decoder as a parameterized function (e.g., a neural network) and learn good parameter configurations with data-driven optimization [1]–[4]. Without further restrictions on the code and decoder, these methods do not work well if the codes have moderate length or if they are unstructured. For linear codes, the problem simplifies considerably because one has to learn only a single decision region instead of one region per codeword. One can take advantage of linearity by using message-passing [2] or syndromes [3], [4]. Still, the problem remains challenging because good codes typically have complicated decision regions due to the large number of neighboring codewords.

This talk focuses on applications of reinforcement learning (RL) [5] to problems in channel coding. Despite impressive results in other fields, RL has yet to received significant attention in this area. In fact, with the exception [10] and recent work by the authors [6], no references were found that discuss RL for channel coding. For a general survey of RL in the general context of communications, see [11].

The unifying idea behind this work is that iterative decoding algorithms can adjust their behavior based on their current state. Thus, they can be modeled as sequential decision processes and RL can be used to optimize their behavior.

The bit-flipping (BF) decoder was introduced in [7], [8] and has been studied extensively in the literature. In [6], a subset of the authors apply RL to optimize a sequential BF decoder where, based on the syndrome, one bit is flipped in each step [6]. Rather than learning a direct mapping from observations to estimated codewords (or bits) in a supervised fashion, decoding is done in steps and the problem is mapped to a Markov decision process (MDP). Then, RL is applied to optimize the choice of which bit to flip. Following [3], [4], this approach is syndrome-based and the state space of the MDP is formed by all possible binary syndromes. This also decouples the decoding problem from the transmitted codeword. We also consider a parallel BF algorithm where the decision to flip a bit is based on the local neighborhood of that bit.

For the second example, we consider the belief-propagation decoding of protograph low-density parity-check (LDPC) codes. For these codes, the standard decoding schedule can be inefficient. In particular, windowed decoding is known to improve the performance–complexity trade-off of spatially-coupled codes [9]. Thus, we employ RL to optimize the decoding schedule and show that the optimized schedule can achieve good performance with significantly fewer operations.

In summary, we believe that RL is a promising technique for optimizing sequential decisions in decoding algorithms.

References

[1] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding," in *Proc. Annual Conf. Information Sciences and Systems (CISS)*, Baltimore, MD, 2017.

[2] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," *IEEE J. Select. Topics in Signal Processing*, vol. 12, no. 1, pp. 119–131, 2018.

[3] L. G. Tallini and P. Cull, "Neural nets for decoding error-correcting codes," in *Proc. IEEE Technical Applications Conf. and Workshops*, Portland, USA, 1995.

[4] A. Bennatan, Y. Choukroun, and P. Kisilev, "Deep learning for decoding of linear codes - a syndrome-based approach," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Vail, CO, 2018.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. A Bradford Book, 1998.

[6] F. Carpi, C. Häger, M. Martalò, R. Raheli, and H. D. Pfister, "Reinforcement learning for channel coding: Learned bit-flipping decoding," in *Proc. Annual Allerton Conf. on Commun., Control, and Comp.*, 2019.

[7] M. Bossert and F. Hergert, "Hard- and soft-decision decoding beyond the half minimum distance—an algorithm for linear codes (corresp.)," *IEEE Trans. Inf. Theory*, vol. 32, no. 5, pp. 709–714, Sept. 1986.

[8] Y. Kou, S. Lin, and M. Fossorier, "Low-density parity-check codes based on finite geometries: a rediscovery and new results," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2711–2736, Nov. 2001.

[9] A. R. Iyengar, M. Papaleo, P. H. Siegel, J. K. Wolf, A. Vanelli-Coralli, and G. E. Corazza, "Windowed decoding of protograph-based LDPC convolutional codes over erasure channels," *IEEE Trans. Inform. Theory*, vol. 58, no. 4, pp. 2303–2320, 2011.

[10] X. Wang, H. Zhang, R. Li, L. Huang, S. Dai, Y. Huangfu, and J. Wang, "Learning to flip successive cancellation decoding of polar codes with LSTM networks," *arXiv:1902.08394*, Feb. 2019.

[11] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *arxiv:1810.07862*, 2018.

Joint Source-Channel Coding of Images with (not very) Deep Learning

David Burth Kurka and Deniz Gündüz
Department of Electrical and Electronic Engineering, Imperial College London, London, UK
{d.kurka, d.gunduz}@imperial.ac.uk

***Abstract*—Almost all wireless communication systems today are designed based on essentially the same digital approach, that separately optimizes the compression and channel coding stages. Using machine learning techniques, we investigate whether end-to-end transmission can be learned from scratch, thus using joint source-channel coding (JSCC) rather than the separation approach. This paper reviews and advances recent developments on our proposed technique, *deep-JSCC*, an autoencoder-based solution for generating robust and compact codes directly from images pixels, being comparable or even superior in performance to state-of-the-art (SoA) separation-based schemes (BPG+LDPC). Additionally, we show that deep-JSCC can be expanded to exploit a series of important features, such as graceful degradation, versatility to different channels and domains, variable transmission rate through successive refinement, and its capability to exploit channel output feedback.**

I. Introduction

Wireless communication systems have traditionally followed a modular model-based design approach, in which highly specialized blocks are designed separately based on expert knowledge accumulated over decades of research. This approach is partly motivated by Shannon's *separation theorem* [1], which gives theoretical guarantees that the separate optimization of source compression and channel coding can, in the asymptotic limit, approach the optimal performance. In this way, we have available today highly specialized source codes, e.g., JPEG2000/BPG for images, MPEG-4/WMA for audio, or H.264 for video, to be used in conjunction with near-capacity-achieving channel codes, e.g., Turbo, LDPC, polar codes.

However, despite its huge impact, optimality of separation holds only under unlimited delay and complexity assumptions; and, even under these assumptions, it breaks down in multi-user scenarios [2], [3], or non-ergodic source or channel distributions [4], [5]. Moreover, unconventional communication paradigms have been emerging, demanding extreme end-to-end low latency and low power (e.g., IoT, autonomous driving, tactile Internet), and operating under more challenging environments that might not follow the traditional models (e.g., channels under bursty interference).

In light of above, our goal is to rethink the problem of wireless communication of lossy sources by using machine learning techniques, focusing particularly on image transmission. For this, we replace the modular *separation-based* design with a single neural network component for encoder and decoder (see Fig.1 for an illustrative diagram), thus performing JSCC, whose parameters are trained from data, rather than being designed. Our solution, the *deep-JSCC*, is applied to the problem of image transmission and can learn strictly from data in an unsupervised manner, as we model our system as an autoencoder [6], [7] with the communication channel incorporated as a non-trainable layer. This approach is motivated by the recent developments in machine learning through the use of deep learning (DL) techniques, and their applications to communication systems in recent years [8]. Autoencoders, in particular, due to the similarity between its architecture and digital communication systems [9], [10] have been used in related problems and pushing the boundaries of communications [11]–[16]. The use of DL for the separate problems of channel coding and image compression have been showing promising results, achieving performance in some cases superior to handcrafted algorithms [17], [18]. We show, however, that by performing JSCC, we can further improve the end-to-end performance.

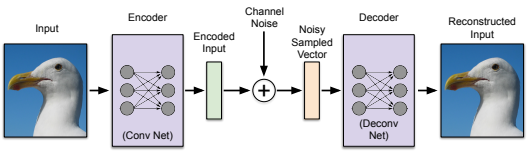


Fig. 1. Machine learning based communication system.

This paper reviews different features that were shown to be achieved with deep-JSCC, namely (a) performance comparable or superior to SoA separation-based schemes; (b) graceful degradation upon deterioration of channel conditions [19]; (c) versatility to adapt to different channels and domains [19]; (d) capacity of successive refinement [20] and (e) ability to exploit channel output feedback in order to improve the communication [21]. Thus, deep-JSCC presents itself as a powerful solution for the transmission of images, enabling communications with excellent performance while achieving low-delay and low-energy, being robust to channel changes, and allowing small and flexible bandwidth transmissions, thus advancing the field of communications by improving existing JSCC and separation-based methods.

This work was supported by the European Research Council (ERC) through project BEACON (No. 677854).

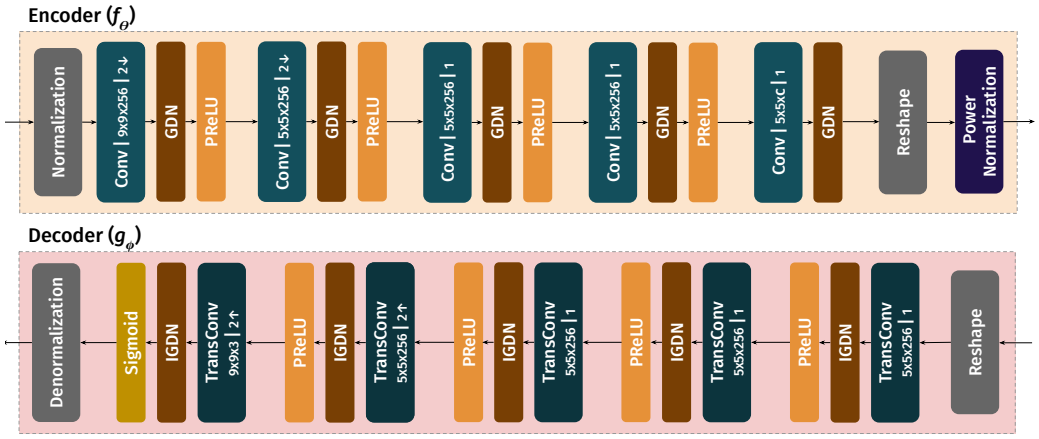


Fig. 2. Encoder and decoder architectures used in experiments.

II. Problem Formulation and Model Description

Consider an input image with height H, width W and C color channels, represented as a vector of pixel intensities $\boldsymbol{x} \in \mathcal{R}^n$; $n = H \times W \times C$ to be transmitted over k uses of a noisy channel, where k/n is the *bandwidth ratio*. An encoder $f_{\theta_i} : \mathcal{R}^n \to \mathcal{C}^{k_i}$ maps $\boldsymbol{x}$ into channel input symbols $\boldsymbol{z}_i \in \mathcal{C}_i^k$ in L blocks, where $\sum_{i=1}^{L} k_i = k$. These symbols are transmitted over a noisy channel, characterized by a random transformation $\eta : \mathcal{C}^{k_i} \to \mathcal{C}^{k_i}$, which may model physical impairments such as noise, fading or interference, resulting in the corrupted channel output $\hat{\boldsymbol{z}}_i = \eta(\boldsymbol{z}_i)$. We consider L distinct decoders, where the channel outputs for the first i blocks are decoded using $g_{\phi_i} : \mathcal{C}^{k_I} \to \mathcal{R}^n$ (where $I = \sum_{j=0}^{i} k_j$), creating reconstructions $\hat{\boldsymbol{x}}_i = g_{\phi_i}(\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_i) \in \mathcal{R}^n$, for $i \in 1, \ldots, L$.

The encoder and decoder(s) are modelled as fully convolutional networks, using generalized normalization transformations (GDN/IGDN) [22], followed by a parametric ReLU (PReLU) [23] activation function (or a sigmoid, in the last decoder block). The channel is incorporated into the model as a non-trainable layer. Fig. 2 presents the architecture and the hyperparameters used in the experiments. Before transmission, the latent vector $\boldsymbol{z}_i'$ generated at the encoder's last convolutional layer is normalized to enforce an average power constraint so that $\frac{1}{k_i}\mathbb{E}[\boldsymbol{z}_i^* \boldsymbol{z}_i] \leq P$, by setting $\boldsymbol{z}_i = \sqrt{k_i P}\frac{\boldsymbol{z}_i'}{\sqrt{\boldsymbol{z}_i'^* \boldsymbol{z}_i'}}$. The model can be optimized to minimize the average distortion between input $\boldsymbol{x}$ and its reconstructions $\hat{\boldsymbol{x}}_i$ at each layer i:

$$(\theta_i^*, \phi_i^*) = \arg\min_{\theta_i, \phi_i} \mathbb{E}_{p(\boldsymbol{x}, \hat{\boldsymbol{x}})}[d(\boldsymbol{x}, \hat{\boldsymbol{x}}_i)], \tag{1}$$

where $d(\boldsymbol{x}, \hat{\boldsymbol{x}}_i)$ is a specified distortion measure, usually the mean squared error (MSE), although other metrics are also considered. When $L > 1$, we have a multi-objective problem. However, we simplify it so that the optimization of multiple layers is done either jointly, by considering a weighted combination of losses, or greedily, by optimizing (θ_i, ϕ_i) successively. Please see [20], [21] for more details.

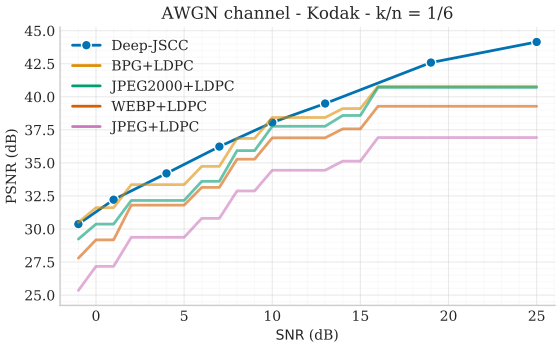


Fig. 3. Deep-JSCC performance compared to digital schemes.

III. Deep-JSCC

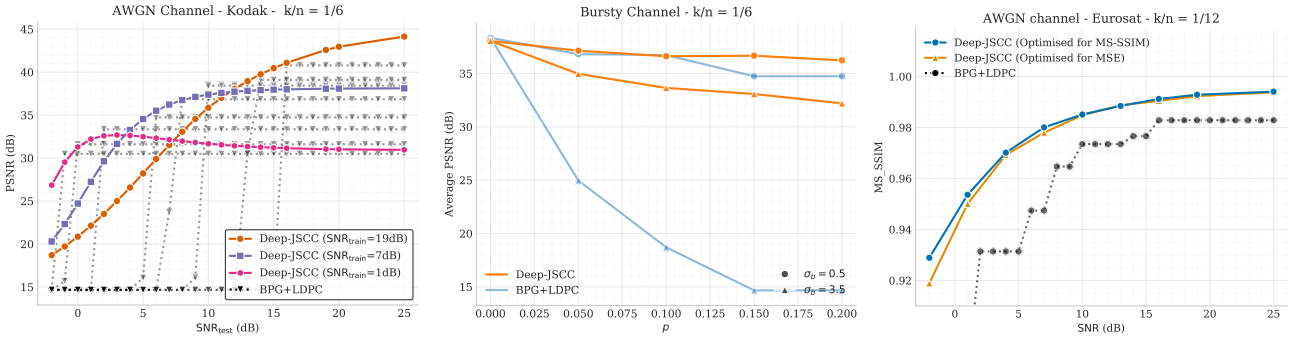
Our first set of results demonstrate the base case when, an image $\boldsymbol{x}$ is encoded by a single encoder and a single decoder, thus $L = 1$. We consider a complex AWGN channel with transfer function given by:

$$\eta_n(\boldsymbol{z}) = \boldsymbol{z} + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{n} \in \mathbb{C}^k$ is independent and identically distributed (i.i.d.) with $\boldsymbol{n} \sim \mathcal{CN}(0, \sigma^2 I)$, where σ^2 is the average noise power. We measure the quality of the channel by the average signal-to-noise ratio (SNR) given by $\text{SNR} = 10\log_{10}\frac{1}{\sigma^2}(dB)$ when $P = 1$ and the systems' performance by the peak SNR (PSNR), given by $\text{PSNR} = 10\log_{10}\frac{255^2}{||\boldsymbol{x}-\hat{\boldsymbol{x}}_i||^2}(dB)$.

Fig. 3 compares deep-JSCC with other well established codecs (BPG, JPEG2000, WebP, JPEG) followed by LDPC channel coding (see [19], [24] for more information on the experimental setup, dataset and alternative schemes considered). We see that the performance of deep-JSCC is either above or comparable to the performance of the SoA schemes, for a wide range of channel SNRs.

These results are obtained by training a different encoder/decoder model for each SNR value evaluated in the case of deep-JSCC, and considering the best performance achieved by the separation-based scheme at each SNR. In



(a) (b) (c)

Fig. 4. (a) effects of graceful degradation for deep-JSCC compared to cliff effect in separation-based scheme; (b) performance of deep-JSCC on a bursty interference channel (c) performance of deep-JSCC trained with MS-SSIM as objective function.

Fig. 4a, we experiment training models at a specific channel SNR, but evaluating it on several SNR_{test} values, also for the separation-based schemes. It can be clearly seen that deep-JSCC presents *graceful degradation*, that is, the performance gradually decreases as channel deteriorates, while the digital scheme presents a *cliff-effect* when the quality of the channel goes below the capacity for which the code was designed, losing all transmission output. Thus, we can see that deep-JSCC not only produces high performing transmissions, but also *analog behavior*, being more robust to non-ergodic channels.

A. Versatility

A big advantage of deep-JSCC being data-driven is the possibility of training for different channel models, objective functions, or specific domains. Previous work [19] show deep-JSCC is able to learn how to operate on a Rayleigh fading channel, which models variations in channel quality over time, due to physical changes in the environment. Remarkably, the model could learn to operate in a fading channel without the need of channel estimation or feedback, which are both common practice in separation-based systems.

We can also consider a channel with 'bursty' noise, which can model the presence of a high variance noise with probability p in addition to the AWGN noise $\boldsymbol{n}$, modeling in practice, an occasional random interference from a nearby transmitter. Formally, this is a Bernoulli-Gaussian noise channel with transfer function:

$$\eta_w(\boldsymbol{z}) = \boldsymbol{z} + \boldsymbol{n} + B(k,p)\boldsymbol{w}, \tag{3}$$

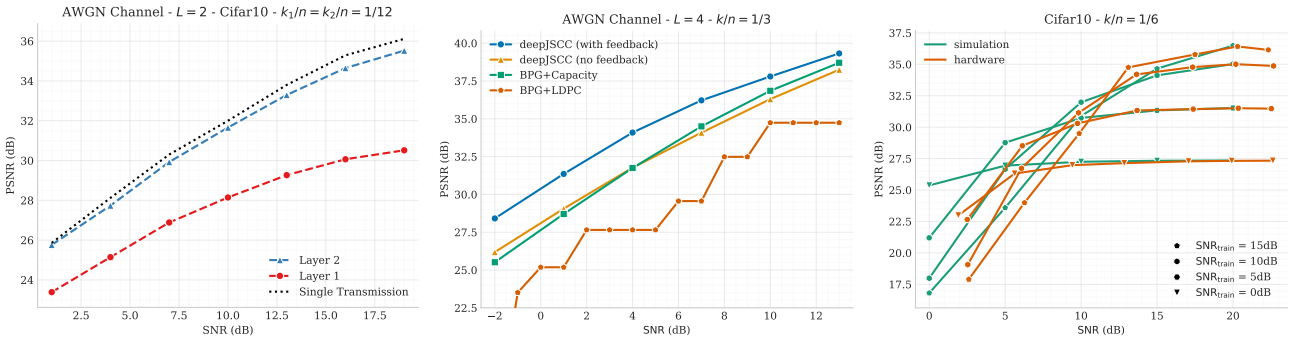
where $B(k,p)$ is the binomial distribution, and $\boldsymbol{w} \sim \mathcal{CN}(0, \sigma_b^2 I)$ the high variance noise component ($\sigma_b^2 >> 0$). Fig. 4b shows the effect of the probability p on the performance when the AWGN component's SNR is 10dB. We consider both a low-power ($\sigma_b = 0.5$) and a high-power ($\sigma_b = 3.5$) burst, and compare the performance with a digital scheme with BPG+LDPC. As expected, the performance degrades as p increases, but deep-JSCC is much more robust against the increasing power of the burst noise. A high-power burst degrades the performance of the digital scheme very quickly, even if the burst probability is very low, completely destroying the signal when $p > 0.15$. Deep-JSCC exhibits graceful degradation even in the presence of bursty noise, another important advantages in practical scenarios, particularly for communications over unlicensed bands, where occasional burst noise is common.

We also experimented training our model to a domain specific task, namely the transmission of satellite image data [25], a plausible application of our model. Here we use the distortion measure of multi-scale structural similarity (MS-SSIM) [26] – a widely accepted image quality measure that better represents human visual perception than pixel-wise differences. Our results, shown in Fig. 4c show that, when considering more specific domains, our model can better adapt to it, significantly increasing the performance gap between deep-JSCC and separation-based schemes.

B. Successive Refinement

Yet another advantage of deep-JSCC is the flexibility to adapt the transmission to different paths or stages. Consider a model with $L > 1$, in which a same image is transmitted progressively in blocks of size k_i, $i = 1, \dots L$ and $\sum_{i=1}^{L} k_i = k$. We aim to be able to reconstruct the complete image after each transmission, with increasing quality, thus performing *successive refinement* [27]–[29]. Progressive transmission can be applied to scenarios in which communication is either expensive or urgent. For example, in surveillance applications, it may be beneficial to quickly send a low-resolution image to detect a potential threat as soon as possible, while a higher resolution description can be later received for further evaluation or archival purposes. Or, in a multi-user communication setting, one could send different number of component for different users, depending on the available bandwidth.

We therefore expand our system, by creating L encoder and decoder pairs, each responsible for a partial transmission $\boldsymbol{z}_i$ and trained jointly (see [20] for implementation details and alternative architectures). Fig. 5a presents results for the case $L = 2$, for $k_1/n = k_2/n = 1/12$ and



(a) (b) (c)

Fig. 5. (a) Successive refinement with $L = 2$; (b) Layered transmission with channel output feedback, for $L = 4$; (c) Comparison between simulated and hardware performance.

shows the performance of each layer for different channel SNRs, for the AWGN channel. Results show that the loss of dividing the transmission into multiple stages is not significant; when compared to a single transmission with $k/n = 1/6$ (dotted black curve in Fig. 5a), the model performs with approximately the same quality for most channel conditions. Moreover, we observe that every layer of the layered transmission scheme preserves all features of the single transmission, such as graceful degradation and adaptability to different channel models.

C. Channel Output Feedback

Another interesting direction to be explored by deep-JSCC is the use of channel output feedback, when it is available. Suppose that alongside the *forward* communication channel considered so far, there is also a *feedback* channel, able to send back to the transmitter an estimation of the channel output $\tilde{\boldsymbol{z}}_{\boldsymbol{i}}$ after its realization. In a multi-layered transmission, this information can be used to inform subsequent layers and enhance the reconstruction at the receiver. Thus, a transmission of a source $\boldsymbol{x}$ is done sequentially in L steps, in which each step i a channel input $\boldsymbol{z}_i$ is generated from input x and feedback $\tilde{\boldsymbol{z}}_{\boldsymbol{i-1}}$ (for $i > 1$), transmitted and decoded to generate successively refined representations $\hat{\boldsymbol{x}}_{\boldsymbol{i}}$ (see [21] for specific architecture and implementation details). There has also been recent advances in the use of channel output feedback to improve the performance of channel coding [30]; however, the design is for a specific blocklength and code rate, whereas the proposed deep-JSCC scheme can transmit large content, such as images.

Fig. 5b shows the results for a scenario considering noiseless feedback (i.e. $\tilde{\boldsymbol{z}}_{\boldsymbol{i}} = \hat{\boldsymbol{z}}_{\boldsymbol{i}}$) and three uses of the feedback channel ($L = 4$), for channel inputs with size $k_i/n = 1/12$, $i = 1, \ldots, 4$. We see that by exploiting the feedback information, deep-JSCC can further increase its performance, establishing its superiority to other schemes. Note that we compare deep-JSCC with feedback with a theoretical capacity achieving channel code, and can still outperform the separation-based scheme.

This architecture enables other communication strategies, such as variable length coding, in which a minimum number of layers $\boldsymbol{z}_{\boldsymbol{i}}$ are transmitted and the quality of the reconstruction is estimated through feedback, until a target quality is achieved and the further transmission is interrupted. This scheme can provide gains of over 50% in bandwidth, when compared to separation-based approaches [21]. Further experiments also demonstrate that our model successfully operates under noisy feedback channels, and even present graceful degradation when the feedback channel changes between training and evaluation.

D. Hardware Implementation

Finally, to validate the real world performance of the proposed architecture, we implemented our basic deep-JSCC on software defined radio platform. We used models trained on the AWGN model, with different SNRs. Results can be seen in Fig. 5c and show that the simulated performance closely matches the hardware performance, especially in higher SNRs.

We also analyzed the execution time of our model. We observed that the average encoding and decoding time per image with deep-JSCC is 6.40ms on GPU, or 15.4ms on CPU, while a scheme with JPEG2000+LDPC and BPG+LDPC takes on average 4.53 and 69.9ms respectively. This shows that, although our model can be further optimized for speed, it already presents competitive times, given its outstanding performance.

IV. Conclusion

This paper reviewed and explored different features of a DL-based architecture for JSCC of images over wireless channels, the deep-JSCC. We have shown that our architecture is extremely versatile to channel models, objective functions and even transmission configurations, being able to perform multi-layered transmission and exploit channel feedback. When compared to traditional digital schemes of transmission, deep-JSCC has shown outstanding performance in different metrics and scenarios, therefore presenting itself as a viable and superior alternative, particularly for low-latency and low-power applications.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[2] ——, "Two-way communication channels," in *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, vol. 1, Berkeley, CA, 1961, pp. 611–644.

[3] D. Gündüz, E. Erkip, A. Goldsmith, and H. V. Poor, "Source and channel coding for correlated sources over multiuser channels," *IEEE Trans. on Information Theory*, vol. 55, no. 9, pp. 3927–3944, Sep. 2009.

[4] S. Vembu, S. Verdu, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 44–54, Jan 1995.

[5] D. Gunduz and E. Erkip, "Joint source-channel codes for MIMO block-fading channels," *IEEE Trans. on Information Theory*, vol. 54, no. 1, pp. 116–134, Jan 2008.

[6] Y. Bengio, "Learning deep architectures for AI," *Found. and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

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

[8] D. Gunduz, P. de Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. van der Schaar, "Machine learning in the air," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2184–2199, Oct 2019.

[9] T. J. O'Shea, K. Karra, and T. C. Clancy, "Learning to communicate: Channel auto-encoders, domain specific regularizers, and attention," in *Proc. of IEEE Int. Symp. on Signal Processing and Information Technology (ISSPIT)*, Dec. 2016, pp. 223–228.

[10] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, Dec 2017.

[11] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 119–131, Feb 2018.

[12] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," in *Proc. of Int. Conf. on Learning Representations (ICLR)*, 2018.

[13] N. Samuel, T. Diskin, and A. Wiesel, "Deep mimo detection," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2017, pp. 1–5.

[14] M. B. Mashhadi, Q. Yang, and D. Gunduz, "Cnn-based analog csi feedback in fdd mimo-ofdm systems," 2019.

[15] A. Felix, S. Cammerer, S. Dorner, J. Hoydis, and S. ten Brink, "OFDM autoencoder for end-to-end learning of communications systems," in *Proc. IEEE Int. Workshop Signal Proc. Adv. Wireless Commun. (SPAWC)*, Jun. 2018.

[16] A. Caciularu and D. Burshtein, "Blind channel equalization using variational autoencoders," in *Proc. IEEE Int. Conf. on Comms. Workshops, Kansas City, MO*, May 2018, pp. 1–6.

[17] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Turbo autoencoder: Deep learning based channel codes for point-to-point communication channels," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 2754–2764.

[18] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 10 771–10 780.

[19] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[20] D. B. Kurka and D. Gündüz, "Successive refinement of images with deep joint source-channel coding," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2019, pp. 1–5.

[21] D. B. Kurka and D. Gündüz, "Deepjscc-f: Deep joint-source channel coding of images with feedback," 2019.

[22] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," *arXiv preprint arXiv:1511.06281*, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[24] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 4774–4778.

[25] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[26] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[27] Y. Steinberg and N. Merhav, "On hierarchical joint source-channel coding," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, Jun. 2004, pp. 365–365.

[28] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[29] K. R. Sloan and S. L. Tanimoto, "Progressive refinement of raster images," *IEEE Transactions on Computers*, vol. 28, no. 11, pp. 871–874, 1979.

[30] H. Kim, Y. Jiang, S. Kannan, S. Oh, and P. Viswanath, "Deepcode: Feedback codes via deep learning," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 9436–9446.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Reinforcement Learning Technique for Finding the Feedback Capacity

(Invited Paper)

Ziv Aharoni
Ben-Gurion University of the Negev
zivah@post.bgu.ac.il

Oron Sabag
California Institute of Technology
oron@caltech.edu

Haim Henry Permuter
Ben-Gurion University of the Negev
haimp@bgu.ac.il

Abstract

One of the classic problems in information theory is solving the feedback capacity of noisy channels with memory. The capacity is expressed analytically by an optimization problem over a multi-letter objective. This is the main obstacle to directly solving the feedback capacity analytically. In the last decade, some channels with memory where solved by formulating the capacity objective as a Markov decision process, and then applying dynamic programming algorithms. However, those solutions were restricted by the channel cardinality and were computationally tractable only for channels with binary alphabet. In this paper, we propose a novel method to compute the feedback capacity of channels with memory using reinforcement learning (RL). The main advantage of this approach is its computational efficiency, even for channels with with large cardinality. The outcome of the RL algorithm sheds light on the properties of the optimal solution, which in our case, is the optimal input distribution of the channel. These insights can be converted into analytic, single-letter capacity objectives by solving corresponding lower and upper bounds. We demonstrate the efficiency of this method by analytically solving the feedback capacity of the well-known Ising channel with cardinality smaller than 9. The proposed method is used to extract the structure of the optimal input distribution, which is followed by an analytic solution for the feedback capacity and a capacity achieving coding scheme. However, we can show that the coding scheme derived for small cardinality is no longer optimal for cardinality larger or equal to 9. Insights on the solution are supplied by a new upper-bound for large cardinality. Also, we present an optimal coding scheme for asymptotic alphabet size. The proposed methodology is a step in the course of developing strong numerical tools for channels with large cardinality. Furthermore, the insights obtained by the analysis of large cardinality sheds light on the behaviour of the optimal solution as the cardinality increases.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Robust Generalization via $\alpha-$Mutual Information

Amedeo Roberto Esposito, Michael Gastpar
School of Computer and Communication Sciences
EPFL
{amedeo.esposito, michael.gastpar}@epfl.ch

Ibrahim Issa
Electrical and Computer Engineering Department
American University of Beirut
ii19@aub.edu.lb

***Abstract*—The aim of this work is to provide bounds connecting two probability measures of the same event using Rényi α-Divergences and Sibson's α-Mutual Information, a generalization of respectively the Kullback-Leibler Divergence and Shannon's Mutual Information. A particular case of interest can be found when the two probability measures considered are a joint distribution and the corresponding product of marginals (representing the statistically independent scenario). In this case a bound using Sibson's $\alpha-$Mutual Information is retrieved, extending a result involving Maximal Leakage to general alphabets. These results have broad applications, from bounding the generalization error of learning algorithms to the more general framework of adaptive data analysis, provided that the divergences and/or information measures used are amenable to such an analysis (*i.e.,* are robust to post-processing and compose adaptively). The generalization error bounds are derived with respect to high-probability events but a corresponding bound on expected generalization error is also retrieved.**

***Index Terms*—Rényi-Divergence, Sibson's Mutual Information, Maximal Leakage, Adaptive Data Analysis**

I. Introduction

Let us consider two probability spaces $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ and let $E \in \mathcal{F}$ be a measurable event. Given some divergence between the two distributions $\hat{D}(\mathcal{P}, \mathcal{Q})$ (e.g., KL, Rényi's $\alpha-$Divergence, ...) our aim is to provide bounds of the following shape:

$$\mathcal{P}(E) \leq f(\mathcal{Q}(E)) \cdot g(\hat{D}(\mathcal{P}, \mathcal{Q})), \tag{1}$$

for some functions f, g. E represents some "undesirable" event (e.g., large generalization error), whose measure under $\mathcal{Q}$ is known and whose measure under $\mathcal{P}$ we wish to bound. To that end, we use some notion of "distance" between $\mathcal{P}$ and $\mathcal{Q}$. Of particular interest is the case where $\Omega = \mathcal{X} \times \mathcal{Y}$, $\mathcal{P} = \mathcal{P}_{XY}$ (the joint distribution), and $\mathcal{Q} = \mathcal{P}_X\mathcal{P}_Y$ (product of the marginals). This allows us to bound the likelihood of $E \subseteq \mathcal{X} \times \mathcal{Y}$ when two random variables X and Y are dependent as a function of the likelihood of E when X and Y are independent (typically easier to analyze). Indeed, an immediate application can be found in bounding the generalization error of a learning algorithm and, when the proper measure is chosen, in adaptive data analysis. In order to be used in adaptive data analysis, such measure needs to be robust to post-processing and to compose adaptively (meaning that we can bound the measure between input and output of the composition of a sequence of algorithms if each of them has bounded measure). Results of this form involving mutual information can be found in [1]–[3]. More recently, a different measure satisfying these properties, maximal leakage [4], has been used in [5], [6]. More specifically, it was shown that Equation (1) holds for the following choice of $f(\mathcal{P}_X\mathcal{P}_Y(E)) = \max_y(\mathcal{P}_X(E_y))$ and $g(\hat{D}(\mathcal{P}_{XY}||\mathcal{P}_X\mathcal{P}_Y)) = \exp(\mathcal{L}\,(X\to Y)) = \mathbb{E}_Y\left(D_\infty(\mathcal{P}_{X|Y}||\mathcal{P}_X)\right) = I_\infty(X;Y)$, where $I_\infty(X;Y)$ is the Sibson mutual information of order infinity. In this work, we derive a general bound in the form of (1) and focus on two interesting special cases. In particular, one specialization of the bound leads to a family of bounds in terms of α-divergences. The other specialization leads to a family of bounds in terms of Sibson's α-mutual information, thus generalizing the previous maximal leakage bound (which corresponds to $\alpha \to \infty$).

II. Background And Definitions

A. Sibson's $\alpha-$Mutual Information

Introduced by Rényi as a generalization of entropy and KL-divergence, α-divergence has found many applications ranging from hypothesis testing to guessing and several other statistical inference problems [7]. Indeed, it has several useful operational interpretations (e.g., the number of bits by which a mixture of two codes can be compressed, the cut-off rate in block coding and hypothesis testing [8], [9] [10, p. 649]). It can be defined as follows [8].

Definition 1. Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\alpha > 0$ be a positive real number different from 1. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$ (such a measure always exists, e.g. $\mu = (\mathcal{P} + \mathcal{Q})/2$)) and denote with p, q the densities of $\mathcal{P}, \mathcal{Q}$ with respect to μ. The $\alpha-$Divergence of $\mathcal{P}$ from $\mathcal{Q}$ is defined as follows:

$$D_\alpha(\mathcal{P}\|\mathcal{Q}) = \frac{1}{\alpha - 1} \ln \int p^\alpha q^{1-\alpha} d\mu. \tag{2}$$

Remark 1. The definition is independent of the chosen measure μ. It is indeed possible to show that $\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{q}{p}\right)^{1-\alpha} d\mathcal{P}$, and that whenever $\mathcal{P} \ll \mathcal{Q}$ or $0 < \alpha < 1$, we have $\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{p}{q}\right)^{\alpha} d\mathcal{Q}$, see [8].

It can be shown that if $\alpha > 1$ and $\mathcal{P} \not\ll \mathcal{Q}$ then $D_\alpha(\mathcal{P}\|\mathcal{Q}) = \infty$. The behaviour of the measure for $\alpha \in \{0, 1, \infty\}$ can be defined by continuity. In general, one has that $D_1(\mathcal{P}\|\mathcal{Q}) = D(\mathcal{P}\|\mathcal{Q})$ but if $D(\mathcal{P}\|\mathcal{Q}) = \infty$ or there exists β such that $D_\beta(\mathcal{P}\|\mathcal{Q}) < \infty$ then $\lim_{\alpha\downarrow 1} D_\alpha(\mathcal{P}\|\|Q) = D(\mathcal{P}\|\mathcal{Q})$ [8, Theorem 5]. For an extensive treatment of α-divergences

and their properties we refer the reader to [8]. Starting from the concept of $\alpha-$divergence, Sibson built a generalization of mutual information that retains many interesting properties. The definition is the following [7]:

Definition 2. Let X and Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$, and with marginal distributions $\mathcal{P}_X$ and $\mathcal{P}_Y$, respectively. For $\alpha > 0$, the Sibson's mutual information of order α between X and Y is defined as:

$$I_\alpha(X;Y) = \min_{Q_Y} D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X Q_Y). \tag{3}$$

Moreover, $\lim_{\alpha\to 1} I_\alpha(X;Y) = I(X;Y)$. On the other hand when $\alpha \to \infty$, we get:

$$I_\infty(X;Y) = \log \mathbb{E}_{\mathcal{P}_Y}\left[\sup_{x:\mathcal{P}_X(x)>0} \frac{\mathcal{P}_{XY}(\{x,Y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{Y\})}\right].$$

For more details on Sibson's α-MI we refer the reader to [7].

B. Learning Theory

In this section, we provide some basic background knowledge on learning algorithms and concepts like generalization error. We are mainly interested in supervised learning, where the algorithm learns a *classifier* by looking at points in a proper space and the corresponding labels.

More formally, suppose we have an instance space $\mathcal{Z}$ and a hypothesis space $\mathcal{H}$. The hypothesis space is a set of functions that, given a data-point $s \in \mathcal{Z}$ give as an output the corresponding label $\mathcal{Y}$. Suppose we are given a training data set $\mathcal{Z}^n \ni S = \{z_1, \ldots, z_n\}$ made of n points sampled in an i.i.d fashion from some distribution $\mathcal{P}$. Given some $n \in \mathbb{N}$, a learning algorithm is a (possibly stochastic) mapping $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ that given as an input a finite sequence of points $S \in \mathcal{Z}^n$ outputs some classifier $h = \mathcal{A}(S) \in \mathcal{H}$. In the simplest setting we can think of $\mathcal{Z}$ as a product between the space of data-points and the space of labels, *i.e.,* $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ and suppose that $\mathcal{A}$ is fed with n data-label pairs $(x, y) \in \mathcal{Z}$. In this work we will view $\mathcal{A}$ as a family of conditional distributions $\mathcal{P}_{H|S}$ and provide a stochastic analysis of its generalization capabilities using the information measures introduced above. The goal is to generate a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that has good performance on both the training set and newly sampled points from $\mathcal{X}$. In order to ensure such property the concept of generalization error is introduced.

Definition 3. Let $\mathcal{P}$ be some distribution over $\mathcal{Z}$. Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function. The error (or risk) of a prediction rule h with respect to $\mathcal{P}$ is defined as

$$L_\mathcal{P}(h) = \mathbb{E}_{Z\sim\mathcal{P}}[\ell(h, Z)], \tag{4}$$

while, given a sample $S = (z_1, \ldots, z_n)$, the empirical error of h with respect to S is defined as

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i). \tag{5}$$

Moreover, given a learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$, its generalization error with respect to S is defined as:

$$\text{gen-err}_\mathcal{P}(\mathcal{A}, S) = |L_\mathcal{P}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))|. \tag{6}$$

The definition above considers general loss functions. An important instance for the case of supervised learning is the $0-1$ loss. Suppose again that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and that $\mathcal{H} = \{h | h : \mathcal{X} \to \mathcal{Y}\}$; given a pair $(x, y) \in \mathcal{Z}$ and a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ the loss is defined as follows:

$$\ell(h, (x, y)) = \mathbb{1}_{h(x)\neq y}, \tag{7}$$

where $\mathbb{1}$ is the indicator function. The corresponding errors become:

$$L_\mathcal{P}(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\mathbb{1}_{h(x)\neq y}] = \mathcal{P}(\{(x, y) : h(x) \neq y\}) \tag{8}$$

and

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{h(x_i)\neq y_i}. \tag{9}$$

Another fundamental concept we will need is the sample complexity of a learning algorithm.

Definition 4. Fix $\epsilon, \delta \in (0, 1)$. Let $\mathcal{H}$ be a hypothesis class. The sample complexity of $\mathcal{H}$ with respect to (ϵ, δ), denoted by $m_\mathcal{H}(\epsilon, \delta)$, is defined as the smallest $m \in \mathbb{N}$ for which there exists a learning algorithm $\mathcal{A}$ such that, for every distribution $\mathcal{P}$ over the domain $\mathcal{X}$ we have that $\mathbb{P}(\text{gen-err}_\mathcal{P}(\mathcal{A}, S) > \epsilon) \leq \delta$. If there is no such m then $m_\mathcal{H}(\epsilon, \delta) = \infty$.

For more details we refer the reader to [11].

III. Main Results

Our main theorem is a general bound on $\mathcal{P}_{XY}(E)$ in terms of $\mathcal{P}_X\mathcal{P}_Y(E)$, parameterized by two real numbers α and α'. For particular choices of α and α', we demonstrate bounds in terms of α-divergence, as well as α-mutual information. The latter is a generalization of the maximal leakage bound in [6].

Theorem 1. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Given $E \in \mathcal{F}$, let $E_y := \{x : (x, y) \in E\}$, i.e., the "fibers" of E with respect to y. Then for any $E \in \mathcal{F}$,*

$$\begin{aligned}\mathcal{P}_{XY}(E) \leq &\left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathcal{P}_X(E_Y)^{\gamma'/\gamma}\right]\right)^{1/\gamma'} \cdot \\ &\left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{\alpha'/\alpha}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right]\right)^{1/\alpha'},\end{aligned} \tag{10}$$

where $\gamma, \alpha, \gamma', \alpha'$ are such that $1 = \frac{1}{\alpha} + \frac{1}{\gamma} = \frac{1}{\alpha'} + \frac{1}{\gamma'}$, and $\alpha, \gamma, \alpha', \gamma' \geq 1$.

Proof. We have that:

$$\mathcal{P}_{XY}(E) = \mathbb{E}_{\mathcal{P}_{XY}}[\mathbb{1}_E] \tag{11}$$

$$= \mathbb{E}_{\mathcal{P}_X\mathcal{P}_Y}\left[\mathbb{1}_E \frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right] \tag{12}$$

$$= \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}\left[\mathbb{1}_{\{X\in E_Y\}}\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right]\right] \tag{13}$$

$$\leq \mathbb{E}_{\mathcal{P}_Y}\Bigg[\left(\mathbb{E}_{\mathcal{P}_X}\left[\mathbb{1}^{\gamma}_{\{X\in E_Y\}}\right]\right)^{1/\gamma}\cdot \left(\mathbb{E}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right)^{1/\alpha}\Bigg] \tag{14}$$

$$= \mathbb{E}_{\mathcal{P}_Y}\left[\mathcal{P}_X(E_Y)^{1/\gamma}\left(\mathbb{E}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right)^{1/\alpha}\right] \tag{15}$$

$$\leq \left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathcal{P}_X(E_Y)^{\gamma'/\gamma}\right]\right)^{1/\gamma'}\cdot \left(\mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{\alpha'/\alpha}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_X\mathcal{P}_Y}\right)^{\alpha}\right]\right]\right)^{1/\alpha'}, \tag{16}$$

where (14) and (16) follow from Holder's inequality, given that $\gamma, \alpha, \gamma', \alpha' \geq 1$ and $\frac{1}{\gamma}+\frac{1}{\alpha}=\frac{1}{\gamma'}+\frac{1}{\alpha'}=1$. □

Remark 2. It is clear from the proof that one can similarly bound $\mathbb{E}[g(X,Y)]$ for any positive function $g(X,Y)$ such that $g(X,Y)$ is $\mathcal{P}_X\mathcal{P}_Y$-integrable. But the shape of the bound becomes more complex as one in general does not have that $g(X,Y)^{\gamma} = g(X,Y)$ for every $\gamma \geq 1$.

Based on the choices of α, α', one can derive different bounds. Two are of particular interests to us and rely on different choices of α'. Choosing $\alpha' = \alpha$ and thus $\gamma' = \gamma$ in Theorem 1, we retrieve:

Corollary 1. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Let $E \in \mathcal{F}$ we have that:*

$$P_{XY}(E) \leq (\mathcal{P}_X\mathcal{P}_Y(E))^{\frac{\alpha-1}{\alpha}}\cdot \exp\left(\frac{\alpha-1}{\alpha}D_{\alpha}(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{P}_Y)\right). \tag{17}$$

Choosing $\alpha' \to 1$, which implies $\gamma' \to +\infty$, we retrieve:

Corollary 2. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Given $E \in \mathcal{F}$, we have that:*

$$P_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{1/\gamma}\cdot \tag{18}$$

$$\mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{1/\alpha}_{\mathcal{P}_X}\left[\left(\frac{dP_{XY}}{d\mathcal{P}_Y\mathcal{P}_X}\right)^{\alpha}\right]\right] \tag{19}$$

$$= \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{\frac{\alpha-1}{\alpha}}\exp\left(\frac{\alpha-1}{\alpha}I_{\alpha}(X;Y)\right), \tag{20}$$

where $I_{\alpha}(X;Y)$ is the Sibson's mutual information of order α [7].

Remark 3. An in-depth study of $\alpha-$mutual information appears in [7], where a slightly different notation is used. For reference, we can restate Eq. (19) in the notation of [7] to obtain:

$$P_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{1/\gamma}\cdot \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}^{1/\alpha}_{\mathcal{P}_X}\left[\left(\frac{dP_{Y|X}}{d\mathcal{P}_Y}\right)^{\alpha}\middle|Y\right]\right]. \tag{21}$$

Moreover, for a fixed α due to the property that Holder's conjugates need to satisfy, we have that $\frac{1}{\gamma}=\frac{\alpha-1}{\alpha}$ and the bound in (20) can also be rewritten as:

$$\mathcal{P}_{XY}(E) \leq \exp\left(\frac{\alpha-1}{\alpha}\left(I_{\alpha}(X;Y)+\log\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)\right). \tag{22}$$

Considering the right hand side of (22), because of the non-decreasability of Sibson's $\alpha-$Mutual Information with respect to α [7] we have that, for $1 \leq \alpha_1 \leq \alpha_2$:

$$\frac{\alpha_1-1}{\alpha_1}I_{\alpha_1}(X;Y) \leq \frac{\alpha_2-1}{\alpha_2}I_{\alpha_2}(X;Y). \tag{23}$$

Thus, choosing a smaller α yields a better dependence on $I_{\alpha}(X;Y)$ in the bound, but given that $\frac{1}{\gamma}=\frac{\alpha-1}{\alpha}$ we also have that $\frac{1}{\gamma_1} \leq \frac{1}{\gamma_2}$ and being $\operatorname{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y) \leq 1$ it implies that

$$\left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{\frac{1}{\gamma_1}} \geq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)^{\frac{1}{\gamma_2}}, \tag{24}$$

with a worse dependence on $\left(\operatorname{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}(E_Y)\right)^{\frac{1}{\gamma}}$ on the bound. This leads to a trade-off between the two quantities. If we focus on Corollary 2, letting $\alpha \to \infty$ we recover a result involving maximal leakage [5], [6], but extending it to general alphabets:

Corollary 3. *Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X}\times\mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Let $E \in \mathcal{F}$ we have that:*

$$P_{XY}(E) \leq \left(\operatorname*{ess\,sup}_{\mathcal{P}_Y}\mathcal{P}_X(E_Y)\right)\exp\left(\mathcal{L}\left(X\to Y\right)\right), \tag{25}$$

where $\mathcal{L}\left(X\to Y\right)$ is the maximal leakage [4].

The bound follows from the fact that $\mathcal{L}\left(X\to Y\right) = I_{\infty}(X;Y)$ [12]. A comparison between the bound for maximal leakage and some analogous result obtained for mutual information (through a different approach [1], [2]) can be found in [6].

IV. Applications

In this section, we consider some applications of the above bounds in the context of the generalization error. In the bounds of interest $\mathcal{P}_X(E_y)$ is typically exponentially decaying with the number of samples and the trade-off between α and γ can be explicitly seen in the sample complexity of a learning algorithm:

Corollary 4. *Let $\mathcal{X}\times\mathcal{Y}$ be the sample space and $\mathcal{H}$ be the set of hypotheses. Let $\mathcal{A}: \mathcal{X}^n\times\mathcal{Y}^n \to \mathcal{H}$ be a learning algorithm*

that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, i.e., $S \sim \mathcal{P}^n$. Let ℓ be the $0-1$ loss function as defined in (7)*. Given $\eta \in (0,1)$, let $E = \{(S,h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$. Then,*

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - 2n\eta^2\right)\right). \quad (26)$$

Proof. Fix $\eta \in (0,1)$ and $\alpha \geq 1$. Let $\frac{1}{\gamma} = \frac{\alpha-1}{\alpha}$. Let us denote with E_h the fiber of E over h for some $h \in \mathcal{H}$, *i.e.,* $E_h = \{S : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Consider $S, \hat{S} \in \{\mathcal{X} \times \mathcal{Y}\}^n$, where $S = ((x_1,y_1),\ldots,(x_n,y_n))$ and $\hat{S} = ((\hat{x}_1,\hat{y}_1),\ldots,(\hat{x}_n,\hat{y}_n))$. If $S, \hat{S}$ differ only in one position j, *i.e.,* $(x_i,y_i) = (\hat{x}_i,\hat{y}_i)\, \forall i \in [n]\backslash\{j\}$ and $(x_j,y_j) \neq (\hat{x}_j,\hat{y}_j)$ we have that for every $h \in \mathcal{H}$,

$$|L_S(h) - L_{\hat{S}}(h)| \leq \frac{1}{n}. \quad (27)$$

By McDiarmid's inequality [13][Sec. 1.1] and Ineq. (27) we have that for every hypothesis $h \in \mathcal{H}$,

$$\mathcal{P}_S(E_h) \leq 2 \cdot \exp(-2n\eta^2). \quad (28)$$

Then it follows from Corollary 2 and Ineq. (28) that:

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(S;\mathcal{A}(S))\right) \left(2\exp(-2n\eta^2)\right)^{\frac{\alpha-1}{\alpha}}. \quad (29)$$

□

Corollary 5. *Let $\mathcal{X} \times \mathcal{Y}$ be the sample space and $\mathcal{H}$ be the set of hypotheses. Let $\mathcal{A} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, i.e., $S \sim \mathcal{P}^n$. Let ℓ be the $0-1$ loss function. Given $\eta \in (0,1)$, let $E = \{(S,h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$ then, in order to ensure a confidence of $\delta \in (0,1)$, i.e., $\mathbb{P}(E) \leq \delta$, we need a number of samples m satisfying:*

$$m \geq \frac{I_\alpha(S;\mathcal{A}(S)) + \log 2 + \gamma \log\left(\frac{1}{\delta}\right)}{2\eta^2}. \quad (30)$$

Proof. From Corollary 4 we have that

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - 2n\eta^2\right)\right).$$

Fix $\delta \in (0,1)$, our aim is to have that:

$$\exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - 2n\eta^2\right)\right) \leq \delta, \quad (31)$$

solving the inequality wrt n gives us Equation (30). □

Smaller α means that $I_\alpha(S;\mathcal{A}(S))$ will be smaller, but it will imply a larger value for $\gamma = \frac{\alpha}{\alpha-1}$ and thus a worse dependency on $\log(1/\delta)$ in the sample complexity. Let $\mathcal{Z}$ be the sample space and $\mathcal{H}$ be the set of hypotheses. An immediate generalization of Corollary 4 follows by considering loss functions such that for every fixed $h \in \mathcal{H}$, the random variable $l(h,Z)$ (induced by Z) is σ^2-sub Gaussian[1] for some $\sigma > 0$.

[1]Given a random variable X we say that it is σ^2-sub-Gaussian if for every $\lambda \in \mathbb{R}$: $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$.

Corollary 6. *Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution $\mathcal{P}$ over $\mathcal{Z}$. Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function such that $\ell(h,Z)$ is σ-sub Gaussian random variable for every $h \in \mathcal{H}$. Given $\eta \in (0,1)$, let $E = \{(S,h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$. Then,*

$$\mathbb{P}(E) \leq \exp\left(\frac{1}{\gamma}\left(I_\alpha(S;\mathcal{A}(S)) + \log 2 - n\frac{\eta^2}{2\sigma^2}\right)\right). \quad (32)$$

Proof. Fix $\eta \in (0,1)$. Let us denote with E_h the fiber of E over h for some $h \in \mathcal{H}$, *i.e.,* $E_h = \{S : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. By assumption we have that $l(h,Z)$ is $\sigma-$sub Gaussian for every h. We can thus use Hoeffding's inequality for every hypothesis $h \in \mathcal{H}$, and retrieve that for every $h \in \mathcal{H}$:

$$\mathcal{P}_S(E_h) \leq 2 \cdot \exp\left(-n\frac{\eta^2}{2\sigma^2}\right). \quad (33)$$

Then it follows from Corollary 2 and Ineq. (33) that:

$$\mathbb{P}(E) \leq \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(S;\mathcal{A}(S))\right)\left(2\exp\left(-n\frac{\eta^2}{2\sigma^2}\right)\right)^{\frac{\alpha-1}{\alpha}}. \quad (34)$$

□

One important characteristic of these bounds is that they involve information-measures satisfying the data processing inequality [7]. This means that all these results about generalization are **robust** to post-processing, *i.e.,* if the outcome of any learning algorithm with bounded I_α is processed further, the value of the information measure cannot increase. Another desirable property that would render the usage of such measures appealing in Adaptive Data Analysis is the Adaptive Composition property [14]. Alas, the lack of a definition of conditional Sibson's MI does not allows us, for the moment, to fully address the issue and verify whether or not the measure composes adaptively (like Mutual Information and Maximal Leakage [2], [6]). Moreover, a comparison between this and other well-known results in the literature can be found in Table I. One can immediately see that the Sibson's MI bound and, in particular, the Maximal Leakage one, are the ones that most resemble the VC-Dimension bound both in terms of excess probability decay and sample complexity.

V. Bounds on Expected Generalization Error

So far, when analyzing the generalization error, we have only considered high probability bounds, what can these results tell us about the **expected** generalization error? In order to provide a meaningful bound, some assumptions on the quantity $\max_h \mathcal{P}_S(|L_S(h) - \mathbb{E}[L(h)]| > \eta)$ are needed (where S is a random vector of length n, sampled in an iid fashion from some distribution $\mathcal{D}$). More precisely, we will assume this probability to be exponentially decreasing with the number of samples n, as it often happens in the literature [13], [15]. The following result is inspired by [11, p. 419] with a slightly different proof.

TABLE I
COMPARISON BETWEEN BOUNDS

	Robust	Adaptive	Bound	Sample Complexity
$\beta-$Stability [15]	No	No	exp. decay in n	$f(\beta,\eta)\times\log\left(\frac{2}{\delta}\right)$
ϵ-DP [14]	Yes	Yes	$\frac{1}{4}\exp\left(\frac{-n\eta^2}{12}\right)$, $\epsilon\leq\eta/2$	$\frac{12\cdot\log(1/4\delta)}{\eta^2}$
MI [1]	Yes	Yes	$(I(X;Y)+1)/(2n\eta^2-1)$	$I(X;Y)/\eta^2\delta$
Maximal Leakage [6]	Yes	Yes	$2\cdot\exp(\mathcal{L}(X\to Y)-2n\eta^2)$	$(\mathcal{L}\left(X\to Y\right)+\log\left(\frac{2}{\delta}\right))/2\eta^2$
α-Sibson's MI	Yes	Unknown	$\exp(\frac{\alpha-1}{\alpha}(I_\alpha(S;\mathcal{A}(S))+\log 2-2n\eta^2))$	$(I_\alpha(X;Y)+\log 2+\gamma\log\left(\frac{1}{\delta}\right))/2\eta^2$
VC-Dim. K [11]			$2\cdot\exp(\log(K)-2n\eta^2)$	$(\log(K)+\log\left(\frac{2}{\delta}\right))/2\eta^2$

Lemma 1. *Let X be a random variable and let $\hat{x}\in\mathbb{R}$. Suppose that exist $a\geq 0$ and $b\geq e$ such that for every $\eta>0$ $\mathcal{P}_X(|X-\hat{x}|\geq\eta)\leq 2b\exp\left(-\eta^2/a^2\right)$ then $\mathbb{E}\left[|X-\hat{x}|\right]\leq a\left(\sqrt{\log 2b}+\frac{1}{2\sqrt{\log 2b}}\right)$.*

Proof.

$$\mathbb{E}\left[|X-\hat{x}|\right]=\int_0^{+\infty}\mathcal{P}_X(|X-\hat{x}|\geq\eta)d\eta \tag{35}$$

$$\leq\int_0^{+\infty}\min\left(1,2b\exp\left(-\eta^2/a^2\right)\right)d\eta \tag{36}$$

$$=\int_0^{\sqrt{a^2\log 2b}}d\eta+\int_{\sqrt{a^2\log 2b}}^{+\infty}2b\exp(-\frac{\eta^2}{a^2})d\eta \tag{37}$$

$$\leq a\left(\sqrt{\log 2b}+\frac{1}{2\sqrt{\log 2b}}\right). \tag{38}$$

□

Theorem 2. *Let $\mathcal{A}:\mathcal{Z}^n\to\mathcal{H}$ be a learning algorithm and let $I_\alpha(S;\mathcal{A}(S))$ be the dependence measure chosen. Suppose that the loss function $l:\mathcal{Z}\times\mathcal{H}\to\mathbb{R}$ is such that $\forall h\mathcal{P}_{S\sim\mathcal{D}^n}(|L_S(h)-\mathbb{E}[L(h)]|>\eta)\leq 2\exp\left(-\frac{\eta^2}{2\sigma^2}n\right)$ for some $\sigma>0$ (e.g. $l(h,Z)$ is σ^2-sub-Gaussian), then:*

$$\mathbb{E}\left[|L_S(H)-\mathbb{E}[L(H)]|\right]\leq \tag{39}$$

$$\sqrt{\frac{2\sigma^2\gamma}{n}}\left(\sqrt{\frac{\log(2)+I_\alpha(S;\mathcal{A}(S))}{\gamma}}+\frac{1}{2\sqrt{\frac{\log 2+I_\alpha(S;\mathcal{A}(S))}{\gamma}}}\right). \tag{40}$$

Proof. The proof is a simple application of Lemma 1 and Corollary 6 with $a=\sqrt{2\gamma\sigma^2}/\sqrt{n}$ and with $b=2^{\frac{1}{\gamma}-1}\exp\left(\frac{I_\alpha(\mathcal{A}(S);S)}{\gamma}\right)$. □

An interesting application of Theorem 2 can be found by considering $\mathcal{L}\left(S\to\mathcal{A}(S)\right)$ and the $0-1$ loss (hence, $1/4$-sub-Gaussian).

Corollary 7. *Let $\mathcal{A}:\mathcal{Z}^n\to\mathcal{H}$. Consider the $0-1$ loss, then $\forall h\mathcal{P}_{S\sim\mathcal{D}^n}(|L_S(h)-\mathbb{E}[L(h)]|>\eta)\leq 2\exp\left(-2\eta^2n\right)$, and:*

$$\mathbb{E}\left[|L_S(H)-\mathbb{E}[L(H)]|\right]\leq \tag{41}$$

$$\frac{1}{\sqrt{2n}}\left(\sqrt{\log 2+\mathcal{L}\left(S\to\mathcal{A}(S)\right)}+\frac{1}{2\sqrt{\log 2+\mathcal{L}\left(S\to\mathcal{A}(S)\right)}}\right). \tag{42}$$

REFERENCES

[1] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," ser. Proceedings of Machine Learning Research, vol. 83. PMLR, 07–09 Apr 2018, pp. 25–55.

[2] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, p. 2521–2530.

[3] 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, vol. 51. PMLR, 09–11 May 2016, pp. 1232–1240.

[4] I. Issa, S. Kamath, and A. B. Wagner, "An operational measure of information leakage," in *2016 Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 234–239.

[5] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *2019 IEEE International Symposium on Information Theory, ISIT Paris, France, July 7-12*, 2019.

[6] A. R. Esposito, M. Gastpar, and I. Issa, "Learning and adaptive data analysis via maximal leakage," in *IEEE Information Theory Workshop, ITW 2019, Visby, Gotland, Sweden, Aug 25-28*, 2019.

[7] S. Verdú, "α-mutual information," in *2015 Information Theory and Applications Workshop, ITA 2015, San Diego, CA, USA, February 1-6, 2015*, 2015, pp. 1–6.

[8] T. van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[9] I. Csiszar, "Generalized cutoff rates and rényi's information measures," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 26–34, Jan 1995.

[10] P. D. Grünwald, *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[11] S. Shalev-Shwartz and S. Ben-David., *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[12] I. Issa, A. B. Wagner, and S. Kamath, "An Operational Approach to Information Leakage," *ArXiv e-prints*, jul 2018.

[13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[14] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. Cambridge, MA, USA: MIT Pressf, 2015.

[15] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, 3 2002.

On Data-Processing and Majorization Inequalities for f-Divergences

Igal Sason
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion-Israel Institute of Technology
Haifa 32000, Israel
E-mail: sason@ee.technion.ac.il

***Abstract*—This work introduces new strong data-processing and majorization inequalities for f-divergences, and it studies some of their applications in information theory and statistics. The full paper version [16] will be published soon in the *Entropy* journal, including all proofs and further results, discussions, and information-theoretic applications. One application refers to the performance analysis of list decoding with either fixed or variable list sizes. Another application is related to a study of the quality of approximating a probability mass function, induced by the leaves of a Tunstall tree, by an equiprobable distribution. The compression rates of finite-length Tunstall codes are further analyzed for asserting their closeness to the Shannon entropy of a memoryless and stationary discrete source.**

Index Terms – Contraction coefficient, data-processing inequalities, f-divergences, hypothesis testing, list decoding, majorization, Rényi information measures, Tsallis entropy, Tunstall trees.

I. Introduction

Divergences are non-negative measures of the dissimilarity between pairs of probability measures which are defined on the same measurable space. They play a key role in the development of information theory, probability theory, statistics, learning, signal processing, and other related fields. One important class of divergence measures is defined by means of convex functions f, and it is called the class of f-divergences. It unifies fundamental and independently-introduced concepts in several branches of mathematics such as the chi-squared test for the goodness of fit in statistics, the total variation distance in functional analysis, the relative entropy in information theory and statistics, and it is also closely related to the Rényi divergence which generalizes the relative entropy. The class of f-divergences was independently introduced in the sixties by Ali and Silvey [2], and Csiszár [5]. This class satisfies pleasing features such as the data-processing inequality, convexity, continuity and duality properties, and it finds nice applications in information theory and statistics (see, e.g., [6], [7], [8], [17], [19], [20], [21]).

The full paper version of this work [16] is a research paper which is focused on the derivation of data-processing and majorization inequalities for f-divergences, and a study of some of their potential applications in information theory and statistics. Preliminaries are next provided.

II. Preliminaries

A. Preliminaries and Related Works

We provide here definitions which serve as a background to the presentation in this paper. We first provide a definition for the family of f-divergences.

Definition 1: [9, p. 4398] Let P and Q be probability measures, let μ be a dominating measure of P and Q (i.e., $P, Q \ll \mu$), and let $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q := \frac{\mathrm{d}Q}{\mathrm{d}\mu}$. The f-divergence from P to Q is given, independently of μ, by

$$D_f(P\|Q) := \int q\, f\Big(\frac{p}{q}\Big)\, \mathrm{d}\mu, \tag{1}$$

where

$$f(0) := \lim_{t \to 0^+} f(t), \tag{2}$$

$$0 f\left(\frac{0}{0}\right) := 0, \tag{3}$$

$$0 f\left(\frac{a}{0}\right) := \lim_{t \to 0^+} t f\left(\frac{a}{t}\right) = a \lim_{u \to \infty} \frac{f(u)}{u}, \quad a > 0. \tag{4}$$

Definition 2: Let Q_X be a probability distribution which is defined on a set $\mathcal{X}$, and that is not a point mass, and let $W_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$ be a stochastic transformation. The contraction coefficient for f-divergences is defined as

$$\mu_f(Q_X, W_{Y|X}) := \sup_{P_X \colon D_f(P_X\|Q_X) \in (0,\infty)} \frac{D_f(P_Y\|Q_Y)}{D_f(P_X\|Q_X)}, \tag{5}$$

where, for all $y \in \mathcal{Y}$,

$$P_Y(y) = (P_X W_{Y|X})(y) := \int_{\mathcal{X}} \mathrm{d}P_X(x)\, W_{Y|X}(y|x), \tag{6}$$

$$Q_Y(y) = (Q_X W_{Y|X})(y) := \int_{\mathcal{X}} \mathrm{d}Q_X(x)\, W_{Y|X}(y|x). \tag{7}$$

Contraction coefficients for f-divergences play a key role in strong data-processing inequalities (see [1], [12], [13]).

Definition 3: Pearson's χ^2-divergence from P to Q is defined to be the f-divergence from P to Q (see Definition 1) with $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$ for all $t > 0$,

$$\chi^2(P\|Q) := D_f(P\|Q) \tag{8}$$

$$= \int \frac{(p-q)^2}{q}\, \mathrm{d}\mu \tag{9}$$

$$= \int \frac{p^2}{q}\, \mathrm{d}\mu - 1 \tag{10}$$

independently of the dominating measure μ (i.e., $P, Q \ll \mu$, e.g., $\mu = P + Q$).

Neyman's χ^2-divergence from P to Q is the Pearson's χ^2-divergence from Q to P, i.e., it is equal to

$$\chi^2(Q\|P) = D_g(P\|Q) \tag{11}$$

with $g(t) = \frac{(t-1)^2}{t}$ or $g(t) = \frac{1}{t} - t$ for all $t > 0$.

For the presentation of our majorization inequalities for f-divergences and related entropy bounds, essential definitions and basic results are next provided (see, e.g., [11]). Let P be a probability mass function defined on a finite set $\mathcal{X}$, let $p_{\max}$ be the maximal mass of P, and let $G_P(k)$ be the sum of the k largest masses of P for $k \in \{1, \ldots, |\mathcal{X}|\}$ (hence, it follows that $G_P(1) = p_{\max}$ and $G_P(|\mathcal{X}|) = 1$).

Definition 4: Consider discrete probability mass functions P and Q defined on a finite set $\mathcal{X}$. It is said that P is majorized by Q (or Q majorizes P), and it is denoted by $P \prec Q$, if $G_P(k) \leq G_Q(k)$ for all $k \in \{1, \ldots, |\mathcal{X}|\}$ (recall that $G_P(|\mathcal{X}|) = G_Q(|\mathcal{X}|) = 1$).

A unit mass majorizes any other distribution; on the other hand, the equiprobable distribution on a finite set is majorized by any other distribution defined on the same set.

Definition 5: Let $\mathcal{P}_n$ denote the set of all the probability mass functions that are defined on $\mathcal{A}_n := \{1, \ldots, n\}$. A function $f\colon \mathcal{P}_n \to \mathbb{R}$ is said to be *Schur-convex* if for every $P, Q \in \mathcal{P}_n$ such that $P \prec Q$, we have $f(P) \leq f(Q)$. Likewise, f is said to be *Schur-concave* if $-f$ is Schur-convex, i.e., $P, Q \in \mathcal{P}_n$ and $P \prec Q$ imply that $f(P) \geq f(Q)$.

Finally, what is the connection between data processing and majorization, and why these types of inequalities are both considered in the same manuscript ? This connection is provided in the following fundamental well-known result (see, e.g., [11, Theorem B.2]):

Proposition 1: Let P and Q be probability mass functions defined on a finite set $\mathcal{A}$. Then, $P \prec Q$ if and only if there exists a doubly-stochastic transformation $W_{Y|X}\colon \mathcal{A} \to \mathcal{A}$ (i.e., $\sum\limits_{x\in\mathcal{A}} W_{Y|X}(y|x) = 1$ for all $y \in \mathcal{A}$, and $\sum\limits_{y\in\mathcal{A}} W_{Y|X}(y|x) = 1$ for all $x \in \mathcal{A}$ with $W_{Y|X}(\cdot|\cdot) \geq 0$) such that

$$Q \to W_{Y|X} \to P.$$

In other words, $P \prec Q$ if and only if in their representation as column vectors, there exists a doubly-stochastic matrix $\mathbf{W}$ (i.e., a square matrix with non-negative entries such that the sum of each column or each row in $\mathbf{W}$ is equal to 1) such that $P = \mathbf{W}Q$.

B. Contributions

This work (see the full paper version in [16]) is focused on the derivation of data-processing and majorization inequalities for f-divergences, and it applies these inequalities to information theory and statistics.

The starting point for obtaining strong data-processing inequalities in this paper relies on the derivation of lower and upper bounds on the difference $D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y)$ where (P_X, Q_X) and (P_Y, Q_Y) denote, respectively, pairs of input and output probability distributions with a given stochastic transformation $W_{Y|X}$ (i.e., $P_X \to W_{Y|X} \to P_Y$, $Q_X \to W_{Y|X} \to Q_Y$). These bounds are expressed in terms of the respective difference in the Pearson's or Neyman's χ^2-divergence, and they hold for all f-divergences (see Theorem 1).

This paper also derives majorization inequalities for f-divergences where part of these inequalities rely on the earlier data-processing inequalities (see Theorem 3). A different approach, which relies on the concept of majorization, serves to derive tight bounds on the maximal value of an f-divergence from a probability mass function P to an equiprobable distribution; the maximization is carried over all P with a fixed finite support where the ratio of their maximal to minimal probability masses does not exceed a given value (see Theorem 4). These bounds lead to accurate asymptotic results which apply to general f-divergences, and they strengthen and generalize recent results of this type with respect to the relative entropy [4], and the Rényi divergence [15].

As an application of the data-processing inequalities for f-divergences, the setup of list decoding is further studied in [16], reproducing in a unified way some known bounds on the list decoding error probability, and deriving new bounds for fixed and variable list sizes.

As an application of the majorization inequalities in this paper, we study in [16] properties of a measure which is used to quantify the quality of approximating probability mass functions, induced by the leaves of a Tunstall tree, by an equiprobable distribution. An application of majorization inequalities for the relative entropy is used to derive a sufficient condition, expressed in terms of the principal and secondary real branches of the Lambert W function, for asserting the proximity of compression rates of finite-length (lossless and variable-to-fixed) Tunstall codes to the Shannon entropy of a memoryless and stationary discrete source.

III. Main Results on f-divergences

A. Data-processing inequalities for f-divergences

Strong data-processing inequalities are provided in the following, bounding the difference $D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y)$ and ratio $\frac{D_f(P_Y\|Q_Y)}{D_f(P_X\|Q_X)}$ where (P_X, Q_X) and (P_Y, Q_Y) denote, respectively, pairs of input and output probability distributions with a given stochastic transformation.

Theorem 1: Let $\mathcal{X}$ and $\mathcal{Y}$ be finite or countably infinite sets, let P_X and Q_X be probability mass functions that are supported on $\mathcal{X}$, and let

$$\xi_1 := \inf_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [0, 1], \tag{12}$$

$$\xi_2 := \sup_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [1, \infty]. \tag{13}$$

Let $W_{Y|X}\colon \mathcal{X} \to \mathcal{Y}$ be a stochastic transformation such that for every $y \in \mathcal{Y}$, there exists $x \in \mathcal{X}$ with $W_{Y|X}(y|x) > 0$,

and let (see (6) and (7))

$$P_Y := P_X W_{Y|X}, \tag{14}$$
$$Q_Y := Q_X W_{Y|X}. \tag{15}$$

Furthermore, let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let the non-negative constant $c_f := c_f(\xi_1, \xi_2)$ satisfy

$$f'_+(v) - f'_+(u) \geq 2c_f\,(v-u), \quad \forall\, u, v \in \mathcal{I},\ u < v \tag{16}$$

where f'_+ denotes the right-side derivative of f, and

$$\mathcal{I} := \mathcal{I}(\xi_1, \xi_2) = [\xi_1, \xi_2] \cap (0, \infty). \tag{17}$$

Then,

a)

$$\begin{aligned} &D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ &\geq c_f(\xi_1,\xi_2)\left[\chi^2(P_X\|Q_X) - \chi^2(P_Y\|Q_Y)\right] &(18)\\ &\geq 0, &(19)\end{aligned}$$

where equality holds in (18) if $D_f(\cdot\|\cdot)$ is Pearson's χ^2-divergence with $c_f \equiv 1$.

b) If f is twice differentiable on $\mathcal{I}$, then the largest possible coefficient in the right side of (16) is given by

$$c_f(\xi_1,\xi_2) = \tfrac{1}{2} \inf_{t\in\mathcal{I}(\xi_1,\xi_2)} f''(t). \tag{20}$$

c) Under the assumption in Item b), the following dual inequality also holds:

$$\begin{aligned} &D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ &\geq c_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big)\left[\chi^2(Q_X\|P_X) - \chi^2(Q_Y\|P_Y)\right] &(21)\\ &\geq 0, &(22)\end{aligned}$$

where $f^*\colon (0,\infty) \to \mathbb{R}$ is the dual convex function which is given by

$$f^*(t) := t\, f\left(\frac{1}{t}\right), \quad \forall\, t > 0, \tag{23}$$

and the coefficient in the right side of (21) satisfies

$$c_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big) = \tfrac{1}{2} \inf_{t\in\mathcal{I}(\xi_1,\xi_2)} \{t^3\, f''(t)\} \tag{24}$$

with the convention that $\frac{1}{\xi_1} = \infty$ if $\xi_1 = 0$. Equality holds in (21) if $D_f(\cdot\|\cdot)$ is Neyman's χ^2-divergence (i.e., $D_f(P\|Q) := \chi^2(Q\|P)$ for all P and Q) with $c_{f^*} \equiv 1$.

d) Under the assumption in Item b), if

$$e_f(\xi_1,\xi_2) := \tfrac{1}{2} \sup_{t\in\mathcal{I}(\xi_1,\xi_2)} f''(t) < \infty, \tag{25}$$

then,

$$\begin{aligned} &D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ &\leq e_f(\xi_1,\xi_2)\left[\chi^2(P_X\|Q_X) - \chi^2(P_Y\|Q_Y)\right]. &(26)\end{aligned}$$

Furthermore,

$$\begin{aligned} &D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ &\leq e_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big)\left[\chi^2(Q_X\|P_X) - \chi^2(Q_Y\|P_Y)\right] &(27)\end{aligned}$$

where the coefficient in the right side of (27) satisfies

$$e_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big) = \tfrac{1}{2} \sup_{t\in\mathcal{I}(\xi_1,\xi_2)} \{t^3\, f''(t)\}, \tag{28}$$

which is assumed to be finite. Equalities hold in (26) and (27) if $D_f(\cdot\|\cdot)$ is Pearson's or Neyman's χ^2-divergence with $e_f \equiv 1$ or $e_{f^*} \equiv 1$, respectively.

e) The lower and upper bounds in (18), (21), (26) and (27) are locally tight. More precisely, let $\{P_X^{(n)}\}$ be a sequence of probability mass functions defined on $\mathcal{X}$ and pointwise converging to Q_X which is supported on $\mathcal{X}$, and let $P_Y^{(n)}$ and Q_Y be the probability mass functions defined on $\mathcal{Y}$ via (14) and (15) with inputs $P_X^{(n)}$ and Q_X, respectively. Suppose that

$$\lim_{n\to\infty} \inf_{x\in\mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1, \tag{29}$$
$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1. \tag{30}$$

If f has a continuous second derivative at unity, then

$$\lim_{n\to\infty} \frac{D_f(P_X^{(n)}\|Q_X) - D_f(P_Y^{(n)}\|Q_Y)}{\chi^2(P_X^{(n)}\|Q_X) - \chi^2(P_Y^{(n)}\|Q_Y)} = \tfrac{1}{2} f''(1), \tag{31}$$

$$\lim_{n\to\infty} \frac{D_f(P_X^{(n)}\|Q_X) - D_f(P_Y^{(n)}\|Q_Y)}{\chi^2(Q_X\|P_X^{(n)}) - \chi^2(Q_Y\|P_Y^{(n)})} = \tfrac{1}{2} f''(1), \tag{32}$$

and these limits indicate the local tightness of the lower and upper bounds in Items a)–d).

Proof: See [16]. ∎

In continuation to [10, Theorem 8], we next provide an upper bound on the contraction coefficient for another subclass of f-divergences. Although the first part of the next result is stated for finite or countably infinite alphabets, it is clear from its proof that it also holds in the general alphabet setting. Connections to the literature are provided in [16].

Theorem 2: Let $f\colon (0,\infty) \to \mathbb{R}$ satisfy the conditions:

- f is a convex function, differentiable at 1, $f(1) = 0$, and $f(0) := \lim_{t\to 0^+} f(t) < \infty$;
- The function $g\colon (0,\infty) \to \mathbb{R}$, defined by $g(t) := \frac{f(t)-f(0)}{t}$ for all $t > 0$, is convex.

Let

$$\kappa(\xi_1,\xi_2) := \sup_{t\in(\xi_1,1)\cup(1,\xi_2)} \frac{f(t) + f'(1)\,(1-t)}{(t-1)^2} \tag{33}$$

where, for P_X and Q_X which are non-identical probability mass functions, $\xi_1 \in [0,1)$ and $\xi_2 \in (1,\infty]$ are given in (12) and (13). Then, in the setting of (14) and (15),

$$\frac{D_f(P_Y\|Q_Y)}{D_f(P_X\|Q_X)} \leq \frac{\kappa(\xi_1,\xi_2)}{f(0) + f'(1)} \cdot \frac{\chi^2(P_Y\|Q_Y)}{\chi^2(P_X\|Q_X)}. \tag{34}$$

Consequently, if Q_X is finitely supported on $\mathcal{X}$,

$$\mu_f(Q_X, W_{Y|X}) \leq \frac{1}{f(0)+f'(1)} \cdot \kappa\left(0, \frac{1}{\min\limits_{x\in\mathcal{X}} Q_X(x)}\right) \cdot \mu_{\chi^2}(Q_X, W_{Y|X}). \tag{35}$$

Proof: See [16]. ∎

We refer the reader to a parametric subclass of f-divergences with interesting properties which is introduced in [16], and which satisfies the conditions of Theorem 2.

B. f-divergence Inequalities via Majorization

Let U_n denote an equiprobable probability mass function on $\{1,\ldots,n\}$ with $n \in \mathbb{N}$, i.e., $U_n(i) := \frac{1}{n}$ for all $i \in \{1,\ldots,n\}$. By majorization theory and Theorem 1, the next result strengthens the Schur-convexity property of the f-divergence $D_f(\cdot\|U_n)$ (see [3, Lemma 1]).

Theorem 3: Let P and Q be probability mass functions which are supported on $\{1,\ldots,n\}$, and suppose that $P \prec Q$. Let $f\colon (0,\infty) \to \mathbb{R}$ be twice differentiable and convex with $f(1)=0$, and let $q_{\max}$ and $q_{\min}$ be, respectively, the maximal and minimal positive masses of Q. Then,

a)

$$\begin{aligned} &ne_f(nq_{\min}, nq_{\max})\left(\|Q\|_2^2 - \|P\|_2^2\right) \\ &\geq D_f(Q\|U_n) - D_f(P\|U_n) \qquad (36)\\ &\geq nc_f(nq_{\min}, nq_{\max})\left(\|Q\|_2^2 - \|P\|_2^2\right) \geq 0, \qquad (37)\end{aligned}$$

where $c_f(\cdot,\cdot)$ and $e_f(\cdot,\cdot)$ are given in (20) and (25), respectively, and $\|\cdot\|_2$ denotes the Euclidean norm. Furthermore, (36) and (37) hold with equality if $D_f(\cdot\|\cdot) = \chi^2(\cdot\|\cdot)$.

b) If $P \prec Q$ and $\frac{q_{\max}}{q_{\min}} \leq \rho$ for an arbitrary $\rho \geq 1$, then

$$0 \leq \|Q\|_2^2 - \|P\|_2^2 \leq \frac{(\rho-1)^2}{4\rho n}. \tag{38}$$

Proof: See [16]. ∎

The next result provides upper and lower bounds on f-divergences from any probability mass function to an equiprobable distribution. It relies on majorization theory, and it follows in part from Theorem 3.

Theorem 4: Let $\mathcal{P}_n$ denote the set of all the probability mass functions that are defined on $\mathcal{A}_n := \{1,\ldots,n\}$. For $\rho \geq 1$, let $\mathcal{P}_n(\rho)$ be the set of all $Q \in \mathcal{P}_n$ which are supported on $\mathcal{A}_n$ with $\frac{q_{\max}}{q_{\min}} \leq \rho$, and let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1)=0$. Then,

a) The set $\mathcal{P}_n(\rho)$, for any $\rho \geq 1$, is a non-empty, convex and compact set.

b) For a given $Q \in \mathcal{P}_n$, which is supported on $\mathcal{A}_n$, the f-divergences $D_f(\cdot\|Q)$ and $D_f(Q\|\cdot)$ attain their maximal values over the set $\mathcal{P}_n(\rho)$.

c) For $\rho \geq 1$ and an integer $n \geq 2$, let

$$u_f(n,\rho) := \max_{Q\in\mathcal{P}_n(\rho)} D_f(Q\|U_n), \tag{39}$$

$$v_f(n,\rho) := \max_{Q\in\mathcal{P}_n(\rho)} D_f(U_n\|Q), \tag{40}$$

let

$$\Gamma_n(\rho) := \left[\frac{1}{1+(n-1)\rho}, \frac{1}{n}\right], \tag{41}$$

and let the probability mass function $Q_\beta \in \mathcal{P}_n(\rho)$ be defined on the set $\mathcal{A}_n$ as follows:

$$Q_\beta(j) := \begin{cases} \rho\beta, & j \in \{1,\ldots,i_\beta\},\\ 1-\big(n+i_\beta(\rho-1)-1\big)\beta, & j = i_\beta+1,\\ \beta, & i_\beta+2 \leq j \leq n \end{cases} \tag{42}$$

where

$$i_\beta := \left\lfloor \frac{1-n\beta}{(\rho-1)\beta} \right\rfloor. \tag{43}$$

Then,

$$u_f(n,\rho) = \max_{\beta\in\Gamma_n(\rho)} D_f(Q_\beta\|U_n), \tag{44}$$

$$v_f(n,\rho) = \max_{\beta\in\Gamma_n(\rho)} D_f(U_n\|Q_\beta). \tag{45}$$

d) For $\rho \geq 1$ and an integer $n \geq 2$, let the non-negative function $g_f^{(\rho)}\colon [0,1] \to \mathbb{R}_+$ be given by

$$g_f^{(\rho)}(x) := xf\left(\frac{\rho}{1+(\rho-1)x}\right) + (1-x)f\left(\frac{1}{1+(\rho-1)x}\right), \tag{46}$$

for all $x \in [0,1]$. Then,

$$\max_{m\in\{0,\ldots,n\}} g_f^{(\rho)}\left(\tfrac{m}{n}\right) \leq u_f(n,\rho) \leq \max_{x\in[0,1]} g_f^{(\rho)}(x), \tag{47}$$

$$\max_{m\in\{0,\ldots,n\}} g_{f^*}^{(\rho)}\left(\tfrac{m}{n}\right) \leq v_f(n,\rho) \leq \max_{x\in[0,1]} g_{f^*}^{(\rho)}(x) \tag{48}$$

with the convex function $f^*\colon (0,\infty) \to \mathbb{R}$ in (23).

e) The right-side inequalities in (47) and (48) are asymptotically tight ($n \to \infty$). Namely,

$$\lim_{n\to\infty} u_f(n,\rho) = \max_{x\in[0,1]} g_f^{(\rho)}(x), \tag{49}$$

$$\lim_{n\to\infty} v_f(n,\rho) = \max_{x\in[0,1]} g_{f^*}^{(\rho)}(x). \tag{50}$$

f) If $g_f^{(\rho)}(\cdot)$ in (46) is differentiable on $(0,1)$ and its derivative is upper bounded by $K_f(\rho) \geq 0$, then for every integer $n \geq 2$

$$0 \leq \lim_{n'\to\infty}\big\{u_f(n',\rho)\big\} - u_f(n,\rho) \leq \frac{K_f(\rho)}{n}. \tag{51}$$

g) Let $f(0) := \lim\limits_{t\to 0} f(t) \in (-\infty,+\infty]$, and let $n \geq 2$ be an integer. Then,

$$\lim_{\rho\to\infty} u_f(n,\rho) = \left(1-\frac{1}{n}\right) f(0) + \frac{f(n)}{n}. \tag{52}$$

Furthermore, if $f(0) < \infty$, f is differentiable on $(0, n)$, and $K_n := \sup_{t \in (0,n)} \big|f'(t)\big| < \infty$, then, for every $\rho \geq 1$,

$$0 \leq \lim_{\rho' \to \infty} \big\{u_f(n, \rho')\big\} - u_f(n, \rho) \leq \frac{2K_n\,(n-1)}{n+\rho-1}. \quad (53)$$

h) For $\rho \geq 1$, let the function f be also twice differentiable, and let M and m be constants such that the following condition holds:

$$0 \leq m \leq f''(t) \leq M, \quad \forall\, t \in \big[\tfrac{1}{\rho}, \rho\big]. \quad (54)$$

Then, for all $Q \in \mathcal{P}_n(\rho)$,

$$0 \leq \tfrac{1}{2} m \big(n\|Q\|_2^2 - 1\big) \quad (55)$$
$$\leq D_f(Q\|U_n) \quad (56)$$
$$\leq \tfrac{1}{2} M \big(n\|Q\|_2^2 - 1\big) \quad (57)$$
$$\leq \frac{M(\rho-1)^2}{8\rho} \quad (58)$$

with equalities in (56) and (57) for the χ^2 divergence (with $M = m = 2$).

i) Let $d > 0$. If $f''(t) \leq M_f \in (0, \infty)$ for all $t > 0$, then $D_f(Q\|U_n) \leq d$ for all $Q \in \mathcal{P}_n(\rho)$, if

$$\rho \leq 1 + \frac{4d}{M_f} + \sqrt{\frac{8d}{M_f} + \frac{16d^2}{M_f^2}}. \quad (59)$$

Proof: See [16]. ∎

Tsallis entropy was introduced in [18] as a generalization of the Shannon entropy (similarly to the Rényi entropy [14]), and it was applied to statistical physics in [18].

Definition 6: [18] Let P_X be a probability mass function defined on a discrete set $\mathcal{X}$. The *Tsallis entropy of order* $\alpha \in (0,1) \cup (1, \infty)$ of X, denoted by $S_\alpha(X)$ or $S_\alpha(P_X)$, is

$$S_\alpha(X) = \frac{\|P_X\|_\alpha^\alpha - 1}{1-\alpha}, \quad (60)$$

where $\|P_X\|_\alpha := \Big(\sum_{x \in \mathcal{X}} P_X^\alpha(x)\Big)^{\frac{1}{\alpha}}$. The Tsallis entropy is continuously extended at orders 0, 1, and ∞; at order 1, it coincides with the Shannon entropy in nats.

Theorem 3 enables to strengthen the Schur-concavity property of the Tsallis entropy (see [11, Theorem 13.F.3.a.]).

Theorem 5: Let P and Q be probability mass functions which are supported on a finite set, and let $P \prec Q$. Then, for all $\alpha > 0$,

a)

$$0 \leq L(\alpha, P, Q) \leq S_\alpha(P) - S_\alpha(Q) \leq U(\alpha, P, Q), \quad (61)$$

where

$$L(\alpha, P, Q) := \begin{cases} \frac{1}{2}\,\alpha q_{\max}^{\alpha-2}\,\big(\|Q\|_2^2 - \|P\|_2^2\big), & \text{if } \alpha \in (0, 2], \\ \frac{1}{2}\,\alpha q_{\min}^{\alpha-2}\,\big(\|Q\|_2^2 - \|P\|_2^2\big), & \alpha \in (2, \infty), \end{cases} \quad (62)$$

$$U(\alpha, P, Q) := \begin{cases} \frac{1}{2}\,\alpha q_{\min}^{\alpha-2}\,\big(\|Q\|_2^2 - \|P\|_2^2\big), & \text{if } \alpha \in (0, 2], \\ \frac{1}{2}\,\alpha q_{\max}^{\alpha-2}\,\big(\|Q\|_2^2 - \|P\|_2^2\big), & \alpha \in (2, \infty), \end{cases} \quad (63)$$

and the bounds in (62) and (63) are attained at $\alpha = 2$.

b)

$$\inf_{P \prec Q} \frac{S_\alpha(P) - S_\alpha(Q)}{L(\alpha, P, Q)} = \sup_{P \prec Q} \frac{S_\alpha(P) - S_\alpha(Q)}{U(\alpha, P, Q)} = 1,$$

where the inf. and sup. in (b) can be restricted to PMFs P and Q $(P \neq Q)$ supported on a binary alphabet.

REFERENCES

[1] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Annals of Probability*, vol. 4, no. 6, pp. 925–939, December 1976.

[2] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistics Society*, Series B, vol. 28, no. 1, pp. 131–142, 1966.

[3] F. Cicalese, L. Gargano and U. Vaccaro, "A note on approximation of uniform distributions from variable-to-fixed length codes," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3772–3777, August 2006.

[4] F. Cicalese, L. Gargano, and U. Vaccaro, "Bounds on the entropy of a function of a random variable and their applications," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2220–2230, April 2018.

[5] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, January 1963.

[6] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, January 1967.

[7] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematicarum Hungarica*, vol. 2, no. 1, pp. 191–213, March 1972.

[8] F. Liese and I. Vajda, *Convex Statistical Distances* (Teubner-Texte Zur Mathematik), vol. 95. Leipzig, Germany, 1987.

[9] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[10] A. Makur and L. Zheng, "Linear bounds between contraction coefficients for f-divergences," *preprint*, July 2018. [Online]. Available at https://arxiv.org/pdf/1510.01844.pdf.

[11] A. W. Marshall, I. Olkin and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, second edition, Springer, 2011.

[12] Y. Polyanskiy and Y. Wu, "Strong data processing inequalities for channels and Bayesian networks," *Convexity and Concentration*, the IMA Volumes in Mathematics and its Applications (Editors: E. Carlen, M. Madiman and E. M. Werner), vol. 161, pp. 211–249, Springer, 2017.

[13] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.

[14] A. Rényi, "On measures of entropy and information," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 547–561, University of California Press, Berkeley, California, USA, 1961.

[15] I. Sason, "Tight bounds on the Rényi entropy via majorization with applications to guessing and compression," *Entropy*, vol. 20, no. 12, paper 896, pp. 1–25, November 2018.

[16] I. Sason, "On data-processing and majorization inequalities for f-divergences with applications," *Entropy*, vol. 21, no. 10, paper 1022, pp. 1–80, October 2019.

[17] W. Stummer and I. Vajda, "On divergences of finite measures and their applicability in statistics and information theory," *Statistics*, vol. 44, no. 2, pp. 169–187, April 2010.

[18] C. Tsallis, "Possible generalization of the Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1–2, pp. 479–487, July 1988.

[19] I. Vajda, *Theory of Statistical Inference and Information*, Kluwer Academic Publishers, 1989.

[20] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," *Information Theory - New Trends and Open Problems* (Editor: G. Longo), pp. 87–123, Springer, 1975.

[21] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Transactions on Information Theory*, vol. 19, no. 3, pp. 275–283, May 1973.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Entanglement-Assisted Capacity of Quantum Channels with Side Information

Uzi Pereg

***Abstract*—Entanglement-assisted communication over a random-parameter quantum channel with either causal or non-causal channel side information (CSI) at the encoder is considered. This describes a scenario where the quantum channel depends on the quantum state of the input environment. While Bob, the decoder, has no access to this state, Alice, the transmitter, performs a sequence of projective measurements on her environment to encode her message. Dupuis [10, 9] established the entanglement-assisted capacity with non-causal CSI. Here, we establish characterization in the causal setting, and also give an alternative proof technique and further observations for the non-causal setting.**

***Index Terms*—Quantum information, Shannon theory, quantum communication, channel capacity, state information, entanglement assistance.**

I. Introduction

A fundamental task in classical information theory is to determine the ultimate transmission rate of communication. Shannon's channel coding theorem [25] states that for a given channel $p_{Y|X}$, the optimal transmission rate is the channel capacity, given by $C(p_{Y|X}) = \max_{p_X} I(X;Y)$. Various classical settings of practical significance can be described by a channel $p_{Y|X,S}$ that depends on a random parameter S when channel side information (CSI) is available at the encoder (see *e.g.* [18, 6] and references therein). The capacity with causal CSI is given by [26]

$$C_{\text{caus}}(p_{Y|X,S}) = \max_{p_T} I(T;Y) \tag{1}$$

with $X = T(S)$, where $T : \mathcal{S} \to \mathcal{X}$ is called a *Shannon strategy* (see also [18, 6]). Whereas, the capacity with non-causal CSI is given by [12]

$$C_{\text{n-c}}(p_{Y|X,S}) = \max_{p_{U,X|S}} [I(U;Y) - I(U;S)] \tag{2}$$

where U is an auxiliary random variable.

The field of quantum information is rapidly evolving in both practice and theory (see *e.g.* [8, 1, 31]). Communication through quantum channels can be separated into different categories. In particular, one may consider a setting where Alice and Bob are provided with entanglement resources [22]. The entanglement-assisted capacity for transmission of classical information over a quantum channel was fully characterized by Bennet *et al.* [2, 3]. As for classical communication without entanglement between the encoder and the decoder, the Holevo-Schumacher-Westmoreland Theorem provides an asymptotic ("multi-letter") formula for the capacity [15, 24], though calculation of such a formula is intractable in general. This is because the Holevo information is not necessarily additive [13], with some exceptions such as entanglement-breaking channels [27].

The entanglement-assisted capacity of a quantum channel with non-causal CSI was determined by Dupuis [10, 9]. Furthermore, Boche, Cai, and Nötzel [4] addressed the classical-quantum channel with CSI at the encoder without entanglement. The classical capacity was determined given causal CSI, and a multi-letter formula was provided given non-causal CSI. Warsi and Coon [28] derived multi-letter bounds for a similar setting, where the side information has a limited rate. Luo and Devetak [21] considered channel simulation with source side information (SSI) at the decoder, and also solved the quantum generalization of the Wyner-Ziv problem [30]. Quantum data compression with SSI is also studied in [7, 5], and with entanglement assistance in [19, 20].

In this paper, we consider a quantum channel with either causal or non-causal CSI. The motivation is as follows. Suppose that Alice wishes to send classical information to Bob through a (fully) quantum channel $\mathcal{N}_{SA\to B}$, where A is the transmitter system, B is the receiver system, and S is the transmitter's environment, which affects the channel as well. Furthermore, suppose that Alice performs a sequence of projective measurements of the environment system S, hence the system is projected onto a particular vector $|s\rangle$ with probability $q(s)$. Using the measurement results, Alice encodes her message and sends her transmission through the channel. Whereas, Bob, who does not have access to the measurement results, "sees" the average channel $\sum_s q(s)\mathcal{N}^{(s)}_{A\to B}$, where $\mathcal{N}^{(s)}_{A\to B}$ is the projection of the channel onto $|s\rangle$. Assuming Alice's measurement projects onto orthogonal vectors, the environment system can be thought of as a classical random parameter $S \sim q(s)$. Therefore, we treat the quantum counterpart of the models in [12] and [26], *i.e.* a random-parameter quantum channel $\mathcal{N}_{S,A\to B}$ with CSI at the encoder.

We give a full characterization of the entanglement-assisted classical capacity and quantum capacity with causal CSI, and also give an alternative proof technique and further observations for the non-causal setting. While Dupuis' analysis with non-causal CSI in is based on the decoupling approach for the transmission of qubits [10, 9], we take a more direct approach. In our analysis, we incorporate the classical binning technique [14] into the quantum packing lemma [16]. Essentially, in the achievability proof, Alice performs classical compression of the parameter sequence, and then transmits both the classical message and the compressed representation using a random phase variation of the superdense coding protocol (see *e.g.* [16,

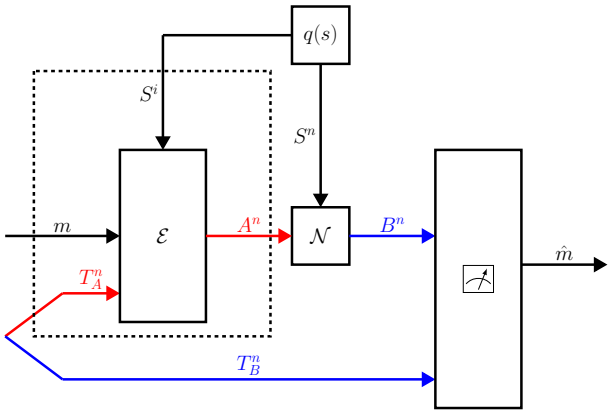


Fig. 1. Coding for a quantum channel $\mathcal{N}_{SA\to B}$ with causal side information at the encoder. The quantum systems of Alice and Bob are marked in red and blue, respectively. The systems inside the dashed-line rectangle are only available at the encoder.

29]). The results are analogous to those in the classical case, although, as usual, the quantum analysis is more involved. As observed in [12, 14], the classical optimization (2) can be restricted to mappings from (U, S) to X that are deterministic. In analogy, we observe that optimization over isometric maps suffices for our problem. With causal CSI, quantum operations are applied in a reversed order, and the Shannon strategy in (1) is replaced with a quantum channel. The full manuscript with proofs can be found in [23].

II. Definitions and Related Work

We begin with basic definitions.

A. States and Information Measures

The state of a quantum system A is given by a density operator ρ on the Hilbert space $\mathcal{H}_A$. A density operator is an Hermitian, positive semidefinite operator, with unit trace, *i.e.* $\rho^\dagger = \rho$, $\rho \succeq 0$, and $\mathrm{Tr}(\rho) = 1$. A measurement is a set of operators $\{\Lambda_j\}$ forming a positive operator-valued measure (POVM), *i.e.* $\Lambda_j \succeq 0$ and $\sum_j \Lambda_j = \mathbb{1}$, where $\mathbb{1}$ is the identity operator. The probability of the measurement outcome j is then $p_A(j) = \mathrm{Tr}(\Lambda_j \rho)$.

Define the quantum entropy of the density operator ρ as $H(\rho) \triangleq -\mathrm{Tr}[\rho \log(\rho)]$. Given a bipartite state σ_{AB}, define the quantum mutual information by

$$I(A;B)_\sigma = H(\sigma_A) + H(\sigma_B) - H(\sigma_{AB})\,. \tag{3}$$

Furthermore, define conditional quantum entropy by $H(A|B)_\sigma = H(\sigma_{AB}) - H(\sigma_B)$.

B. Quantum Channel

A random-parameter quantum channel is defined as a linear, completely positive, trace preserving map $\mathcal{N}_{SA\to B}$, corresponding to a quantum physical evolution. The channel parameter S can also be thought of as a classical system at state

$$\rho_S = \sum_{s\in\mathcal{S}} q(s)|s\rangle\langle s| \tag{4}$$

where $\{|s\rangle\}_{s\in\mathcal{S}}$ is an orthonormal basis of the Hilbert space $\mathcal{H}_S$. A quantum channel has a Kraus representation

$$\mathcal{N}_{SA\to B}(\rho) = \sum_j N_j \rho_{SA} N_j^\dagger \tag{5}$$

for all ρ_{SA}, where the operators N_j satisfy $\sum_j N_j^\dagger N_j = \mathbb{1}$. The projection on $|s\rangle$ is then given by

$$\mathcal{N}^{(s)}_{A\to B}(\rho) = \sum_j N_j^{(s)} \rho N_j^{(s)\,\dagger} \tag{6}$$

where $N_j^{(s)} \equiv \langle s|N_j|s\rangle$. A quantum channel is called isometric if it can be expressed as $\mathcal{N}_{SA\to B}(\rho) = N\rho_{SA}N^\dagger$, . where the operator N is an isometry, *i.e.* $N^\dagger N = \mathbb{1}$ [29, Section 4.6.3]. We assume that both the random parameter state and the quantum channel have a product form. That is, $\rho_{S^n} = \rho_S^{\otimes n}$, and $\mathcal{N}_{S^nA^n\to B^n} \equiv \mathcal{N}_{SA\to B}^{\otimes n}$.

C. Coding

We define a code to transmit classical information provided that the encoder and the decoder share unlimited entanglement. The entangled system pairs are denoted by (T_A^n, T_B^n).

Definition 1. A $(2^{nR}, n)$ entanglement-assisted classical code with causal CSI at the encoder consists of the following: a message set $[1 : 2^{nR}]$, where 2^{nR} is assumed to be an integer, a pure entangled state $\Psi_{T_A^n,T_B^n}$, a sequence of n encoding maps (channels) $\mathcal{E}^{m,s^i}_{T_{A,i}\to A_i}$, $m \in [1 : 2^{nR}]$, $s^i \in \mathcal{S}^i$, for $i \in [1 : n]$, and a decoding POVM $\{\Lambda^m_{B^nT_B^n}\}_{m\in[1:2^{nR}]}$. We denote the code by $(\mathcal{E}, \Psi, \Lambda)$.

The communication scheme is depicted in Figure 1. The sender Alice has the systems T_A^n, A^n and the receiver Bob has the systems T_B^n, B^n, where T_A^n and T_B^n are entangled. Alice chooses a classical message $m \in [1 : 2^{nR}]$. At time $i \in [1 : n]$, given the sequence of past and present parameters $s^i \in \mathcal{S}^i$, she applies the encoding channel $\mathcal{E}^{m,s^i}_{T_A\to A_i}$ to her share of the entangled state $\Psi_{T_{A,i},T_{B,i}}$, and then transmits the system A_i over the channel. In other words, Alice uses an encoding channel $\overline{\mathcal{E}}^{m,s^n}_{T_A^n\to A^n}$ of the following form,

$$\overline{\mathcal{E}}^{m,s^n} \triangleq \mathcal{E}^{m,s_1} \otimes \mathcal{E}^{m,s_1,s_2} \otimes \cdots \otimes \mathcal{E}^{m,s^n}\,, \tag{7}$$

and then transmits the systems A^n over n channel uses of $\mathcal{N}_{SA\to B}$.

Bob receives the channel output systems B^n, combines them with the entangled system T_B^n, and performs the POVM $\{\Lambda^m_{B^nT_B^n}\}_{m\in[1:2^{nR}]}$. The conditional probability of error, given that the message m was sent, is given by

$$P^{(n)}_{e|m}(\mathcal{E},\Psi,\Lambda) = \sum_{s^n\in\mathcal{S}^n} q^n(s^n)\mathrm{Tr}\Big[(\mathbb{1} - \Lambda^m_{B^nT_B})\,(\mathcal{N}^{(s^n)}_{A^n\to B^n}\otimes\mathbb{1})(\overline{\mathcal{E}}^{m,s^n}\otimes\mathbb{1})(\Psi_{T_A^n,T_B^n})\Big]\,. \tag{8}$$

A $(2^{nR}, n, \varepsilon)$ entanglement-assisted classical code satisfies $P_{e|m}^{(n)}(\mathcal{E}, \Psi, \Lambda) \leq \varepsilon$ for all $m \in [1 : 2^{nR}]$. A rate $R > 0$ is called achievable if for every $\varepsilon > 0$ and sufficiently large n, there exists a $(2^{nR}, n, \varepsilon)$ code. The entanglement-assisted classical capacity $\mathbb{C}_{\text{caus}}(\mathcal{N})$ is defined as the supremum of achievable rates.

The entanglement-assisted *quantum capacity* is denoted by $\mathbb{Q}_{\text{caus}}(\mathcal{N})$. We skip the definition due to lack of space.

With non-causal CSI, Alice is aware of the entire sequence S^n a priori, hence, she may use any encoding channel $\mathcal{E}_{T_A^n \to A^n}^{m,s^n}$. The entanglement-assisted classical capacity $\mathbb{C}_{\text{n-c}}(\mathcal{N})$ and quantum capacity $\mathbb{Q}_{\text{n-c}}(\mathcal{N})$ with non-causal CSI are defined accordingly.

D. Related Work

We briefly review known results for a quantum channel that does not depend on a random parameter, *i.e.* $\mathcal{N}_{A\to B}^{(s)} = \mathcal{N}_{A\to B}^{(0)}$ for $s \in \mathcal{S}$. Define

$$\mathsf{C}(\mathcal{N}^{(0)}) \triangleq \max_{|\phi\rangle_{AA'}} I(A;B)_\rho \tag{9}$$

with $\rho_{AB} \equiv (\mathbb{1} \otimes \mathcal{N}^{(0)})(|\phi\rangle\langle\phi|_{AA'})$.

Theorem 1 (see [2, 3]). The entanglement-assisted classical capacity of a quantum channel $\mathcal{N}_{A\to B}^{(0)}$ is given by

$$\mathbb{C}(\mathcal{N}^{(0)}) = \mathsf{C}(\mathcal{N}^{(0)}) \,. \tag{10}$$

Given an unlimited supply of entanglement, the teleportation protocol can send a qubit using two classical bits, while the super-dense coding protocol can send two classical bits using one qubit [22]. This implies the following.

Corollary 2 (see [2, 3]). The entanglement-assisted quantum capacity of a quantum channel $\mathcal{N}_{A\to B}^{(0)}$ is given by

$$\mathbb{Q}(\mathcal{N}^{(0)}) = \frac{1}{2}\mathsf{C}(\mathcal{N}^{(0)}) \,. \tag{11}$$

III. Main Results

We give our results on the random-parameter quantum channel $\mathcal{N}_{SA\to B}$ with causal or non-causal CSI at the encoder.

A. Causal CSI

We begin with our main result for the causal case. Define

$$\mathsf{C}_{\text{caus}}(\mathcal{N}) \triangleq \max_{\theta_{KA'},\, \mathcal{F}_{K\to A}^{(s)}} I(K;B)_\omega \tag{12}$$

where the maximization is over the quantum state $\theta_{KA'}$ and the set of quantum channels $\{\mathcal{F}_{K\to A}^{(s)}\}_{s\in\mathcal{S}}$, with

$$\omega_{AA'}^s = (\mathcal{F}^{(s)} \otimes \mathbb{1})(\theta_{KA'}) \tag{13}$$

$$\omega_{ASA'} = \sum_{s\in\mathcal{S}} q(s)|s\rangle\langle s| \otimes \omega_{AA'}^s \tag{14}$$

$$\omega_{AB} = (\mathbb{1} \otimes \mathcal{N})(\omega_{ASA'}) \,. \tag{15}$$

Before we state the capacity theorem, we give the following lemma.

Lemma 3. The maximization in (12) can be restricted to pure states $\theta_{KA'} = |\xi_{KA'}\rangle\langle\xi_{KA'}|$ and isometric channels $\mathcal{F}_{K\to A}^{(s)}(\rho_A) = F^{(s)}\rho_A F^{(s)\,\dagger}$.

The proof of Lemma 3 is given in [23], using state purification and isomeric channel extension. Now, we give our main result.

Theorem 4. The entanglement-assisted classical capacity and quantum capacity of the random-parameter quantum channel $\mathcal{N}_{SA\to B}$ with causal CSI at the encoder are given by

$$\mathbb{C}_{\text{caus}}(\mathcal{N}) = \mathsf{C}_{\text{caus}}(\mathcal{N}) \text{ and } \mathbb{Q}_{\text{caus}}(\mathcal{N}) = \frac{1}{2}\mathsf{C}_{\text{caus}}(\mathcal{N}) \tag{16}$$

respectively.

To prove achievability, we apply the random coding techniques from [2, 3] to the virtual channel $\mathcal{M}_{K\to B}$, defined by

$$\mathcal{M}(\rho_K) = \sum_{s\in\mathcal{S}} q(s)\mathcal{N}^{(s)}\left(\mathcal{F}^{(s)}(\rho_K)\right) \,. \tag{17}$$

To prove the converse part, we bound the classical *randomness-distribution* rate of a correlated pair M, M'. Using the Alicki-Fannes-Winter inequality [29], we show that $R - \varepsilon_n \leq \frac{1}{n}\sum_{i=1}^n I(K_i;B_i)_\omega \leq \max_{\theta_{KA'},\, \mathcal{F}_{K\to A}^{(s)}} I(K;B)_\omega$, with $K_i = (M, M', S^{i-1}, A'^{i-1}, T_A, T_B)$. The details are given in [23].

B. Non-Causal CSI

The entanglement-assisted capacity of a quantum channel with non-causal CSI was determined by Dupuis [10, 9]. Here, we use an alternative proof approach, which yields an equivalent formulation and further observations. Define

$$\mathsf{C}_{\text{n-c}}(\mathcal{N}) \triangleq \max_{\theta_{KA'},\, \mathcal{F}_{K\to A}^{(s)}} [I(A;B)_\omega - I(A;S)_\omega] \tag{18}$$

where the maximization is as in (13). Before we state the capacity theorem, we note that the property in Lemma 3 holds for (18) as well. Not only this property simplifies calculation of the capacity formula, but it is also useful in our proof for the theorem below.

Theorem 5 (also in [10, 9]). The entanglement-assisted classical capacity and quantum capacity of the random-parameter quantum channel $\mathcal{N}_{SA\to B}$ with non-causal CSI at the encoder are given by

$$\mathbb{C}_{\text{n-c}}(\mathcal{N}) = \mathsf{C}_{\text{n-c}}(\mathcal{N}) \text{ and } \mathbb{Q}_{\text{n-c}}(\mathcal{N}) = \frac{1}{2}\mathsf{C}_{\text{n-c}}(\mathcal{N}) \tag{19}$$

respectively.

In Section IV, we give the outline of our alternative proof for the direct part. The full proof for both the direct and converse parts is given in [23].

C. Discussion

We give a few remarks on the results above. There is clear similarity between the capacity formulas (2) and (18) given non-causal CSI. In particular, it can be seen that the classical variables U and X in (2) are replaced by the quantum systems A and A' in (18), respectively. For the classical formula (2),

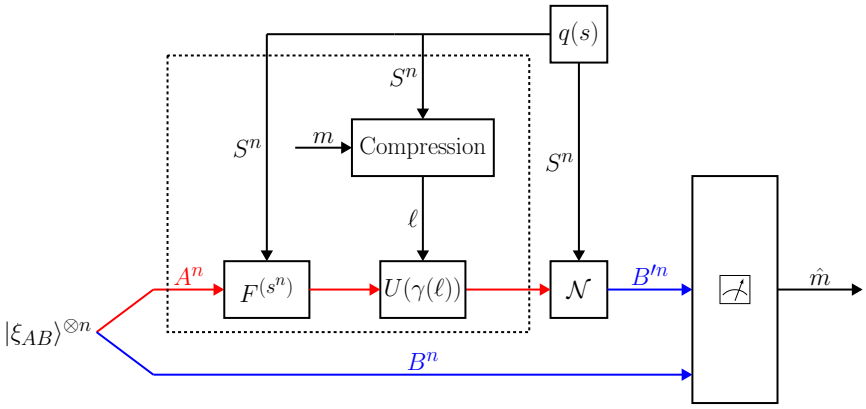


Fig. 2. Coding scheme with non-causal CSI, combining classical compression and generalized super-dense coding. The quantum systems of Alice and Bob are marked in red and blue, respectively. The blocks inside the dashed-line rectangle correspond to Alice's operations.

as shown in [12, 14], the maximization can be restricted to distributions $p_{U,X|S} = p_{U|S}p_{X|U,S}$ such that $p_{X|U,S}$ is a 0-1 probability law, based on simple convexity arguments. The property stated in Lemma 3 can thus be viewed as the quantum counterpart.

As for causal CSI, the capacity formula (1) for a classical channel can also be expressed as in (2), constrained such that U and S are statistically independent [17], and the direct part can be proved by modifying the proof for non-causal CSI accordingly. In analogy, for a quantum channel, the classical variable U is replaced with the quantum system K in (12), where K and S are in a product state. Nonetheless, we observe that in the analysis, the causality requirement also dictates that Alice applies the encoding operations in a different order.

IV. Proof Outline for Theorem 5

We present the coding scheme with non-causal CSI and describe the proof in broad strokes. The details are given in [23]. Using the non-causal counterpart of Lemma 3 (see [23, Lemma 8]), it suffices to consider a pure entangled state $|\xi_{AB}\rangle$ and a set of isometric channels, $\mathcal{F}^{(s)}_{K\to A}(\rho_K) = F^{(s)}\rho_K F^{(s)\,\dagger}$, $s \in \mathcal{S}$. Then, suppose that Alice and Bob share the joint state $|\xi_{AB}\rangle^{\otimes n}$. Define $|\varphi^s_{AB}\rangle = (F^{(s)} \otimes \mathbb{1})|\xi_{AB}\rangle$, and consider the Schmidt decomposition of the state,

$$|\varphi^s_{A,B}\rangle = \sum_{x\in\mathcal{X}} \sqrt{p_{X|S}(x|s)}|x\rangle \otimes |\psi_{x,s}\rangle \tag{20}$$

where $p_{X|S}$ is a conditional probability distribution, $\{|x\rangle\}$ is an orthonormal basis of $\mathcal{H}_A$, and $|\psi_{x,s}\rangle$ are orthonormal vectors in $\mathcal{H}_B$.

1) Code Construction: Encoding is performed in two stages, first classical compression of the parameter sequence S^n, and then, application of quantum operators depending on the result in the first stage.

(i) Classical Compression: Let $\widetilde{R} > R$. For every message $m \in [1 : 2^{nR}]$, generate a sub-codebook $\mathscr{B}(m) = \{x^n(\ell) : \ell \in [(m-1)2^{n(\widetilde{R}-R)} + 1 : m2^{n(\widetilde{R}-R)}]\}$ independently at random, with $x^n(\ell) \sim \prod_{i=1}^n p_X(x_i)$.

(ii) Quantum Operators: Using the Heisenberg-Weyl operators $\{\Sigma(a,b) = X(a)Z(b)\}$ of dimension D (see [29, Subsection 3.7.2]), we define for every $s^n \in \mathcal{S}^n$ and every conditional type class $\mathcal{T}_n(t|s^n)$ in $\mathcal{X}^n$, the operators

$$U(\gamma) = \bigoplus_t (-1)^{c_t}\Sigma(a_t, b_t)\,,$$
$$a_t, b_t \in \{0, 1, \ldots, D_t - 1\}\,,\ c_t = 0, 1\,. \tag{21}$$

with $D_t = |\mathcal{T}_n(t|s^n)|$ and $\gamma = ((a_t, b_t, c_t)_t)$. Then, choose $2^{n\widetilde{R}}$ vectors $\gamma(\ell)$, $\ell \in [1 : 2^{n\widetilde{R}}]$, uniformly at random.

2) Encoding and Decoding: The coding scheme is depicted in Figure 2. To send a message $m \in [1 : 2^{nR}]$, given a parameter sequence $s^n \in \mathcal{S}^n$, Alice performs the following.

(i) Find a sequence $x^n(\ell) \in \mathscr{B}(m)$ such that s^n and $x^n(\ell)$ are jointly typical. If there is none, choose arbitrarily.
(ii) Apply the operators $F^{(s_1)}, F^{(s_2)}, \ldots, F^{(s_n)}$, and $U(\gamma(\ell))$.
(iii) Send the systems A^n through the channel.

Bob receives the systems B'^n at state $\omega_{B'^nB^n}$ and applies a POVM. We use the quantum packing lemma [16, Lemma 2] to show that there exsits a POVM $\{\Lambda_\ell\}_{\ell\in[1:2^{n\tilde{R}}]}$ that decodes ℓ reliably, provided that $\widetilde{R} < I(B';B)_\omega - \varepsilon'$. Once Bob has a measurement result $\hat{\ell}$, he decodes the message as the corresponding sub-codebook. That is, Bob declares the message to be $\hat{m} \in [1 : 2^{nR}]$ such that $x^n(\hat{\ell}) \in \mathscr{B}(\hat{m})$.

Then, by the classical covering lemma (see *e.g.* [11, Lemma 3.3]), we have that the probability of error tends to zero as $n \to \infty$, provided that

$$R < I(B;B')_\omega - I(B;S)_\omega - \varepsilon_1\,. \tag{22}$$

Following similar analysis as in [29, Sec. 21.4], we use the ricochet property to show that Alice's unitary operations can be reflected to Bob's side. That is, there exist systems A_1, A_1', B_1 at state ω_{A_1,A_1',B_1} as in (13), and such that $I(B;B')_\omega = I(A_1;B_1)_\omega$ and $I(B;S)_\omega = I(A_1;S)_\omega$.

Acknowledgment

We gratefully thank Mark M. Wilde (Louisiana State University) for raising our attention to previous work by Dupuis [10, 9].

The work was supported by the German Federal Ministry of Education and Research (Minerva Stiftung) and the Viterbi scholarship of the Technion.

References

[1] C. H. Bennett and G. Brassard. "Quantum cryptography: public key distribution and coin tossing." *Theor. Comput. Sci.* 560.12 (2014), pp. 7–11.

[2] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal. "Entanglement-assisted classical capacity of noisy quantum channels". *Phys. Rev. Lett.* 83.15 (Oct. 1999), p. 3081.

[3] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal. "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem". *IEEE Trans. Inf. Theory* 48.10 (Oct. 2002), pp. 2637–2655.

[4] H. Boche, N. Cai, and J. Nötzel. "The classical-quantum channel with random state parameters known to the sender". *J. Physics A: Math. and Theor.* 49.19 (Apr. 2016), p. 195302.

[5] H. C. Cheng, E. P. Hanson, N. Datta, and M. H. Hsieh. "Duality between source coding with quantum side information and cq channel coding". *Proc. IEEE Int. Symp. Inf. Theory (ISIT'2019)*. Paris, France, July 2019, pp. 1142–1146.

[6] C. Choudhuri, Y. H. Kim, and U. Mitra. "Causal state communication". *IEEE Trans. Inf. Theory* 59.6 (June 2013), pp. 3709–3719.

[7] I. Devetak and A. Winter. "Classical data compression with quantum side information". *Phys. Rev. A* 68 (4 Oct. 2003), p. 042301.

[8] J. P Dowling and G. J. Milburn. "Quantum technology: the second quantum revolution". *Philos. Trans. Royal Soc. London. Series A: Math., Phys. and Eng. Sciences* 361.1809 (2003), pp. 1655–1674.

[9] F. Dupuis. "The capacity of quantum channels with side information at the transmitter". *Proc. IEEE Int. Symp. Inf. Theory (ISIT'2009)*. June 2009, pp. 948–952.

[10] Frédéric Dupuis. "Coding for quantum channels with side information at the transmitter". *arXiv preprint arXiv:0805.3352* (2008).

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

[12] S. I. Gel'fand and M. S. Pinsker. "Coding for channel with random parameters". *Probl. Control Inform. Theory* 9.1 (Jan. 1980), pp. 19–31.

[13] M. B. Hastings. "Superadditivity of communication capacity using entangled inputs". *Nature Physics* 5.4 (Mar. 2009), p. 255.

[14] C. Heegard and A. E. Gamal. "On the capacity of computer memory with defects". *IEEE Trans. Inf. Theory* 29.5 (Sept. 1983), pp. 731–739.

[15] A. S. Holevo. "The capacity of the quantum channel with general signal states". *IEEE Trans. Inf. Theory* 44.1 (Jan. 1998), pp. 269–273.

[16] M. Hsieh, I. Devetak, and A. Winter. "Entanglement-assisted capacity of quantum multiple-access channels". *IEEE Trans. Inf. Theory* 54.7 (July 2008), pp. 3078–3090.

[17] S. Jafar. "Capacity With Causal and Noncausal Side Information: A Unified View". *IEEE Trans. Inf. Theory* 52.12 (Dec. 2006), pp. 5468–5474.

[18] G. Keshet, Y. Steinberg, and N. Merhav. "Channel coding in the presence of side information". *Foundations and Trends in Communications and Information Theory* 4.6 (Jan. 2007), pp. 445–586.

[19] Z. B. Khanian and A. Winter. "Distributed compression of correlated classical-quantum sources or: the price of ignorance". *Proc. IEEE Int. Symp. Inf. Theory (ISIT'2019)*. Paris, France, July 2019, pp. 1152–1156.

[20] Z. B. Khanian and A. Winter. "Entanglement-assisted quantum data compression". *Proc. IEEE Int. Symp. Inf. Theory (ISIT'2019)*. Paris, France, July 2019, pp. 1147–1151.

[21] Z. Luo and I. Devetak. "Channel Simulation With Quantum Side Information". *IEEE Trans. Inf. Theory* 55.3 (Mar. 2009), pp. 1331–1342.

[22] M. A. Nielsen and I. Chuang. *Quantum computation and quantum information*. 2002.

[23] U. Pereg. "Entanglement-Assisted Capacity of Quantum Channels with Side Information". `arXiv:1909.09992` (Sept. 2019). URL: https://arxiv.org/pdf/1909.09992.pdf.

[24] B. Schumacher and M. D. Westmoreland. "Sending classical information via noisy quantum channels". *Phys. Rev. A* 56.1 (July 1997), p. 131.

[25] C.E. Shannon. "A mathematical theory of communication". *Bell Syst. Tech. J* 27 (July 1948), pp. 379–423, 623–656.

[26] C. E. Shannon. "Channels with side Information at the transmitter". *IBM J. Res. Dev.* 2.4 (Oct. 1958), pp. 289–293.

[27] Peter W Shor. "Additivity of the classical capacity of entanglement-breaking quantum channels". *J. Math. Phys.* 43.9 (May 2002), pp. 4334–4340.

[28] N. A. Warsi and J. P. Coon. "Coding for classical-quantum channels with rate limited side information at the encoder: information-spectrum approach". *IEEE Trans. Inf. Theory* 63.5 (May 2017), pp. 3322–3331.

[29] M. M. Wilde. *Quantum information theory*. 2nd ed. Cambridge University Press, 2017.

[30] A. Wyner and J. Ziv. "The rate-distortion function for source coding with side information at the decoder". *IEEE Trans. Inf. Theory* 22.1 (Jan. 1976), pp. 1–10.

[31] W. Zhang, D. S. Ding, Y. B. Sheng, L. Zhou, B. S. Shi, and G. C. Guo. "Quantum secure direct communication with quantum memory". *Phys. Rev. Lett.* 118.22 (2017), p. 220501.

Error Exponents of Mismatched Likelihood Ratio Testing

Parham Boroumand
University of Cambridge
pb702@cam.ac.uk

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—We study the problem of mismatched likelihood ratio test. We analyze the type-I and II error exponents when the actual distributions generating the observation are different from the distributions used in the test. We derive the worst-case error exponents when the actual distributions generating the data are within a relative entropy ball of the test distributions. In addition, we study the sensitivity of the test for small relative entropy balls.**

I. Introduction and Prelimenaries

Consider the binary hypothesis testing problem [1] where an observation $\boldsymbol{x} = (x_1, \ldots, x_n)$ is generated from two possible distributions P_1^n and P_2^n defined on the probability simplex $\mathcal{P}(\mathcal{X}^n)$. We assume that P_1^n and P_2^n are product distributions, i.e., $P_1^n(\boldsymbol{x}) = \prod_{i=1}^n P_1(x_i)$, and similarly for P_2^n. For simplicity, we assume that both $P_1(x) > 0$ and $P_2(x) > 0$ for each $x \in \mathcal{X}$.

Let $\phi : \mathcal{X}^n \to \{1, 2\}$ be a hypothesis test that decides which distribution generated the observation $\boldsymbol{x}$. We consider deterministic tests ϕ that decide in favor of P_1^n if $\boldsymbol{x} \in \mathcal{A}_1$, where $\mathcal{A}_1 \subset \mathcal{X}^n$ is the decision region for the first hypothesis. We define $\mathcal{A}_2 = \mathcal{X}^n \setminus \mathcal{A}_1$ to be the decision region for the second hypothesis. The test performance is measured by the two possible pairwise error probabilities. The type-I and type-II error probabilities are defined as

$$\epsilon_1(\phi) = \sum_{\boldsymbol{x} \in \mathcal{A}_2} P_1^n(\boldsymbol{x}), \qquad \epsilon_2(\phi) = \sum_{\boldsymbol{x} \in \mathcal{A}_1} P_2^n(\boldsymbol{x}). \tag{1}$$

A hypothesis test is said to be optimal whenever it achieves the optimal error probability tradeoff given by

$$\alpha_\beta = \min_{\phi : \epsilon_2(\phi) \leq \beta} \epsilon_1(\phi). \tag{2}$$

The likelihood ratio test defined as

$$\phi_\gamma(\boldsymbol{x}) = \mathbb{1}\left\{ \frac{P_2^n(\boldsymbol{x})}{P_1^n(\boldsymbol{x})} \geq e^{n\gamma} \right\} + 1. \tag{3}$$

was shown in [2] to attain the optimal tradeoff (2) for every γ. The type of a sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ is $\hat{T}_{\boldsymbol{x}}(a) = \frac{N(a|\boldsymbol{x})}{n}$, where $N(a|\boldsymbol{x})$ is the number of occurrences of the symbol $a \in \mathcal{X}$ in the string. The likelihood ratio test can also be expressed as a function of the type of the observation $\hat{T}_{\boldsymbol{x}}$ as [3]

$$\phi_\gamma(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{ D(\hat{T}_{\boldsymbol{x}} \| P_1) - D(\hat{T}_{\boldsymbol{x}} \| P_2) \geq \gamma \big\} + 1. \tag{4}$$

where $D(P\|Q) = \sum_{\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$ is the relative entropy between distributions P and Q.

In this paper, we are interested in the asymptotic exponential decay of the pairwise error probabilities. Therefore, it is sufficient to consider deterministic tests. The optimal error exponent tradeoff (E_1, E_2) is defined as

$$E_2(E_1) \triangleq \sup \big\{ E_2 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \quad \epsilon_1(\phi) \leq e^{-nE_1} \quad \text{and} \quad \epsilon_2(\phi) \leq e^{-nE_2} \big\}. \tag{5}$$

By using the Sanov's Theorem [3], [4], the optimal error exponent tradeoff (E_1, E_2), attained by the likelihood ratio test, can be shown to be [5], [6]

$$E_1(\phi_\gamma) = \min_{Q \in \mathcal{Q}_1(\gamma)} D(Q\|P_1), \tag{6}$$

$$E_2(\phi_\gamma) = \min_{Q \in \mathcal{Q}_2(\gamma)} D(Q\|P_2), \tag{7}$$

where

$$\mathcal{Q}_1(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_1) - D(Q\|P_2) \geq \gamma \big\}, \tag{8}$$

$$\mathcal{Q}_2(\gamma) = \big\{ Q \in \mathcal{P}(\mathcal{X}) : D(Q\|P_1) - D(Q\|P_2) \leq \gamma \big\}. \tag{9}$$

The minimizing distribution in (6), (7) is the tilted distribution

$$Q_\lambda(x) = \frac{P_1^{1-\lambda}(x) P_2^\lambda(x)}{\sum_{a \in \mathcal{X}} P_1^{1-\lambda}(a) P_2^\lambda(a)}, \quad 0 \leq \lambda \leq 1 \tag{10}$$

whenever γ satisfies $-D(P_1\|P_2) \leq \gamma \leq D(P_2\|P_1)$. In this case, λ is the solution of

$$D(Q_\lambda\|P_1) - D(Q_\lambda\|P_2) = \gamma. \tag{11}$$

Instead, if $\gamma < -D(P_1\|P_2)$, the optimal distribution in (6) is $Q_\lambda(x) = P_1(x)$ and $E_1(\phi_\gamma) = 0$, and if $\gamma > D(P_2\|P_1)$, the optimal distribution in (7) is $Q_\lambda(x) = P_2(x)$ and $E_2(\phi_\gamma) = 0$.

Equivalently, the dual expressions of (6) and (7) can be derived by substituting the minimizing distribution (10) into the Lagrangian yielding [4], [5]

$$E_1(\phi_\gamma) = \max_{\lambda \geq 0} \lambda\gamma - \log\Big(\sum_{x \in \mathcal{X}} P_1^{1-\lambda}(x) P_2^\lambda(x) \Big), \tag{12}$$

$$E_2(\phi_\gamma) = \max_{\lambda \geq 0} -\lambda\gamma - \log\Big(\sum_{x \in \mathcal{X}} P_1^{\lambda}(x) P_2^{1-\lambda}(x) \Big). \tag{13}$$

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 TEC2016-78434-C3-1-R.

The Stein regime is defined as the highest error exponent under one hypothesis when the error probability under the other hypothesis is at most some fixed $\epsilon \in (0, \frac{1}{2})$ [3]

$$E_2^{(\epsilon)} \triangleq \sup\big\{E_2 \in \mathbb{R}_+ : \exists \phi, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \quad \epsilon_1(\phi) \le \epsilon \quad \text{and} \quad \epsilon_2(\phi) \le e^{-nE_2}\big\}. \tag{14}$$

The optimal $E_2^{(\epsilon)}$, given by [3]

$$E_2^{(\epsilon)} = D(P_1\|P_2), \tag{15}$$

can be achieved by setting the threshold in (4) to be $\gamma = -D(P_1\|P_2) + \frac{C_2}{\sqrt{n}}$, where C_2 is a constant that depends on distributions P_1, P_2 and ϵ.

In this work, we revisit the above results in the case where the distributions used by the likelihood ratio test are not known precisely, and instead, fixed distributions $\hat{P}_1$ and $\hat{P}_2$ are used for testing. In particular, we find the error exponent tradeoff for fixed $\hat{P}_1$ and $\hat{P}_2$ and we study the worst-case tradeoff when the true distributions generating the observation are within a certain distance of the test distributions. The literature in robust hypothesis testing is vast (see e.g., [7]–[9] and references therein). Robust hypothesis testing consists of designing tests that are robust to the inaccuracy of the distributions generating the observation. Instead, we study the error exponent tradeoff performance of the likelihood ratio test for fixed test distributions. The proofs of our results can be found in [10].

II. Mismatched Likelihood Ratio Testing

Let $\hat{P}_1(x)$ and $\hat{P}_2(x)$ be the test distributions used in the likelihood ratio test with threshold $\hat{\gamma}$ given by

$$\hat{\phi}_{\hat{\gamma}}(\hat{T}_{\boldsymbol{x}}) = \mathbb{1}\big\{D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_1) - D(\hat{T}_{\boldsymbol{x}}\|\hat{P}_2) \ge \hat{\gamma}\big\} + 1. \tag{16}$$

For simplicity, we assume that both $\hat{P}_1(x) > 0$ and $\hat{P}_2(x) > 0$ for each $x \in \mathcal{X}$. We are interested in the achievable error exponent of the mismatched likelihood ratio test, i.e.,

$$\hat{E}_2(\hat{E}_1) \triangleq \sup\big\{\hat{E}_2 \in \mathbb{R}_+ : \exists \hat{\gamma}, \exists n_0 \in \mathbb{Z}_+ \text{ s.t. } \forall n > n_0 \quad \epsilon_1(\hat{\phi}_{\hat{\gamma}}) \le e^{-n\hat{E}_1} \text{ and } \epsilon_2(\hat{\phi}_{\hat{\gamma}}) \le e^{-n\hat{E}_2}\big\}. \tag{17}$$

Theorem 1. *For fixed $\hat{P}_1, \hat{P}_2 \in \mathcal{P}(X)$ the optimal error exponent tradeoff in* (17) *is given by*

$$\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = \min_{Q \in \hat{\mathcal{Q}}_1(\hat{\gamma})} D(Q\|P_1) \tag{18}$$

$$\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = \min_{Q \in \hat{\mathcal{Q}}_2(\hat{\gamma})} D(Q\|P_2) \tag{19}$$

where

$$\hat{\mathcal{Q}}_1(\hat{\gamma}) = \big\{Q \in \mathcal{P}(\mathcal{X}) : D(Q\|\hat{P}_1) - D(Q\|\hat{P}_2) \ge \hat{\gamma}\big\}, \tag{20}$$

$$\hat{\mathcal{Q}}_2(\hat{\gamma}) = \big\{Q \in \mathcal{P}(\mathcal{X}) : D(Q\|\hat{P}_1) - D(Q\|\hat{P}_2) \le \hat{\gamma}\big\}. \tag{21}$$

The minimizing distributions in (18) *and* (19) *are*

$$\hat{Q}_{\lambda_1}(x) = \frac{P_1(x)\hat{P}_1^{-\lambda_1}(x)\hat{P}_2^{\lambda_1}(x)}{\sum_{a\in\mathcal{X}} P_1(a)\hat{P}_1^{-\lambda_1}(a)\hat{P}_2^{\lambda_1}(a)}, \quad \lambda_1 \ge 0, \tag{22}$$

$$\hat{Q}_{\lambda_2}(x) = \frac{P_2(x)\hat{P}_2^{-\lambda_2}(x)\hat{P}_1^{\lambda_2}(x)}{\sum_{a\in\mathcal{X}} P_2(a)\hat{P}_2^{-\lambda_2}(a)\hat{P}_1^{\lambda_2}(a)}, \quad \lambda_2 \ge 0 \tag{23}$$

respectively, where λ_1 is chosen so that

$$D(\hat{Q}_{\lambda_1}\|\hat{P}_1) - D(\hat{Q}_{\lambda_1}\|\hat{P}_2) = \hat{\gamma}, \tag{24}$$

whenever $D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2) \le \hat{\gamma}$, and otherwise, $\hat{Q}_{\lambda_1}(x) = P_1(x)$ and $\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = 0$. Similarly, $\lambda_2 \ge 0$ is chosen so that

$$D(\hat{Q}_{\lambda_2}\|\hat{P}_1) - D(\hat{Q}_{\lambda_2}\|\hat{P}_2) = \hat{\gamma}, \tag{25}$$

whenever $D(P_2\|\hat{P}_1) - D(P_2\|\hat{P}_2) \ge \hat{\gamma}$, and otherwise, $\hat{Q}_{\lambda_2}(x) = P_2(x)$ and $\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = 0$. Furthermore, the dual expressions for the type-I and type-II error exponents are

$$\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = \max_{\lambda\ge 0} \lambda\hat{\gamma} - \log\Big(\sum_{x\in\mathcal{X}} P_1(x)\hat{P}_1^{-\lambda}(x)P_2^{\lambda}(x)\Big), \tag{26}$$

$$\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = \max_{\lambda\ge 0} -\lambda\hat{\gamma} - \log\Big(\sum_{x\in\mathcal{X}} P_1^{\lambda}(x)P_2(x)\hat{P}_2^{-\lambda}(x)\Big). \tag{27}$$

Remark 1*:* For mismatched likelihood ratio testing, the optimizing distributions $\hat{Q}_{\lambda_1}, \hat{Q}_{\lambda_2}$ can be different, since the decision regions only depend on the mismatched distributions. However, if $\hat{P}_1, \hat{P}_2$ are tilted with respect to P_1 and P_2, then both $\hat{Q}_{\lambda_1}, \hat{Q}_{\lambda_2}$ are also tilted respect to P_1 and P_2. This implies the result in [11], where for any set of mismatched distributions $\hat{P}_1, \hat{P}_2$ that are tilted with respect to generating distributions, the mismatched likelihood ratio test achieves the optimal error exponent tradeoff in (5).

Theorem 2. *In the Stein regime, the mismatched likelihood ratio test achieves*

$$\hat{E}_2^{(\epsilon)} = \min_{Q\in\hat{\mathcal{Q}}_2(\hat{\gamma})} D(Q\|P_2), \tag{28}$$

with threshold

$$\hat{\gamma} = D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2) + \frac{\hat{C}_2}{\sqrt{n}}, \tag{29}$$

and $\hat{C}_2$ is a constant that depends on distributions $P_1, \hat{P}_1, \hat{P}_2$, and ϵ.

Remark 2*:* Note that since P_1 satisfies the constraint in (28) then $\hat{E}_2^{(\epsilon)} \le E_2^{(\epsilon)}$. In fact, if $\hat{P}_1, \hat{P}_2$ are tilted respect to P_1, P_2 then this inequality is met with equality. Moreover, it is easy to find a set of data and test distributions where $\hat{E}_2^{(\epsilon)} < E_2^{(\epsilon)}$.

III. Mismatched Likelihood Ratio Testing with Uncertainty

In this section, we analyze the worst-case error exponents tradeoff when the actual distributions P_1, P_2 are close to the mismatched test distributions $\hat{P}_1$ and $\hat{P}_2$. More specifically,

$$P_1 \in \mathcal{B}(\hat{P}_1, R_1), \quad P_2 \in \mathcal{B}(\hat{P}_2, R_2) \tag{30}$$

where the D-ball

$$\mathcal{B}(Q, R) = \big\{P \in \mathcal{P}(\mathcal{X}) : D(Q\|P) \le R\big\} \tag{31}$$

is a ball centered at distribution Q containing all distributions whose relative entropy is smaller or equal than radius R. This model was used in robust hypothesis testing in [12]. Figure 1 depicts the mismatched probability distributions and the mismatched likelihood ratio test as a hyperplane dividing the probability space into the two decision regions.

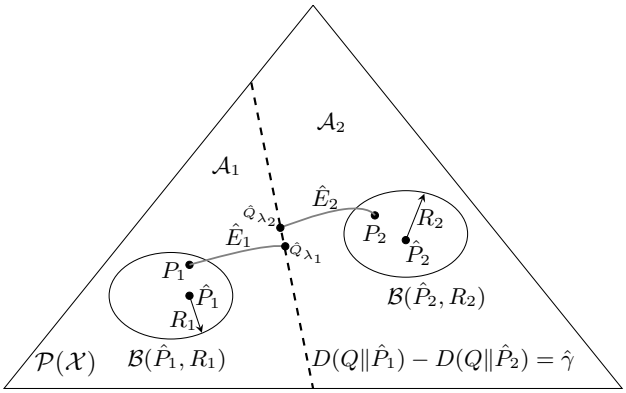


Fig. 1. Mismatched likelihood ratio test over distributions in D-balls.

We study the worst-case error-exponent performance of mismatched likelihood ratio testing when the distributions generating the observation fulfill (30). In particular, we are interested in the least favorable distributions P_1^L, P_2^L in $\mathcal{B}(\hat{P}_1, R_1), \mathcal{B}(\hat{P}_2, R_2)$, i.e., the distributions achieving the lowest error exponents $\hat{E}_1^L(R_1), \hat{E}_2^L(R_2)$.

Theorem 3. *For every $R_1, R_2 \geq 0$ let the least favorable exponents $\hat{E}_1^L(R_1), \hat{E}_2^L(R_2)$ defined as*

$$\hat{E}_1^L(R_1) = \min_{P_1 \in \mathcal{B}(\hat{P}_1, R_1)} \min_{Q \in \hat{\mathcal{Q}}_1(\hat{\gamma})} D(Q\|P_1), \tag{32}$$

$$\hat{E}_2^L(R_2) = \min_{P_2 \in \mathcal{B}(\hat{P}_2, R_2)} \min_{Q \in \hat{\mathcal{Q}}_2(\hat{\gamma})} D(Q\|P_2), \tag{33}$$

where $\hat{\mathcal{Q}}_1(\hat{\gamma}), \hat{\mathcal{Q}}_2(\hat{\gamma})$ are defined in (20), (21). *Then, for any distribution pair $P_1 \in \mathcal{B}(\hat{P}_1, R_1), P_2 \in \mathcal{B}(\hat{P}_2, R_2)$, the corresponding error exponent pair $(\hat{E}_1, \hat{E}_2)$ satisfies*

$$\hat{E}_1^L(R_1) \leq \hat{E}_1(\hat{\phi}_{\hat{\gamma}}), \quad \hat{E}_2^L(R_2) \leq \hat{E}_2(\hat{\phi}_{\hat{\gamma}}). \tag{34}$$

Furthermore, the optimization problem in (32) *is convex with optimizing distributions*

$$Q_{\lambda_1}^L(x) = \frac{P_1^L(x)\hat{P}_1^{-\lambda_1}(x)\hat{P}_2^{\lambda_1}(x)}{\sum_{a\in\mathcal{X}} P_1^L(a)\hat{P}_1^{-\lambda_1}(a)\hat{P}_2^{\lambda_1}(a)}, \tag{35}$$

$$P_1^L(x) = \beta_1 Q_{\lambda_1}^L(x) + (1-\beta_1)\hat{P}_1(x), \tag{36}$$

where $\lambda_1 \geq 0, 0 \leq \beta_1 \leq 1$ are chosen such that

$$D(Q_{\lambda_1}^L\|\hat{P}_1) - D(Q_{\lambda_1}^L\|\hat{P}_2) = \hat{\gamma}, \tag{37}$$

$$D(\hat{P}_1\|P_1^L) = R_1, \tag{38}$$

when

$$\max_{P_1 \in \mathcal{B}(\hat{P}_1, R_1)} D(P_1\|\hat{P}_1) - D(P_1\|\hat{P}_2) \leq \hat{\gamma}. \tag{39}$$

Otherwise, we can find a least favorable distribution $P_1^L \in \mathcal{B}(\hat{P}_1, R_1)$ such that $\hat{E}_1(\hat{\phi}_{\hat{\gamma}})$ for this distribution is $\hat{E}_1(\hat{\phi}_{\hat{\gamma}}) = 0$. Similarly, the optimization (33) *is convex with optimizing distributions*

$$Q_{\lambda_2}^L(x) = \frac{P_2^L(x)\hat{P}_2^{-\lambda_2}(x)\hat{P}_1^{\lambda_2}(x)}{\sum_{a\in\mathcal{X}} P_2^L(a)\hat{P}_2^{-\lambda_2}(a)\hat{P}_1^{\lambda_2}(a)}, \tag{40}$$

$$P_2^L(x) = \beta_2 Q_{\lambda_2}^L(x) + (1-\beta_2)\hat{P}_2(x), \tag{41}$$

where $\lambda_2 \geq 0, 0 \leq \beta_2 \leq 1$ are chosen such that

$$D(Q_{\lambda_2}^L\|\hat{P}_2) - DQ_{\lambda_2}^L\|\hat{P}_1) = \hat{\gamma}, \tag{42}$$

$$D(\hat{P}_2\|P_2^L) = R_2, \tag{43}$$

whenever,

$$\min_{P_2 \in \mathcal{B}(\hat{P}_2, R_2)} D(P_2\|\hat{P}_1) - D(P_2\|\hat{P}_2) \geq \hat{\gamma}. \tag{44}$$

Otherwise, we can find a distribution $P_2^L \in \mathcal{B}(\hat{P}_2, R_2)$ such that $\hat{E}_2(\hat{\phi}_{\hat{\gamma}})$ for this distribution is $\hat{E}_2(\hat{\phi}_{\hat{\gamma}}) = 0$.

The worst-case achievable error exponents of mismatched likelihood ratio testing for data distributions in a D-ball are essentially the minimum relative entropy between two sets of probability distributions. Specifically, the minimum relative entropy $\mathcal{B}(\hat{P}_1, R_1)$ and $\hat{\mathcal{Q}}_2(\hat{\gamma})$ gives $\hat{E}_1^L(R_1)$, and similarly for $\hat{E}_2^L(R_2)$.

IV. Mismatched Likelihood Ratio Testing Sensitivity

In this section, we study how the worst-case error exponents $(\hat{E}_1^L, \hat{E}_2^L)$ behave when the D-ball radii R_1, R_2 are small. In particular, we derive a Taylor series expansion of the worst-case error exponent. This approximation can also be interpreted as the worst-case sensitivity of the test, i.e., how does the test perform when actual distributions are very close to the mismatched distributions.

Theorem 4. *For every $R_i \geq 0$, $\hat{P}_i \in \mathcal{P}(\mathcal{X})$ for $i = 1, 2$, and*

$$-D(\hat{P}_1\|\hat{P}_2) \leq \hat{\gamma} \leq D(\hat{P}_2\|\hat{P}_1), \tag{45}$$

we have

$$\hat{E}_i^L(R_i) = E_i(\hat{\phi}_{\hat{\gamma}}) - S_i(\hat{P}_1, \hat{P}_2, \hat{\gamma})\sqrt{R_i} + o\big(\sqrt{R_i}\big), \tag{46}$$

where

$$S_i^2(\hat{P}_1, \hat{P}_2, \hat{\gamma}) = 2Var_{\hat{P}_i}\left(\frac{\hat{Q}_\lambda(X)}{\hat{P}_i(X)}\right) \tag{47}$$

and $\hat{Q}_\lambda(X)$ is the minimizing distribution in (10) *for test $\hat{\phi}_{\hat{\gamma}}$.*

Lemma 5. *For every $\hat{P}_1, \hat{P}_2 \in \mathcal{P}(\mathcal{X})$, and $\hat{\gamma}$ satisfying* (45)

$$\frac{\partial}{\partial\hat{\gamma}} S_1(\hat{P}_1, \hat{P}_2, \hat{\gamma}) \geq 0, \quad \frac{\partial}{\partial\hat{\gamma}} S_2(\hat{P}_1, \hat{P}_2, \hat{\gamma}) \leq 0. \tag{48}$$

This lemma shows that $S_1(\hat{P}_1, \hat{P}_2, \hat{\gamma})$ is a non-decreasing function of $\hat{\gamma}$, i.e., as $\hat{\gamma}$ increases from $-D(\hat{P}_1\|\hat{P}_2)$ to $D(\hat{P}_2\|\hat{P}_1)$, the worst-case exponent $\hat{E}_1^L(R_1)$ becomes more sensitive to mismatch with likelihood ratio testing. Conversely, $S_2(\hat{P}_1, \hat{P}_2, \hat{\gamma})$ is a non-increasing function of $\hat{\gamma}$, i.e., as $\hat{\gamma}$

increases from $-D(\hat{P}_1\|\hat{P}_2)$ to $D(\hat{P}_2\|\hat{P}_1)$, the worst-case exponent $\hat{E}_2^L(R_2)$ becomes less sensitive (more robust) to mismatch with likelihood ratio testing. Moreover, when $\lambda = \frac{1}{2}$, we have

$$\hat{Q}_{\frac{1}{2}}(x) = \frac{\sqrt{\hat{P}_1(x)\hat{P}_2(x)}}{\sum_{a\in\mathcal{X}}\sqrt{\hat{P}_1(a)\hat{P}_2(a)}}, \tag{49}$$

and then $S_1(\hat{P}_1, \hat{P}_2, \hat{\gamma}) = S_2(\hat{P}_1, \hat{P}_2, \hat{\gamma})$. In addition, $\hat{Q}_{\frac{1}{2}}$ minimizes $E_1(\hat{\phi}_{\hat{\gamma}}) + E_2(\hat{\phi}_{\hat{\gamma}})$ yielding [13]

$$E_1(\hat{\phi}_{\hat{\gamma}}) + E_2(\hat{\phi}_{\hat{\gamma}}) = \min_{Q\in\mathcal{P}(\mathcal{X})} D(Q\|\hat{P}_1) + D(Q\|\hat{P}_2) \tag{50}$$

$$= 2B(\hat{P}_1, \hat{P}_2) \tag{51}$$

where $B(\hat{P}_1, \hat{P}_2)$ is the Bhattacharyya distance between the mismatched distributions $\hat{P}_1$ and $\hat{P}_2$. This suggests that having equal sensitivity (or robustness) for both hypotheses minimizes the sum of the exponents.

Example 1. When $\gamma = 0$ the likelihood ratio test becomes the maximum-likelihood test, which is known to achieve the lowest average probability of error in the Bayes setting for equal priors. For fixed priors π_1, π_2, the error probability in the Bayes setting is $\bar{\epsilon} = \pi_1\epsilon_1 + \pi_2\epsilon_2$, resulting in the following error exponent [3]

$$\bar{E} = \lim_{n\to\infty} \frac{1}{n} \log \bar{\epsilon} = \min\{E_1, E_2\}. \tag{52}$$

Consider $\hat{P}_1 = \text{Bern}(0.1)$, $\hat{P}_2 = \text{Bern}(0.8)$. Also, assume $R_1 = R_2 = R$. Figure 2 shows the worst-case error exponent in the Bayes setting given by $\min\{\hat{E}_1^L, \hat{E}_2^L\}$ by solving (32) and (33) as well as $\min\{\tilde{E}_1^L, \tilde{E}_2^L\}$ using the approximation in (46). We can see that the approximation is good for small R. Moreover, it can be seen that error exponents are very sensitive to mismatch for small R, i.e., the slope of the worst-case exponent goes to infinity as R approaches to zero.

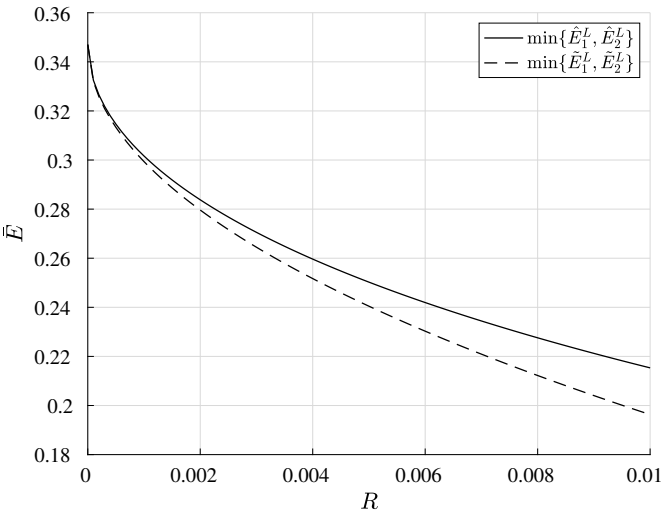


Fig. 2. Worst-case achievable Bayes error exponent.

REFERENCES

[1] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, Springer Texts in Statistics. Springer, New York, third edition, 2005.

[2] 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.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, July 2006.

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 95, 01 2010.

[5] R. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 405–417, July 1974.

[6] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, no. 2, pp. 369–401, 04 1965.

[7] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, no. 6, pp. 1753–1758, 12 1965.

[8] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, March 1985.

[9] H. V. Poor, *An introduction to signal detection and estimation*, Springer, 2013.

[10] P. Boroumand and A. Guillén i Fàbregas, "Error exponents of mismatched likelihood ratio testing," *http://arxiv.org/abs/2001.03917*, 2020.

[11] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1587–1603, Mar. 2011.

[12] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 413–421, Jan 2009.

[13] 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.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Properties of a Recent Upper Bound to the Mismatch Capacity

Ehsan Asadi Kangarshahi
University of Cambridge
ea460@cam.ac.uk

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—We study several properties of the upper bound on the mismatch capacity problem we recently proposed. In particular, we show that the bound can be cast as a convex-concave saddlepoint problem enabling efficient computation. Moreover, as opposed to multiple achievability bounds in the literature, we show that the multiletter version of this bound does not yield any improvement. In addition, for binary-input channels, we show a necessary condition for the mismatch capacity to be strictly smaller than the channel capacity.**

I. Introduction and Preliminaries

We consider reliable communication over a discrete memoryless channel (DMC) W with a given decoding metric [1], [2]. This problem arises when the decoder uses a suboptimal decoding rule due to limited computational resources, or imperfect channel estimation. Moreover, it is shown in [2] that important problems in information theory, like zero-error capacity of a channel can be cast as instances of mismatched decoding. Multiple achievability results have been reported in the literature [1]–[4] (see also [5]). These results are derived by random-coding techniques, i.e. analyzing the average probability of error of mismatched decoder over an ensemble of codebooks. On the other hand, the only single-letter converse was given in [6], where it was claimed that for binary-input DMCs, the mismatch capacity was the achievable rate derived in [3], [4]. Reference [7] provided a counterexample to this converse invalidating its claim. Multiletter converse results were proposed in [8].

We assume that the input and output alphabets are $\mathcal{X} = \{1, 2, \cdots, J\}$ and $\mathcal{Y} = \{1, 2, \cdots, K\}$, respectively, with $J, K < \infty$. We denote the channel transition probability by $W(k|j), k \in \mathcal{Y}, j \in \mathcal{X}$. A codebook $\mathcal{C}_n$ is defined as a set of M sequences $\mathcal{C}_n = \big\{\boldsymbol{x}(1), \boldsymbol{x}(2), \cdots, \boldsymbol{x}(M)\big\}$, where $\boldsymbol{x}(m) = \big(x_1(m), x_2(m), \cdots, x_n(m)\big) \in \mathcal{X}^n$, for $m \in \{1, 2, \cdots, M\}$. A message $m \in \{1, 2, \cdots, M\}$ is chosen equiprobably and $\boldsymbol{x}(m)$ is sent over the channel. The channel produces a noisy observation $\boldsymbol{y} = (y_1, y_2, \cdots, y_n) \in \mathcal{Y}^n$ according to $W^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n W(y_i|x_i)$. Upon observing $\boldsymbol{y} \in \mathcal{Y}^n$ the decoder produces an estimate of the transmitted message $\hat{m} \in \{1, 2, \cdots, M\}$. The decoder that minimizes the error probability is the maximum-likelihood (ML) decoder, that produces the message estimate $\hat{m}$ according to

$$\hat{m} = \underset{i \in \{1,2,\cdots,M\}}{\arg\max}\ W^n\big(\boldsymbol{y}|\boldsymbol{x}(i)\big). \tag{1}$$

Rate $R > 0$ is achievable if for any $\epsilon > 0$ there exists a sequence of length-n codebooks $\{\mathcal{C}_n\}_{n=1}^{\infty}$ such that $|\mathcal{C}_n| \geq 2^{n(R-\epsilon)}$, and $\liminf_{n\to\infty} P_e(\mathcal{C}_n) = 0$. The capacity of W, denoted by $C(W)$, is defined as the largest achievable rate.

In multiple practical scenarios, it is not possible to use a decoder based on W^n and instead, the decoder produces the message estimate $\hat{m}$ as

$$\hat{m} = \underset{i \in \{1,2,\cdots,M\}}{\arg\max}\ d\big(\boldsymbol{x}(i), \boldsymbol{y}\big), \tag{2}$$

where,

$$d\big(\boldsymbol{x}(i), \boldsymbol{y}\big) = \sum_{\ell=1}^{n} d\big(x_\ell(i), y_\ell\big) \tag{3}$$

The mismatch capacity $C_d(W)$ is defined as the largest achievable rate with decoder (2). Recently, we have shown that $C_d(W)$ is upper bounded by the following quantity,

$$\bar{R}_d(W) = \max_{P_X} \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} I(P_X, P_{\hat{Y}|X}) \tag{4}$$

where $I(P_X, P_{\hat{Y}|X}) \triangleq I(X; \hat{Y})$ and the set $\mathcal{M}_{\max}(d)$ is given in the following definition.

Definition 1: Let $P_{Y\hat{Y}|X}$ be a joint conditional distribution and define the set $\mathcal{S}(k_1, k_2) \triangleq \big\{i \in \mathcal{X} | i = \arg\max_{i' \in \mathcal{X}} d(i', k_2) - d(i', k_1)\big\}$. We say that $P_{Y\hat{Y}|X}$ is a maximal joint conditional distribution if for all $(j, k_1, k_2) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$,

$$P_{Y\hat{Y}|X}(k_1, k_2|j) = 0 \ \text{ if } \ j \notin \mathcal{S}(k_1, k_2). \tag{5}$$

For a given decoding metric d, we define the set of maximal joint conditional distributions to be $\mathcal{M}_{\max}(d)$.

In this paper we study some properties of the upper bound (4). Specifically, in Section II, we show that computing our upper bound is a convex-concave saddlepoint problem and we derive the optimality KKT conditions. In Section III, we show that the multiletter version of the upper bound coincides with the single-letter one. In Section IV, we derive a sufficient condition for $C_d(W) < C(W)$ for binary-input channels.

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 TEC2016-78434-C3-1-R.

II. Convexity Analysis

In this section, we show that the optimization (4) is a convex-concave saddlepoint problem. First, we argue that the constraints induce a convex set.

Lemma 1: For any channel W and metric d, the set of joint conditional distributions $P_{Y\hat{Y}|X}$ satisfying both $P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d)$ and $P_{Y|X} = W$, is a convex set.

Proof: Let $P_{Y\hat{Y}|X}$ and $P'_{Y\hat{Y}|X}$ both satisfy the above constraints. Now for any $0 < \lambda < 1$ we have,

$$\lambda P_{Y|X} + (1-\lambda) P'_{Y|X} = W. \tag{6}$$

In addition, if for some k_1, k_2 we have $j \notin \mathcal{S}(k_1, k_2)$, both $P_{Y\hat{Y}|X}(k_1, k_2|j)$ and $P'_{Y\hat{Y}|X}(k_1, k_2|j)$ are equal to zero, and so is any linear combination of them. Therefore,

$$\lambda P_{Y\hat{Y}|X} + (1-\lambda) P'_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d). \tag{7}$$

■

Moreover, $I(P_X, P_{\hat{Y}|X})$ is convex in terms of $P_{\hat{Y}|X}$, and concave in terms of P_X. Since $P_{\hat{Y}|X}$ is a linear function of $P_{Y\hat{Y}|X}$, we get that $I(P_X, P_{\hat{Y}|X})$ is convex in terms of $P_{Y\hat{Y}|X}$. Therefore from the minimax theorem [9] we get,

$$\bar{R}_d(W) = \max_{P_X} \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} I(P_X, P_{\hat{Y}|X}) \tag{8}$$

$$= \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} \max_{P_X} I(P_X, P_{\hat{Y}|X}) \tag{9}$$

$$= \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} C(P_{\hat{Y}|X}). \tag{10}$$

The rest of this section is devoted to deriving the KKT conditions for the optimization problem in (4). Given that $I(P_X, P_{\hat{Y}|X})$ is convex in $P_{Y\hat{Y}|X}$, and concave in P_X, then the KKT conditions are sufficient for global optimality. For convenience, we define $f(P_X, P_{Y\hat{Y}|X}) \triangleq I(P_X, P_{\hat{Y}|X})$ and rewrite the optimization problem in (4) as,

$$\bar{R}_d(W) = \max_{P_X} \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} f(P_X, P_{Y\hat{Y}|X}). \tag{11}$$

Let $\widehat{P}_X, \widehat{P}_{Y\hat{Y}|X}$ be the optimal input and joint conditional distributions in (11) and $\hat{q}_{\hat{Y}}$ be the output distribution induced by $\widehat{P}_X$ and $\widehat{P}_{\hat{Y}|X}$. Then for $\widehat{P}_X$ we have the following constraints:

$$\widehat{P}_X(j) \geq 0, \quad \forall j \in \mathcal{X} \tag{12}$$

$$\sum_{j \in \mathcal{X}} \widehat{P}_X(j) = 1. \tag{13}$$

Let $\mu_j, j = 1, 2, \cdots, J$ be the Lagrange multipliers corresponding the inequalities in (12) and ρ be the Lagrange multiplier corresponding to (13). Therefore, from stationarity we have,

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} = \rho + \mu_j \tag{14}$$

and from the complementary slackness [10] we have $\mu_j \widehat{P}_X(j) = 0$ and from the dual feasibility we have $\mu_j \geq 0$ which leads to the separation of the equations of in two cases. If $\widehat{P}_X(j) > 0$

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} = \rho, \tag{15}$$

while when $\widehat{P}_X(j) = 0$ we have

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} \leq \rho. \tag{16}$$

Note that, because there is no other constraint on μ_j, all of the KKT conditions are summarized in (16) and (15). Moreover, computing the derivatives in (15) and (16) gives

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} = \sum_{k \in \mathcal{Y}} \widehat{P}_{\hat{Y}|X}(k|j) \log \frac{\widehat{P}_{\hat{Y}|X}(k|j)}{\hat{q}_{\hat{Y}}(k)} - 1. \tag{17}$$

As for $\widehat{P}_{Y\hat{Y}|X}$, we have the following constraints. For all $j, k_1, k_2 \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$,

$$\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) \geq 0, \tag{18}$$

$$\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) = 0, \text{ if } j \notin \mathcal{S}(k_1, k_2) \tag{19}$$

where (18) corresponds to $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j)$ being a distribution and (19) corresponds to $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) \in \mathcal{M}_{\max}(d)$. Moreover from the constraint $P_{Y|X} = W$ we get for all $j, k_1 \in \mathcal{X} \times \mathcal{Y}$

$$\sum_{k_2} \widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) = W(k_1|j). \tag{20}$$

For the ease of notation, we skip the step of explicitly considering a Lagrange multiplier for (18). However, after simplification, The following KKT conditions are equivalent to the full KKT conditions considering a Lagrange multiplier for (18). Details follow similarly to the above derivation. If we use a Lagrange multiplier λ_{j,k_1} for each of the conditions in (20), we have when $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) > 0$

$$\frac{\partial}{\partial P_{Y\hat{Y}|X}(k_1, k_2|j)} f(\widehat{P}_X, P_{Y\hat{Y}|X}) \bigg|_{P_{Y\hat{Y}|X} = \widehat{P}_{Y\hat{Y}|X}} = \lambda_{j,k_1} \tag{21}$$

and when $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) = 0$ and $j \in \mathcal{S}(k_1, k_2)$ we have

$$\frac{\partial}{\partial P_{Y\hat{Y}|X}(k_1, k_2|j)} f(\widehat{P}_X, P_{Y\hat{Y}|X}) \bigg|_{P_{Y\hat{Y}|X} = \widehat{P}_{Y\hat{Y}|X}} \geq \lambda_{j,k_1}. \tag{22}$$

Explicitly computing the derivative gives

$$\frac{\partial}{\partial P_{Y\hat{Y}|X}(k_1, k_2|j)} f(\widehat{P}_X, P_{Y\hat{Y}|X}) \bigg|_{P_{Y\hat{Y}|X} = \widehat{P}_{Y\hat{Y}|X}} \tag{23}$$

$$= \widehat{P}_X(j) \log \frac{\widehat{P}_{\hat{Y}|X}(k_2|j)}{\hat{q}_{\hat{Y}}(k_2)}. \tag{24}$$

Summarizing, for the KKT optimality conditions of we get the following inequalities

1) For $\widehat{P}_X(j) > 0$,

$$\sum_{k\in\mathcal{Y}} \widehat{P}_{\hat{Y}|X}(k|j) \log \frac{\widehat{P}_{\hat{Y}|X}(k|j)}{\hat{q}_{\hat{Y}}(k)} = 1+\rho, \tag{25}$$

2) For $\widehat{P}_X(j) = 0$,

$$\sum_{k\in\mathcal{Y}} \widehat{P}_{\hat{Y}|X}(k|j) \log \frac{\widehat{P}_{\hat{Y}|X}(k|j)}{\hat{q}_{\hat{Y}}(k)} \leq 1+\rho, \tag{26}$$

3) For $\widehat{P}_{Y\hat{Y}|X}(k_1,k_2|j) > 0$,

$$\widehat{P}_X(j) \log \frac{\widehat{P}_{\hat{Y}|X}(k_2|j)}{\hat{q}_{\hat{Y}}(k_2)} = \lambda_{j,k_1}, \tag{27}$$

4) For $\widehat{P}_{Y\hat{Y}|X}(k_1,k_2|j) = 0$ and $j \in \mathcal{S}(k_1,k_2)$,

$$\widehat{P}_X(j) \log \frac{\widehat{P}_{\hat{Y}|X}(k_2|j)}{\hat{q}_{\hat{Y}}(k_2)} \geq \lambda_{j,k_1}. \tag{28}$$

In the next section, we employ the above KKT conditions to analyze the multiletter version of our bound.

III. Multiletter Bound

In this section, we study the multiletter extension of the bound (4). In particular, we show that the multiletter version cannot improve on its single-letter counterpart. We define the ℓ-letter decoding metric $d^{(\ell)} : \mathcal{X}^\ell \times \mathcal{Y}^\ell \to \mathbb{R}$ as follows

$$d^{(\ell)}\big((x_1,x_2,\cdots,x_\ell),(y_1,y_2,\cdots,y_\ell)\big) = \sum_{i=1}^{\ell} d(x_i,y_i). \tag{29}$$

This decoding metric definition is consistent with the additive decoder we have defined in (3). We denote $\boldsymbol{j} \in \mathcal{X}^\ell$ and $\boldsymbol{k} \in \mathcal{Y}^\ell$ as the ℓ-letter inputs and outputs, respectively. Let $W^{(\ell)}$ denote a DMC over input alphabet $\mathcal{X}^\ell$ and output alphabet $\mathcal{Y}^\ell$ with the channel rule $W^{(\ell)}\big((y_1,y_2,\cdots,y_\ell)|(x_1,x_2,\cdots,x_\ell)\big) = \prod_{i=1}^{\ell} W(y_i|x_i)$. Additionally, we define $P_X^{(\ell)}$ and $P_{Y\hat{Y}|X}^{(\ell)}$ accordingly

$$P_X^{(\ell)}(x_1,\dots x_\ell) = \prod_{i=1}^{\ell} P_X(x_i) \tag{30}$$

$$P_{Y\hat{Y}|X}^{(\ell)}\big((y_1,y_2,\cdots,y_\ell),(\hat{y}_1,\hat{y}_2,\cdots,\quad \hat{y}_\ell)|(x_1,x_2,\cdots,x_\ell)\big)$$
$$= \prod_{i=1}^{\ell} P_{Y\hat{Y}|X}(y_i,\hat{y}_i|x_i) \tag{31}$$

X^ℓ and $Y^\ell, \hat{Y}^\ell$ denote random variables defined on alphabets $\mathcal{X}^\ell$, $\mathcal{Y}^\ell$ and $\mathcal{Y}^\ell$, respectively. Moreover, $\mathcal{S}^{(\ell)}(\boldsymbol{k}_1,\boldsymbol{k}_2)$ is defined as

$$\mathcal{S}^{(\ell)}(\boldsymbol{k}_1,\boldsymbol{k}_2) \triangleq \big\{\boldsymbol{i}\in\mathcal{X}^\ell \,|\, \boldsymbol{i} = \arg\max_{\boldsymbol{i}'\in\mathcal{X}^\ell} d^{(\ell)}(\boldsymbol{i}',\boldsymbol{k}_2) - d^{(\ell)}(\boldsymbol{i}',\boldsymbol{k}_1)\big\}. \tag{32}$$

In the following lemma we characterize the sets $\mathcal{S}^{(\ell)}(\boldsymbol{k}_1,\boldsymbol{k}_2)$ and relate them to $\mathcal{S}(k_{1,i},k_{2,i}), i = 1,2,\cdots,\ell$.

Lemma 2: For $\boldsymbol{j} \in \mathcal{X}^\ell, \boldsymbol{k}_1 \in \mathcal{Y}^\ell, \boldsymbol{k}_2 \in \mathcal{Y}^\ell$ we have that $\boldsymbol{j} \in \mathcal{S}^{(\ell)}(\boldsymbol{k}_1,\boldsymbol{k}_2)$ if and only if for all $1 \leq i \leq \ell$ we have

$$j_i \in \mathcal{S}(k_{1,i},k_{2,i}). \tag{33}$$

Proof: We have

$$\arg\max_{\boldsymbol{j}\in\mathcal{X}^\ell} d^{(\ell)}(\boldsymbol{j},\boldsymbol{k}_2) - d^{(\ell)}(\boldsymbol{j},\boldsymbol{k}_1) \tag{34}$$

$$= \arg\max_{\boldsymbol{j}\in\mathcal{X}^\ell} \sum_{i=1}^{\ell} d(j_i,k_{2,i}) - d(j_i,k_{2,i}) \tag{35}$$

$$= \arg\max_{(j_1,j_2,\cdots,j_\ell)\in\mathcal{X}^\ell} \sum_{i=1}^{\ell} d(j_i,k_{2,i}) - d(j_i,k_{2,i}) \tag{36}$$

From (36) we get that if $(j_1,j_2,\cdots,j_\ell) \in \mathcal{S}(\boldsymbol{k}_1,\boldsymbol{k}_2)$ then for all $1 \leq i \leq \ell$ we should have $j_i \in \mathcal{S}(k_{1,i},k_{2,i})$. Therefore,

$$\mathcal{S}^{(\ell)}(\boldsymbol{k}_1,\boldsymbol{k}_2) = \mathcal{S}(k_{1,1},k_{2,1}) \times \mathcal{S}(k_{1,2},k_{2,2}) \times \cdots \times \mathcal{S}(k_{1,\ell},k_{2,\ell}). \tag{37}$$

■

For the above ℓ-letter alphabets and distributions, the construction and analysis of the bound remains unchanged. Therefore, (4) remains valid for its ℓ-letter extension, which can be written as

$$\bar{R}_d^{(\ell)}(W) \triangleq \frac{1}{\ell}\bar{R}_{d^{(\ell)}}(W^{(\ell)}) \tag{38}$$

$$= \frac{1}{\ell}\max_{p_{X^\ell}} \min_{\substack{P_{Y^\ell\hat{Y}^\ell|X^\ell}\in\mathcal{M}_{\max}(d^{(\ell)}) \\ P_{Y^\ell|X^\ell}=W^{(\ell)}}} I(p_{X^\ell}, P_{Y^\ell\hat{Y}^\ell|X^\ell}). \tag{39}$$

We have the following result.

Proposition 1:

$$\bar{R}_d^{(\ell)}(W) = \bar{R}_d(W). \tag{40}$$

Proof: Given that $I(P_X,P_{\hat{Y}|X})$ is convex in $P_{Y\hat{Y}|X}$, and concave in P_X, the KKT conditions are also sufficient for global optimality. Similarly, $f(P_{X^\ell},P_{Y^\ell\hat{Y}^\ell|X^\ell})$ is convex in P_{X^ℓ} and concave in $P_{Y^\ell\hat{Y}^\ell|X^\ell}$. Here we use the optimality conditions derived in the previous section to show that if $\widehat{P}_X, \widehat{P}_{Y\hat{Y}|X}$ are the optimal distributions for the single-letter bound then $\widehat{P}_X^{(\ell)}, \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ defined in (30) and (31) are optimal distributions for the multiletter version. As a result, if we find a feasible pair $P_{Y^\ell\hat{Y}^\ell|X^\ell}, P_{X^\ell}$ such that when fixing $P_{Y^\ell\hat{Y}^\ell|X^\ell}$, the input distribution P_{X^ℓ} is a maximizer of $f(\cdot,P_{Y^\ell\hat{Y}^\ell|X^\ell})$, and when fixing P_{X^ℓ}, the joint conditional distribution $P_{Y^\ell\hat{Y}^\ell|X^\ell}$ is a minimizer of $f(p_{X^\ell},\cdot)$, then the pair $(P_{Y^\ell\hat{Y}^\ell|X^\ell}, P_{X^\ell})$ is a saddlepoint.

We need to show that if $\widehat{P}_X, \widehat{P}_{Y\hat{Y}|X}$ is a saddlepoint for the single-letter case, then, $\widehat{P}_X^{(\ell)}, \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ is a saddlepoint for the multiletter bound. Based on the aforementioned argument, it is sufficient to show that $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ is a minimizer of (39) by fixing

$\widehat{P}_X^{(\ell)}$. This is because it is known that $\frac{1}{\ell}C(\widehat{P}_{\hat{Y}|X}^{(\ell)}) = C(P_{\hat{Y}|X})$, i.e., the product distribution $\widehat{P}_X^{(\ell)}$ achieves $C(\widehat{P}_{\hat{Y}|X}^{(\ell)})$.

In the following lemma, we prove that by fixing $\widehat{P}_X^{(\ell)}$, then $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ satisfies the KKT conditions and hence, it is a minimizer of (39). Before stating the result we recall that the multiletter counterparts of the single-letter KKT conditions given in (27) and (28) hold. Moreover, as in the single-letter case, the multiletter KKT conditions are sufficient for global optimality, because the function $f(\widehat{P}_X^{(\ell)}, \cdot)$ is concave. Using Lemma 3 below completes the proof. ∎

Lemma 3: Let $\widehat{P}_X, \widehat{P}_{Y\hat{Y}|X}$ be a saddlepoint for optimization problem (4). Set $P_{X^\ell} = \widehat{P}_X^{(\ell)}$. Then, the joint conditional distribution $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ is a minimizer of

$$\min_{\substack{P_{Y^\ell\hat{Y}^\ell|X^\ell} \in \mathcal{M}_{\max}(d^{(\ell)}) \\ P_{Y^\ell|X^\ell} = W^{(\ell)}}} f\big(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell}\big). \tag{41}$$

Proof: We should show that by setting $P_{X^\ell} = \widehat{P}_X^{(\ell)}$, the multiletter versions of the KKT conditions (27) and (28) hold for $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$. Generalizing the conditions of (27) and (28) to the multiletter case, and setting $P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$, we should show that for all $\boldsymbol{j}, \boldsymbol{k}_1 \in \mathcal{X}^\ell \times \mathcal{Y}^\ell$ there exist $\lambda_{\boldsymbol{j},\boldsymbol{k}_1}$ such that the conditions below are fulfilled. If we show this, then the Lemma is proved because these are precisely the conditions for the minimizer of (41).

i) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j}) > 0$ we must have,

$$\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} = \lambda_{\boldsymbol{j},\boldsymbol{k}_1}. \tag{42}$$

ii) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j}) = 0$ and $\boldsymbol{j} \in \mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ we must have that,

$$\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} \geq \lambda_{\boldsymbol{j},\boldsymbol{k}_1}. \tag{43}$$

Similarly to (23), the derivative in (42) and (43) is,

$$\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} = \widehat{P}_X^{(\ell)}(\boldsymbol{j}) \log \frac{\widehat{P}_{\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1|\boldsymbol{j})}{\hat{q}_{\hat{Y}}^{(\ell)}(\boldsymbol{k}_1)} \tag{44}$$

which, by using that $P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$, $\widehat{P}_X^{(\ell)}$ and $\hat{q}_{\hat{Y}}^{(\ell)}$ are product distributions, gives,

$$\widehat{P}_X^{(\ell)}(\boldsymbol{j}) \log \frac{\widehat{P}_{\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1|\boldsymbol{j})}{\hat{q}_{\hat{Y}}^{(\ell)}(\boldsymbol{k}_1)} = \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \log \frac{\widehat{P}_{\hat{Y}|X}(k_{2,i}|j_i)}{\hat{q}_{\hat{Y}}(k_{2,i})}\right) \tag{45}$$

In order to show that there exist some coefficients $\lambda_{\boldsymbol{j},\boldsymbol{k}_1}$ satisfying both (42) and (43), we make a particular choice and show that this specific choice satisfies both (42) and (43). To this end, define

$$\lambda_{\boldsymbol{j},\boldsymbol{k}_1} = \begin{cases} 0 & \widehat{P}_X(\boldsymbol{j}) = 0 \\ \prod_{i=1}^{\ell} \widehat{P}_X(j_i) \left(\sum_{i=1}^{\ell} \frac{\lambda_{j_i,k_{1,i}}}{\widehat{P}_X(j_i)}\right) & \widehat{P}_X(\boldsymbol{j}) \neq 0 \end{cases} \tag{46}$$

where $\lambda_{j_i,k_{1,i}}$ is the single-letter Lagrange multiplier corresponding to j_i and $k_{1,i}$.

Now, excluding the cases where $\widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell) = 0$ where from (45), (42) and (43) the KKT conditions clearly hold, we have two cases i) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j}) > 0$, then for all $1 \leq i \leq \ell$ we must have $\widehat{P}_{Y\hat{Y}|X}(k_{1,i}, k_{2,i}|j_i) > 0$ and therefore, (27) is valid. We have to verify that this implies that (42) is also valid. Thus,

$$\begin{aligned} &\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} \\ &= \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \log \frac{\widehat{P}_{\hat{Y}|X}(k_{2,i}|j_i)}{\hat{q}_{\hat{Y}}(k_{2,i})}\right) && (47) \\ &= \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \frac{\lambda_{j_i,k_{1,i}}}{\widehat{P}_X(j_i)}\right) && (48) \\ &= \lambda_{\boldsymbol{j},\boldsymbol{k}_1} && (49) \end{aligned}$$

where (48) holds from the single-letter optimality in (27). ii) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j}) = 0$ and $\boldsymbol{j} \in \mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$, as a result of Lemma 2, we have that $\mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ is a product set, i.e., for all $1 \leq i \leq \ell$,

$$j_i \in \mathcal{S}(k_{1,i}, k_{2,i}). \tag{50}$$

Moreover, either $\widehat{P}_{Y\hat{Y}|X}(k_{1,i}, k_{2,i}|j_i) > 0$ where (27) is satisfied or $\widehat{P}_{Y\hat{Y}|X}(k_{1,i}, k_{2,i}|j_i) = 0$ where (28) is satisfied. Now, with these assumptions, we should verify that (43) holds. We have,

$$\begin{aligned} &\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} \\ &= \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \log \frac{\widehat{P}_{\hat{Y}|X}(k_{2,i}|j_i)}{\hat{q}_{\hat{Y}}(k_{2,i})}\right) && (51) \\ &\geq \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \frac{\lambda_{j_i,k_{1,i}}}{\widehat{P}_X(j_i)}\right) && (52) \\ &= \lambda_{\boldsymbol{j},\boldsymbol{k}_1} && (53) \end{aligned}$$

where (52) is true because of the single-letter optimality in (27) and (28).

∎

IV. Binary-input Channels

In [2], the authors state that for any DMC and decoding metric $d(x,y)$, the mismatch capacity $C_d(W)$ remains unaltered for a decoder with metric $\tilde{d}(x,y) = d(x,y)+a(x)+b(y)$, where $a(x), b(y)$ are functions of the input and output, respectively. This property suggests that for binary-input channels, the mismatch capacity $C_d(W)$ is only a function of the metric difference $d(1,y) - d(2,y)$. In this section, we show a necessary condition for $C_d(W) < C(W)$ for binary-input channels based on the above observation.

Definition 2: We say that two sequences $\{\alpha_i\}_{i=1}^K$ and $\{\beta_i\}_{i=1}^K$ have the same order if for all $1 \le i_1, i_2 \le K$

$$\alpha_{i_1} \ge \alpha_{i_2} \Rightarrow \beta_{i_1} \ge \beta_{i_2}. \tag{54}$$

We have the following result.

Theorem 1: Assume that $W(k|j) > 0$, for all $j = 1, 2$, $k = 1, \dots K$. If the sequences $\big\{\log W(k|1) - \log W(k|2)\big\}_{k=1}^K$ and $\big\{d(1,k) - d(2,k)\big\}_{k=1}^K$ do not have the same order, then $\bar{R}_d(W) < C(W)$.

Proof: Without loss of generality, we assume that the sequence $\big\{d(1,k) - d(2,k)\big\}_{k=1}^K$ is non-decreasing, i.e., for $k_1 \le k_2$,

$$d(1,k_1) - d(2,k_1) \le d(1,k_2) - d(2,k_2). \tag{55}$$

This assumption simplifies the evaluation of the sets $\mathcal{S}(\cdot,\cdot)$. For $k_1 = k_2$ we have $\mathcal{S}(k_1,k_2) = \{1,2\}$. Moreover, when $k_1 < k_2$ from (55) and Definition 1, we have that $1 \in \mathcal{S}(k_1,k_2)$ and $2 \in \mathcal{S}(k_2,k_1)$.

We prove a slightly stronger result. In particular, we prove that the condition $C_d(W) = C(W)$ implies that sequences

$$\Big\{\widehat{P}_X(1)\log\frac{W(k|1)}{\hat{q}_{\hat{Y}}(k)}\Big\}_{k=1}^K,\ \Big\{-\widehat{P}_X(2)\log\frac{W(k|2)}{\hat{q}_{\hat{Y}}(k)}\Big\}_{k=1}^K \tag{56}$$

both should have the same order as the decoding metric difference sequence $\{d(1,k) - d(2,k)\}_{k=1}^K$, where recall that the notation $\widehat{P}_X$ refers to the capacity-achieving distribution of W.

Now assume that $C_d(W) = C(W)$. Therefore, $\widehat{P}_X, P_{Y\hat{Y}|X} = P_{YY|X}$ must be a saddlepoint of (9). As a result, the KKT conditions in (27) and (28) must hold. Observe that

$$P_{YY|X}(k_1,k_2|j) = \begin{cases} W(k_1|j) & k_1 = k_2 \\ 0 & k_1 \ne k_2. \end{cases} \tag{57}$$

Therefore, combining the KKT conditions in (27) (28) we have,

1) If $k_1 = k_2$, for both $j = 1, 2$ we have

$$\widehat{P}_X(j)\log\frac{W(k_1|j)}{\hat{q}_Y(k_1)} = \lambda_{j,k_1} \tag{58}$$

2) If $k_1 < k_2$ we know $1 \in \mathcal{S}(k_1,k_2)$ and $2 \in \mathcal{S}(k_2,k_1)$

$$\widehat{P}_X(1)\log\frac{W(k_2|1)}{\widehat{q}_Y(k_2)} \ge \lambda_{1,k_1} \tag{59}$$

$$\widehat{P}_X(2)\log\frac{W(k_1|2)}{\widehat{q}_Y(k_1)} \ge \lambda_{2,k_2} \tag{60}$$

Therefore, we get that if $k_1 < k_2$

$$\widehat{P}_X(1)\log\frac{W(k_2|1)}{\widehat{q}_Y(k_2)} \ge \lambda_{1,k_1} = \widehat{P}_X(1)\log\frac{W(k_1|1)}{\hat{q}_Y(k_1)} \tag{61}$$

$$\widehat{P}_X(2)\log\frac{W(k_1|2)}{\widehat{q}_Y(k_1)} \ge \lambda_{2,k_2} = \widehat{P}_X(2)\log\frac{W(k_2|2)}{\hat{q}_Y(k_2)}. \tag{62}$$

Thus, we get that $\Big\{\widehat{P}_X(1)\log\frac{W(k|1)}{\hat{q}_Y(k)}\Big\}_{k=1}^K$ and $-\Big\{\widehat{P}_X(2)\log\frac{W(k|2)}{\hat{q}_Y(k)}\Big\}_{k=1}^K$ are both non-decreasing sequences and so is any linear combination of them with positive coefficients. Therefore, since

$$\begin{aligned}\log W(k|1) - \log W(k|2) = &\frac{1}{\widehat{P}_X(1)}\Big(\widehat{P}_X(1)\log\frac{W(k|1)}{\hat{q}_Y(k)}\Big) \\ &- \frac{1}{\widehat{P}_X(2)}\Big(\widehat{P}_X(2)\log\frac{W(k|2)}{\hat{q}_Y(k)}\Big)\end{aligned} \tag{63}$$

we conclude that the sequence $\{\log W(k|1) - \log W(k|2)\}_{k=1}^K$ is a non-decreasing sequence.

■

References

[1] N. Merhav, G. Kaplan, A. Lapidoth, and S. S. Shitz, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, pp. 1953–1967, Nov. 1994.

[2] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theory*, vol. 41, pp. 35–43, Jan. 1995.

[3] J. Y. N. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, Massachusetts Institute of Technology, 1983.

[4] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, pp. 5–12, Jan. 1981.

[5] J. Scarlett, "Reliable communication under mismatched decoding," Ph.D. dissertation, Ph. D. dissertation, University of Cambridge, 2014, [Online: http://itc.upf.edu/biblio/1061, 2014.

[6] V. B. Balakirsky, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1889–1902, 1995.

[7] J. Scarlett, A. Somekh-Baruch, A. Martinez, and A. Guillén i Fàbregas, "A counter-example to the mismatched decoding converse for binary-input discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 61, pp. 5387–5395, Oct. 2015.

[8] A. Somekh-Baruch, "Converse theorems for the DMC with mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 64, pp. 6196–6207, Sept. 2018.

[9] J. von Neumann, "Zur Theorie der Gesellschaftsspiele," *Math. Ann.*, vol. 100, pp. 295–320, 1928.

[10] S. Boyd and L. Vandenberghe, "Convex Optimization," *Cambridge University Press*, 2004.

Fundamental Limits of Wireless Caching under Uneven-Capacity Channels

Eleftherios Lampiris, Jingjing Zhang, Osvaldo Simeone, Petros Elia

***Abstract*— This work identifies the fundamental limits of cache-aided coded multicasting in the presence of the well-known 'worst-user' bottleneck. This stems from the presence of receiving users with uneven channel capacities, which often forces the rate of transmission of each multicasting message to be reduced to that of the slowest user. This bottleneck, which can be detrimental in general wireless broadcast settings, motivates the analysis of coded caching over a standard Single-Input-Single-Output (SISO) Broadcast Channel (BC) with K cache-aided receivers, each with a generally different channel capacity. For this setting, we design a communication algorithm that is based on superposition coding that capitalizes on the realization that the user with the worst channel may not be the real bottleneck of communication. We then proceed to provide a converse that shows the algorithm to be near optimal, identifying the fundamental limits of this setting within a multiplicative factor of 4. Interestingly, the result reveals that, even if several users are experiencing channels with reduced capacity, the system can achieve the same optimal delivery time that would be achievable if all users enjoyed maximal capacity.**

I. Introduction

The seminal work in [1] showed how adding caches to receiving nodes can substantially reduce the time required to deliver content. Specifically, reference [1] studied the case in which a transmitter with access to a library of N unit-sized files serves – via a wired, single-stream, unit-capacity bottleneck link – K cache-aided receivers/users. Each user is equipped with a cache of size equal to a fraction $\gamma \in [0,1]$ of the size of the library, so that $K\gamma$ is the cumulative cache size normalized by the library size. For this setting, the authors of [1] proposed a novel cache placement algorithm and a novel multicast transmission policy that delivers any set of K files to the receivers with (normalized) delay at most

$$T_{MN} = \frac{K(1-\gamma)}{K\gamma+1} \tag{1}$$

thus revealing a speed-up factor of $K\gamma+1$ compared to the delay $K(1-\gamma)$ corresponding to a standard scheme that serves each user in turn.

Eleftherios is with the Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany (lampiris@tu-berlin.de). Petros is with the Communication Systems Department at EURECOM, Sophia Antipolis, 06410, France (elia@eurecom.fr). Their work is supported by the ANR project ECOLOGICAL-BITS-AND-FLOPS and the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929. (ERC project DUALITY). Jingjing and Osvaldo are with the Department of Informatics, King's College London, London, UK (jingjing.1.zhang@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk). Their work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731). This work was conducted while Eleftherios was employed by EURECOM.

The delay (1) is obtained by a *coded caching* approach that is based on the transmission of a sequence of multicast messages that convey information to several users at a time (even if these users requested different content), with users decoding their desired information by means of cache-aided interference cancellation. In this scheme, each multicast message consists of a XOR X_σ that carries information to a subset $\sigma \subset [K] \triangleq [1,2,\dots,K]$ of $|\sigma| = K\gamma+1$ users at a time.

While the promised speedup factor of $K\gamma+1$ in (1) is proportional to the normalized *cumulative cache size of the network*, it was quickly realized that a variety of bottlenecks severely hamper this performance. These include the subpacketization bottleneck [2]–[8], the uneven cache sizes bottleneck [9]–[12], and the bottleneck studied here that arises from uneven channel capacities between the transmitter and the users. This last bottleneck is particularly relevant in wireless scenarios with multicasting. Such networks produce "slower" users that can force the multicast rates to be reduced down to a level that can be decoded by these users. This can diminish the coded caching gains and could pose a serious limitation to any effort to implement cache-aided coded multicasting in wireless settings.

***Example* 1.** *Let us consider the wireless Single-Input-Single-Output (SISO) Broadcast Channel (BC) with K users, each equipped with a cache of normalized size γ, and let us further assume that all users have maximal normalized unit capacity, except for one user that has a normalized link capacity equal to $\frac{1}{K}+\gamma < 1$. It is easy to see that a (naive) transmission of the sequence of the XORs from [1] would induce the delay*

$$T = \frac{1-\gamma}{\frac{1+K\gamma}{K}} + \frac{(K-K\gamma-1)(1-\gamma)}{1+K\gamma} \tag{2}$$

$$= 2T_{MN} - (1-\gamma) \approx 2T_{MN} \tag{3}$$

which is approximately double the delay T_{MN} in (1) *that we would have if all users enjoyed unit normalized link capacities. It is also worth noting that approximately the same delay T in* (2) *would be obtained if we treated the slow user separately from the rest using time sharing. Essentially, whether with a naive or with a separated approach that excludes the slow user from coded caching, a single slow user can cause the worst-case delivery time to double, and the overall multicasting gain to be cut in half.*

A. Related Work

The importance of the uneven-channel bottleneck in coded caching has been acknowledged in a large number of recent

works that seek to understand and ameliorate this limitation [13]–[29]. For example, reference [13] focuses on the uneven link-capacity SISO BC where each user experiences a distinct channel strength, and proposes algorithms that outperform the naive implementation of the algorithm of [1] whereby each coded message is transmitted at a rate equal to the rate of the worst user whose message appears in the corresponding XOR operation. Under a similar setting, the work in [16] considered feedback-aided user selection that can maximize the sum-rate as well as increase a fairness criterion that ensures that each user receives their requested file in a timely manner. In the related context of the erasure BC where users have uneven probabilities of erasures, references [17] and [18] showed how an erasure at some users can be exploited as side information at the remaining users in order to increase system performance. Related work can also be found in [19]–[21].

The uneven-capacity bottleneck was also studied in the presence of multiple transmit antennas [15], [30]. Reference [15] exploited transmit diversity to ameliorate the impact of the worst-user capacity, and showed that employing $\mathcal{O}(\ln K)$ transmit antennas can allow for a transmission sum-rate that scales with K. Similarly, the work in [30] considered multiple transmit and multiple receive antennas, and designed topology-dependent cache-placement to ameliorate the worst-user effect.

In a related line of work, the papers [22] and [23] studied the cache-aided topological interference channel where K cache-aided transmitters are connected to K cache-aided receivers, and each transmitter is connected to one receiver via a direct "strong" link and to each of the other receivers via "weak" links. Under the assumption of no channel state information at the transmitters (CSIT), the authors showed how the lack of CSIT can be ameliorated by exploiting the topology of the channel and the multicast nature of the transmissions.

Recently, significant effort has been made toward understanding the behavior of coded caching in the finite Signal-to-Noise Ratio (SNR) regime with realistic (and thus often uneven) channel qualities. In this direction, the work in [24] showed that a single-stream coded caching message beamformed by an appropriate transmit vector can outperform some existing multi-stream coded caching methods in the low-SNR regime, while references [25], [26] (see also [27]) revealed the importance of jointly considering caching with multicast beamformer design. Moreover, the work in [28] studied the connection between rate and subpacketization in the multi-antenna environment, accounting for the unevenness naturally brought about by fading.

Our work is in the spirit of all the above papers, and it can be seen specifically as an extension of [14] which focused on the case of two link-strength levels, as well as the work of [29], where though the closely related scheme places focus on minimizing the power.

B. Overview of Results

In this paper, we study a cache-aided SISO BC where each receiver k experiences a link of some normalized capacity $\alpha_k \in [0,1]$. We establish the optimal worst-case delivery time $T(K,\gamma,\{\alpha_k\})$ within a factor of at most 4 for any number of K users, fractional cache capacity γ, and capacity set $\{\alpha_k\}$. Key to this result is a new algorithm that uses superposition coding, where (assuming without loss of generality that the users are labeled from weaker to stronger, i.e., such that $\alpha_k \leq \alpha_{k+1}$) we split the power into $K - K\gamma - 1$ layers, and in layer k, we transmit *only* XORs whose weakest user is user k. While this design indeed encodes some XORs at lower rates (matching the capacity of the worst user for that message), it also allows the simultaneous transmission of other XORs in the remaining power layers. The main result reveals that the optimal performance (1) achievable when $\alpha_k = 1$, for all $k \in [K] \triangleq [1,2,\ldots,K]$, is in fact achievable even if each user k has reduced link capacity such that the condition

$$\alpha_k \gtrsim 1 - e^{-k\gamma}, \quad \forall k \in [K] \tag{4}$$

is satisfied. This quantifies the intuitive fact that systems with smaller caches can be better immune to the negative effects of channel unevenness.

II. System Model

We consider the K-user wireless SISO BC, with the transmitter having access to a library of N files $\{W^n\}_{n=1}^N$, each of normalized unit size, and the K receivers having a cache whose size is equal to a fraction $\gamma \in [0,1]$ of the library size. Communication takes place in two distinct phases, namely the pre-fetching and the delivery phases. In the first phase, the caches of the users are filled with content from the library without any knowledge of future requests or of channel capacities. Then, during the delivery phase, each user k requests[1] a single file W^{d_k}, after which the transmitter – with knowledge of the requests and the link capacities – delivers the requested content. After transmission, at each user $k \in [K]$, the received signal takes the form

$$y_k = \sqrt{P^{\alpha_k}} h_k x + z_k, \tag{5}$$

where P represents the transmitting power; $x \in \mathbb{C}$ is the power-normalized transmitted signal satisfying $\mathbb{E}\{|x|^2\} \leq 1$; $h_k \in \mathbb{C}$ is the channel coefficient of user k; $z_k \sim \mathbb{C}\mathcal{N}(0,1)$ represents the Gaussian noise at user k; and $\alpha_k \in (0,1]$ is such that at each user $k \in [K]$ the average SNR equals

$$\mathbb{E}\{|y_k|^2\} = P^{\alpha_k}. \tag{6}$$

Under the simplified Generalized Degrees of Freedom (GDoF) framework of [31]–[33], condition (6) amounts to a (normalized, by a factor $\log P$) user rate of $r_k = \alpha_k \in [0,1]$. Without loss of generality, $\alpha = 1$ corresponds to the highest possible channel strength. We assume an arbitrary set of such normalized capacities $\boldsymbol{\alpha} \triangleq \{\alpha_k\}_{k=1}^K$ and we assume them without loss of generality to be ordered in ascending order ($\alpha_k \leq \alpha_{k+1}$).

The objective is to design the caching and communication scheme that minimizes the worst-case delivery time $T(K,\gamma,\boldsymbol{\alpha})$ for any capacity vector $\boldsymbol{\alpha}$.

[1] We are interested in the worse-case delivery time and thus we will assume that each user will ask for a different file.

III. Main Results

Before presenting the main results, we remind the reader that the naive implementation of coded caching which sequentially transmits the sequence of XORs X_σ to all subsets $\sigma \in [K]$ of $|\sigma| = K\gamma+1$ users, requires a worst-case delivery time

$$T_{uc}(K,\gamma,\boldsymbol{\alpha}) = \frac{1}{\binom{K}{K\gamma}} \sum_{\sigma\subseteq[K],\ |\sigma|=K\gamma+1} \max_{i\in\sigma}\left\{\frac{1}{\alpha_i}\right\}. \quad (7)$$

This follows since this conventional uncoded scheme allocates, for each XOR X_σ, a transmission time $T_\sigma = \max_{w\in\sigma}\left\{\frac{1}{\alpha_w}\right\}$ to allow the weakest user in σ to decode the message[2].

We now proceed with the main result.

Theorem 1. *In the K-user SISO BC with receiver channel strengths $\{\alpha_k\}_{k=1}^K (\alpha_k \le \alpha_{k+1})$ and with receivers equipped with a cache of normalized size γ, the worst-case delivery time*

$$T_{sc}(K,\gamma,\boldsymbol{\alpha}) = \max_{w\in[K]}\left\{\frac{1}{\alpha_w}\cdot\frac{\binom{K}{K\gamma+1}-\binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}}\right\} \quad (8)$$

is achievable and is within a multiplicative factor of at most 4 from the optimal delay $T^*(K,\gamma,\boldsymbol{\alpha})$.

Proof. The achievability part of the scheme is described as Algorithm 1 in Section IV, while the converse and the derivation of the gap to optimal are presented in Section V. □

One of the main conclusions from the above result is summarized in the following corollary.

Corollary 1. *In the same K-user SISO BC with γ-sized caches and (ordered) capacities $\{\alpha_k\}_{k=1}^K$, the baseline performance*

$$T(K,\gamma,\boldsymbol{\alpha}=\mathbf{1}) = T_{MN} = \frac{K(1-\gamma)}{1+K\gamma} \quad (9)$$

associated to the ideal case $\alpha_k = 1\ \forall k\in[K]$, can be achieved even if the capacities satisfy the inequalities

$$\alpha_k \ge \alpha_{th,k} \triangleq 1 - \frac{\binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1}} \approx 1 - e^{-k\gamma}, \quad \forall k\in[K]. \quad (10)$$

Proof. The proof is direct from Eq. (8), after using the Sterling approximation $\binom{n}{k} \approx \left(\frac{n}{k}\right)^k$ and the limit

$$\lim_{K\to\infty}\left(1-\frac{b}{K}\right)^K = e^{-b}. \quad (11)$$

□

Given any user k, $\alpha_{th,k} = 1 - \frac{\binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1}}$ provides a threshold channel capacity that allows the algorithm to achieve the baseline unit-capacity performance T_{MN}.

IV. Placement and Delivery Algorithms

We here present the superposition-based communication scheme with the corresponding cache placement, transmission, and decoding that achieves the delay in Theorem 1.

[2]This is a well known expression that has been calculated in a variety of works such as in [13], [24].

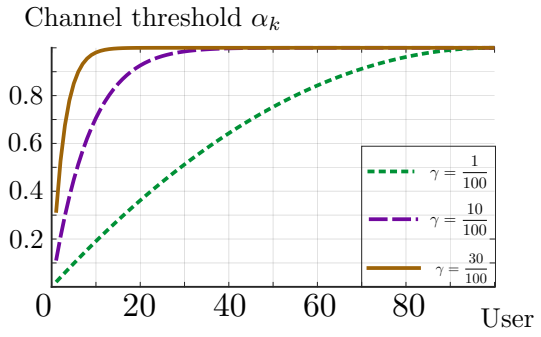


Fig. 1. The plot presents the threshold $\alpha_{th,k}$ for the case of $K = 100$ users. We can see that as γ decreases, an ever increasing fraction of users can have a further reduced channel capacity without any performance degradation with respect to the maximal-capacity delay.

A. Cache Placement

During the placement phase, we apply directly the placement algorithm of [1] without exploiting any knowledge of the channel capacities. To this end, each file $W^n, n\in[N]$, is subpacketized into $S = \binom{K}{K\gamma}$ subfiles

$$W^n \to \{W^n_\tau, \tau\subset[K],\ |\tau| = K\gamma\} \quad (12)$$

and the cache $\mathcal{Z}_k$ of user k is filled as

$$\mathcal{Z}_k = \{W^n_\tau : \tau\subset[K], \forall n\in[N]\} \quad (13)$$

which, as can easily be shown, adheres to the cache-size constraint.

B. Delivery Algorithm

After each user $k\in[K]$ requests a file W^{d_k} as in [1], the transmitter delivers the $\binom{K}{K\gamma+1}$ XORs

$$X_\sigma = \bigoplus_{k\in\sigma} W^{d_k}_{\sigma\setminus\{k\}} \quad (14)$$

for all subsets σ of users of size $|\sigma| = K\gamma+1$. To this end, in every communication slot, we split the available transmission power into $K - K\gamma - 1$ "power layers". In power layer k we encode XORs from the set

$$\mathcal{X}_k \triangleq \{X_\sigma : \min\{\sigma\} = k\}. \quad (15)$$

This contains all the XORs intended for set of users σ for which the slowest user is user k i.e., all the XORs intended for user k except those desired by any user whose channel is weaker than user k. It can be easily shown[3] that the sets $\mathcal{X}_k$ are disjoint; that for any $k \le K - K\gamma - 1$, we have

$$|\mathcal{X}_k| = \binom{K-k+1}{K\gamma+1} - \binom{K-k}{K\gamma+1} = \binom{K-k}{K\gamma} \quad (16)$$

XOR messages in power layer k and that the total number of XOR messages in the first k power layers is

$$\left|\bigcup_{m=1}^{k}\mathcal{X}_m\right| = \binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}. \quad (17)$$

[3]The last equality follows directly from Pascal's triangle.

For example, Layer 1 (which will correspond to the highest-powered layer) contains all the XORs in set $\mathcal{X}_1$ i.e., all the XORs that are intended for the weakest user (user 1). Similarly Layer 2 will contain the XORs from $\mathcal{X}_2$, i.e., those XORs that are intended for user 2, but not for user 1, and so on. The power allocation for each XOR is designed so that the weakest user of the XOR can decode it, implying that any other user that needs to decode that same XOR is able to do so. The chosen power allocation seeks to minimize the overall delay.

Algorithm 1: Delivery based on Superposition Coding

1 Let $\alpha_k \leq \alpha_{k+1},\ \forall k \in [K]$

2 Find $w \in [K]$ such that

$$w = \arg\max_{k\in[K]} \left\{ \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\alpha_k} \right\}. \quad (18)$$

3 Set $\beta_0 = 0$ and for $k \in [K - K\gamma - 1]$ set

$$\beta_k = \frac{|\cup_{i=1}^{k} \mathcal{X}_k|}{|\cup_{i=1}^{w} \mathcal{X}_k|}\alpha_w = \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w. \quad (19)$$

for *all* $k \in [K - K\gamma - 1]$ **do**

4 Encode x_k selected from $\mathcal{X}_k$ without replacement

5 with power

$$P_k = P^{-\beta_{k-1}} - P^{-\beta_k} \quad (20)$$

6 and rate

$$r_k = \beta_k - \beta_{k-1} = \frac{\binom{K-k}{K\gamma}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w. \quad (21)$$

7 Transmit $x_k, \forall k \in [K]$ simultaneously.

The process is described in the form of pseudo-code in Algorithm 1. The algorithm begins by identifying (Step 2) the bottleneck user

$$w = \arg\max_{k\in[K]} \left\{ \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\alpha_k} \right\}. \quad (22)$$

This is defined as the user k that takes the longest time to decode all power layers from 1 to k. Then Step 3 calculates the power layer coefficients $\beta_i, i \in \{0, 1, .., K - K\gamma - 1\}$ for each power layer as explained below. In Step 4, for every $k \in [K - K\gamma - 1]$, a new XOR is selected from set $\mathcal{X}_k$, and is encoded in message x_k, with power $P_k = P^{-\beta_{k-1}} - P^{-\beta_k}$ (Step 5) and rate $\frac{\binom{K-k}{K\gamma}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w$ (Step 6). Finally in Step 7 all the $x_k, \forall k \in [K]$ are transmitted simultaneously using superposition coding.

C. Decoding

In the received signal

$$y_k = h_k\sqrt{P^{\alpha_k}} \sum_{m_1=1}^{k} x_{m_1} + h_k\sqrt{P^{\alpha_k}} \sum_{m_2=k+1}^{K-K\gamma-1} x_{m_2} \quad (23)$$

at user $k \in [K]$, the second term $\sum_{m_2=k+1}^{K-K\gamma-1} x_{m_2}$ contains the lower power layers, which carry no valuable information for user k and are treated as noise. This part of the message is transmitted with power $P^{-\beta_k}$, where $\beta_k = \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w$. Due to the power and rate allocation for each of these messages (cf. Eq. (20) and Eq. (21)), using successive interference cancellation[4] (SIC), receiver k can decode the first term that encodes the messages that potentially contain information that is valuable for user k.

D. Delay Calculation

The total delay of the scheme is

$$T_{sc}(K, \gamma, \boldsymbol{\alpha}) = \max_{k\in[K-K\gamma-1]} \left\{ \frac{|\mathcal{X}_k|}{\binom{K}{K\gamma}} \cdot \frac{1}{r_k} \right\} \quad (24)$$

$$= \frac{1}{\alpha_w} \cdot \frac{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}}. \quad (25)$$

This corresponds to the maximum delay required to deliver all XORs $X_\sigma \in \mathcal{X}_k$ across all values of $k \in [K - K\gamma - 1]$.

V. Converse and Gap to Optimality

In this section, we provide a lower bound on the optimal delay for any given set of parameters $K, \gamma, \boldsymbol{\alpha}$, and then we prove that the achievable delay $T_{sc} \triangleq T_{sc}(K, \gamma, \boldsymbol{\alpha})$ from Theorem 1 is within a factor of at most 4 from the optimal delay $T^*(k, \gamma, \boldsymbol{\alpha})$.

To lower bound the minimum delay $T^*(k, \gamma, \boldsymbol{\alpha})$, we consider an augmented system where the capacities of the first w users, with w selected as (18), are increased to $\alpha_k = \alpha_w \triangleq \alpha$, for all $k \in [w]$, while the capacities of the remaining users are increased to 1. For this system, the delay is lower bounded as

$$T_{\text{aug}} \geq \overbrace{\frac{1}{\alpha}}^{t_1} \overbrace{\frac{1}{2}\frac{w(1-\gamma)}{1+w\gamma}}^{t_2}, \quad (26)$$

where term t_1 corresponds to the channel capacity of the first w users, while term t_2 corresponds to a lower bound on the minimum possible worst-case delivery time[5] associated to a system with w cache-aided users (cf. [36]).

To bound the ratio T_{sc}/T_{aug}, we first consider the case of $w\gamma < 1$ for which we have the inequalities

$$\frac{T_{sc}}{T_{\text{aug}}} \leq \frac{\frac{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}}}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \leq \frac{w(1-\gamma)}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \leq 4, \quad (27)$$

where we used the inequality $\frac{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}} \leq w(1-\gamma)$ which we prove in the online version of this work [37].

[4] In successive interference cancellation, a user first decodes the highest powered message by treating the remaining messages as noise, then proceeds to remove this – known at this point – message and decodes the second message by treating the remaining as noise, and so on until all messages have been decoded.

[5] In fact, as we know from [36], this factor is slightly smaller than $\frac{1}{2}$.

When $w\gamma \geq 1$, the bound – after a few basic algebraic manipulations – takes the form

$$\frac{T_{sc}}{T_e} = \frac{\frac{\binom{K}{K\gamma+1}-\binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}}}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \leq \frac{\frac{\binom{K}{K\gamma+1}}{\binom{K}{K\gamma}}}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \tag{28}$$

$$= \frac{\frac{K(1-\gamma)}{1+K\gamma}}{\frac{1}{2}\frac{w(1-\gamma)}{1+w\gamma}} = 2\frac{K(1+w\gamma)}{w(1+K\gamma)} = 2+2\frac{K-w}{w+Kw\gamma} \tag{29}$$

$$< 2\left(1+\frac{K+w}{w+wK\gamma}\right) < 2\left(1+\frac{K}{wK\gamma}\right) \leq 4, \tag{30}$$

which concludes the proof.

VI. Conclusions and Ramifications

In this work, we studied a cache-aided SISO BC in which users have different channel capacities. This model is motivated by the well-known worst-user bottleneck of coded caching, which, when left untreated, can severely deteriorate coded caching gains. The new algorithm establishes, together with the converse, the fundamental limits of performance within a factor of 4, revealing that it is in fact possible to achieve the full-capacity performance even in the presence of many users with degraded link strengths.

Pivotal to our approach is the identification of a 'bottleneck (threshold) user', which may not necessarily be the user with the worst channel. From an operational point of view, this reveals that to increase performance, we must not necessarily focus on enhancing only the weakest users, but rather should focus on altering this bottleneck threshold.

References

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. on Inf. Theory*, vol. 60, pp. 2856–2867, May 2014.

[2] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. on Inf. Theory*, vol. 62, no. 10, pp. 5524–5537, 2016.

[3] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communication (JSAC), Special Issue on Caching*, June 2018.

[4] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," *IEEE Transactions on Information Theory*, vol. 63, pp. 5821–5833, Sep. 2017.

[5] L. Tang and A. Ramamoorthy, "Coded caching schemes with reduced subpacketization from linear block codes," *IEEE Transactions on Information Theory*, vol. 64, pp. 3099–3120, April 2018.

[6] C. Shangguan, Y. Zhang, and G. Ge, "Centralized coded caching schemes: A hypergraph theoretical approach," *IEEE Transactions on Information Theory*, vol. 64, pp. 5755–5766, Aug 2018.

[7] H. H. S. C and P. Krishnan, "Low subpacketization coded caching via projective geometry for broadcast and D2D networks," *CoRR*, vol. abs/1902.08041, 2019.

[8] X. Zhang and M. Ji, "A new design framework on device-to-device coded caching with optimal rate and significantly less subpacketizations," *CoRR*, vol. abs/1901.07057, 2019.

[9] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Coded caching for heterogeneous systems: An optimization perspective," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[10] B. Asadi, L. Ong, and S. J. Johnson, "Centralized caching with unequal cache sizes," in *IEEE Inf. Theory Workshop (ITW)*, Nov 2018.

[11] M. Mohammadi Amiri, Q. Yang, and D. Gündüz, "Decentralized caching and coded delivery with distinct cache capacities," *IEEE Transactions on Communications*, vol. 65, pp. 4657–4669, Nov 2017.

[12] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *IEEE Information Theory Workshop (ITW)*, 2018.

[13] L. Zheng, Z. Wang, Q. Yan, Q. Chen, and X. Tang, "On the coded caching based wireless video transmission scheme," in *IEEE/CIC Inter. Conf. on Comm. in China (ICCC)*, pp. 1–6, July 2016.

[14] J. Zhang and P. Elia, "Wireless coded caching: A topological perspective," in *IEEE Int. Symp. on Inf. Theory (ISIT)*, June 2017.

[15] K. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 548–562, Jan 2018.

[16] A. Destounis, M. Kobayashi, G. Paschos, and A. Ghorbel, "Alpha fair coded caching," in *15th International Symp. on Modeling and Opt. in Mobile, Ad Hoc, and Wireless Networkss (WiOpt)*, pp. 1–8, May 2017.

[17] A. Ghorbel, M. Kobayashi, and S. Yang, "Content delivery in erasure broadcast channels with cache and feedback," *IEEE Transactions on Information Theory*, vol. 62, pp. 6407–6422, Nov 2016.

[18] M. Mohammadi Amiri and D. Gündüz, "Cache-aided content delivery over erasure broadcast channels," *IEEE Transactions on Communications*, vol. 66, pp. 370–381, Jan 2018.

[19] S. Kamel, M. Sarkiss, and M. Wigger, "Decentralized joint cache-channel coding over erasure broadcast channels," in *IEEE Middle East and North Africa Comm. Conf. (MENACOMM)*, pp. 1–6, April 2018.

[20] S. Kim, S. Mohajer, and C. Suh, "Coding across heterogeneous parallel erasure broadcast channels is useful," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1883–1887, June 2017.

[21] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *IEEE Transactions on Information Theory*, vol. 64, pp. 6996–7016, Nov 2018.

[22] E. Lampiris, J. Zhang, and P. Elia, "Cache-aided cooperation with no CSIT," in *IEEE Int. Symp. on Inf. Theory (ISIT)*, June 2017.

[23] E. Piovano, H. Joudeh, and B. Clerckx, "Generalized degrees of freedom of the symmetric cache-aided MISO broadcast channel with partial CSIT," *IEEE Transactions on Information Theory*, pp. 1–1, 2019.

[24] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Transactions on Information Theory*, vol. 65, pp. 2792–2807, May 2019.

[25] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multi-antenna interference management for coded caching," *arXiv preprint arXiv:1711.03364*, 2017.

[26] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multicast beamformer design for coded caching," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1914–1918, June 2018.

[27] J. Zhao, M. M. Amiri, and D. Gündüz, "A low-complexity cache-aided multi-antenna content delivery scheme," in *IEEE Int. Workshop on Signal Processing Advances in Wireless Comm. (SPAWC)*, July 2019.

[28] M. Salehi, A. Tölli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-rate trade-off in multi-antenna coded caching," *arXiv preprint arXiv:1905.04349*, 2019.

[29] M. M. Amiri and D. Gündüz, "Caching and coded delivery over gaussian broadcast channels for energy efficiency," *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 1706–1720, Aug 2018.

[30] I. Bergel and S. Mohajer, "Cache-aided communications with multiple antennas at finite SNR," *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 1682–1691, Aug 2018.

[31] S. A. Jafar and S. Vishwanath, "Generalized Degrees of Freedom of the symmetric Gaussian K user Interference Channel," *IEEE Transactions on Information Theory*, vol. 56, pp. 3297–3303, July 2010.

[32] A. Gholami Davoodi and S. A. Jafar, "Aligned image sets under channel uncertainty: Settling conjectures on the collapse of degrees of freedom under finite precision CSIT," *IEEE Transactions on Information Theory*, vol. 62, pp. 5603–5618, Oct 2016.

[33] A. Gholami Davoodi and S. A. Jafar, "Generalized degrees of freedom of the symmetric K user interference channel under finite precision CSIT," *IEEE Trans. Inf. Theory*, vol. 63, pp. 6561–6572, Oct 2017.

[34] E. Lampiris and P. Elia, "Bridging two extremes: Multi-antenna coded caching with reduced subpacketization and CSIT," in *IEEE Int. Workshop on Signal Processing Advances in Wireless Comm. (SPAWC)*, 2019.

[35] E. Lampiris and P. Elia, "Achieving full multiplexing and unbounded caching gains with bounded feedback resources," *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[36] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," *IEEE Transactions on Information Theory*, vol. 65, pp. 647–663, Jan 2019.

[37] E. Lampiris, J. Zhang, O. Simeone, and P. Elia, "Fundamental limits of wireless caching under uneven-capacity channels," *arXiv preprint arXiv:1908.04036*, 2019.

Efficient Error Probability Simulation of Coded Modulation over Fading Channels

Josep Font-Segura
Universitat Pompeu Fabra
josep.font@ieee.org

Alfonso Martinez
Universitat Pompeu Fabra
alfonso.martinez@ieee.org

Albert Guillén i Fàbregas
Universitat Pompeu Fabra
ICREA and University of Cambridge
guillen@ieee.org

***Abstract*—We use importance sampling to estimate the random-coding union (RCU) bound to the achievable error probability in coded-modulation wireless channels. We provide closed-form expressions of the exponentially-tilted distributions to generate the required samples, and illustrate the technique for coded BPSK modulation over the i.i.d. Rayleigh fading channel.**

I. Introduction

Evaluating the error probability of the transmission of coded data over a continuous-output channel is a common problem in digital communications. Efficient simulation methods of high-performance codes were proposed in, e. g., [1] for low density parity check (LDPC) codes. Together with other powerful codes such as polar codes and turbo codes, LDPC codes assume large code lengths. This assumption is yet not compatible with the ultra-high reliability and low latency requirements for next-generation wireless systems.

Instead of considering a good code, we study the random-coding union (RCU) bound to the achievable error probability [2, Eq. (62)]. Let $\boldsymbol{x}$ denote a transmitted codeword of length n drawn from a constellation $\mathcal{X}$, and let $\boldsymbol{y}$ be the received sequence taking values over $\mathbb{C}^n$. Random-coding arguments show the existence of a code of M codewords, transmitted over a memoryless channel with conditional density $W(y|x)$, whose error probability, the probability of decoding in favor of the wrong codeword, is at most the RCU, given by

$$\mathrm{rcu}_n = \int Q^n(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x}) \min\big\{1, (M-1)\mathrm{pep}_n(\boldsymbol{x},\boldsymbol{y})\big\}\,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}, \tag{1}$$

where the pairwise error probability $\mathrm{pep}_n(\boldsymbol{x},\boldsymbol{y})$ reads

$$\mathrm{pep}_n(\boldsymbol{x},\boldsymbol{y}) = \int Q^n(\overline{\boldsymbol{x}})\mathbb{1}\{W^n(\boldsymbol{y}|\overline{\boldsymbol{x}}) \geq W^n(\boldsymbol{y}|\boldsymbol{x})\}\,\mathrm{d}\overline{\boldsymbol{x}}, \tag{2}$$

and $\mathbb{1}\{\cdot\}$ is the indicator function. The expressions (1) and (2) are expectations with respect to the joint probability density

$$Q^n(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})Q^n(\overline{\boldsymbol{x}}). \tag{3}$$

The exact computation of the RCU bound is cumbersome even for simple channels and moderate values of n. Instead of resorting to approximations (e.g., [3]–[5]), we explore fast and accurate simulation to estimate (1).

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 TEC2016-78434-C3-1-R.

II. Importance Sampling

We first note that both expectations (1) and (2) can be cast as follows. Let $f(\boldsymbol{z})$ be a non-negative function of some random variable $\boldsymbol{Z}$ with density $P(\boldsymbol{z})$. The standard Monte Carlo estimate of a quantity $p_n = \mathrm{E}\big[f(\boldsymbol{Z})\big]$ involves drawing N samples $\boldsymbol{z}_i$ from $P(\boldsymbol{z})$ and computing the average

$$\hat{p}_{n,N} = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{z}_i). \tag{4}$$

The Monte Carlo estimator (4) is unbiased as its expected value satisfies $\mathrm{E}[\hat{p}_{n,N}] = p_n$. Besides, when $f(\boldsymbol{z})$ in (4) is an indicator function, it can be inferred that the number of samples needed to estimate p_n to a given accuracy level grows as $N \propto p_n^{-1}$, [6, Sec. 4.1]. Since the RCU bound decays exponentially with the codeword length n, this implies an exponential growth in the required number of samples.

Alternatively, importance sampling was proposed in [7] to diminish the sampling size in estimating the error probability of a communication scheme. Instead of estimating p_n as (4), this variance-reducing method involves generating i.i.d. samples from another distribution $\bar{P}(\boldsymbol{z})$ [7] to estimate p_n as

$$\hat{p}_{n,N} = \frac{1}{N}\sum_{i=1}^{N} \omega(\boldsymbol{z}_i)f(\boldsymbol{z}_i), \tag{5}$$

where the weights $\omega(\boldsymbol{z})$ that account for the distribution mismatch are given by the ratio $\omega(\boldsymbol{z}) = P(\boldsymbol{z})/\bar{P}(\boldsymbol{z})$.

A good choice for $\bar{P}(\boldsymbol{z})$ is known to be the exponential tilting [6] that exploits the exponential decay of p_n. For any value $s \geq 0$ and a function $g_n(\boldsymbol{z})$, we define the exponentially-tilted distribution

$$\bar{P}_{s,g}(\boldsymbol{z}) = P(\boldsymbol{z})e^{sg_n(\boldsymbol{z})-\kappa_n(s)} \tag{6}$$

in terms of the cumulant generating function [8] of $g_n(\boldsymbol{z})$,

$$\kappa_n(s) = \log \mathrm{E}\big[e^{sg_n(\boldsymbol{Z})}\big]. \tag{7}$$

The importance-sampling estimator (5) then becomes

$$\hat{p}_{n,N} = \hat{\alpha}_{n,N}(s) \cdot e^{\kappa_n(s)}, \tag{8}$$

where

$$\hat{\alpha}_{n,N}(s) = \frac{1}{N}\sum_{i=1}^{N} e^{-sg_n(\boldsymbol{z}_i)}f(\boldsymbol{z}_i) \tag{9}$$

and the samples $\boldsymbol{z}_i$ are independently drawn from $\bar{P}_{s,g}(\boldsymbol{z})$.

Roughy speaking, the importance-sampling estimator approximates the pre-exponential factor α_n in the quantity $p_n = \alpha_n(s) \cdot e^{\kappa_n(s)}$ by $\hat{\alpha}_{n,N}$, instead of directly estimating p_n. The importance-sampling estimator (8) is also unbiased [6, Sec. 4.2] with a normalized sample variance

$$\sigma_n^2 = \frac{\mathrm{E}\big[e^{\kappa_n(s)-sg_n(\boldsymbol{Z})}f(\boldsymbol{Z})^2\big] - p_n^2}{p_n^2} \tag{10}$$

that is now reduced by properly choosing the parameters involved in the exponential tilting, namely $s \geq 0$ and $g_n(\boldsymbol{z})$. A good choice of s is the minimizer

$$\hat{s}_n = \arg\min_{s\geq 0} \kappa_n(s), \tag{11}$$

whereas the choice of $g_n(\boldsymbol{z})$ depends on the structure of $f(\boldsymbol{z})$. We next apply the exponentially-tilted importance-sampling method described in this section to estimate (1).

III. Error Probability Estimation

We first note that for a fixed transmitted codeword $\boldsymbol{x}$ and received sequence $\boldsymbol{y}$, a nested estimator of the pairwise error probability (2) is needed. A good choice of $g_n(\overline{\boldsymbol{x}})$ for the importance-sampling estimate of the pairwise error probability in (2) with integration variable $\overline{\boldsymbol{x}}$ is the log-likelihood ratio

$$\ell_n(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{x}}) = \log \frac{W^n(\boldsymbol{y}|\overline{\boldsymbol{x}})}{W^n(\boldsymbol{y}|\boldsymbol{x})}. \tag{12}$$

As stated later, this choice helps capturing the correct exponential decay of the pairwise error probability in terms of n. The cumulant generating function of $\ell_n(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{X}})$ is given by

$$\kappa_{n,\tau}(\boldsymbol{x},\boldsymbol{y}) = \log \mathrm{E}\big[e^{\tau\cdot\ell_n(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{X}})}\big] \tag{13}$$

and leads to the following tilted distribution $\bar{P}_\tau(\overline{\boldsymbol{x}}|\boldsymbol{y})$ in (6) for the estimation of $\mathrm{pep}_n(\boldsymbol{x},\boldsymbol{y})$

$$\bar{P}_\tau^n(\overline{\boldsymbol{x}}|\boldsymbol{y}) = \frac{1}{\mu_n(\boldsymbol{y})} Q^n(\overline{\boldsymbol{x}}) W^n(\boldsymbol{y}|\overline{\boldsymbol{x}})^\tau, \tag{14}$$

where $\mu_n(\boldsymbol{y})$ is a normalizing factor. We remark that while the log-likelihood $\ell_n(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{x}})$ depends on the transmitted codeword $\boldsymbol{x}$, the conditional distribution (14) for the codeword $\overline{\boldsymbol{x}}$ depends only on the received sequence $\boldsymbol{y}$ through the tilted channel density $W(\boldsymbol{y}|\overline{\boldsymbol{x}})^\tau$.

The importance-sampling estimator of the pairwise error probability generates N_1 independent samples $\overline{\boldsymbol{x}}_j$ from the conditional probability distribution (14), computes the average

$$\hat{\gamma}_{\tau,N_1}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{N_1}\sum_{j=1}^{N_1} e^{-\tau\cdot\ell_n(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{x}}_j)} f_{\mathrm{pep}}(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{x}}_j), \tag{15}$$

where we defined

$$f_{\mathrm{pep}}(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{x}}) = \mathbb{1}\big\{\ell_n(\boldsymbol{x},\boldsymbol{y},\overline{\boldsymbol{x}}) \geq 0\big\}, \tag{16}$$

and finally obtains the final estimate

$$\hat{\mathrm{pep}}_{n,N_1}(\boldsymbol{x},\boldsymbol{y}) = \hat{\gamma}_{\tau,N_1}(\boldsymbol{x},\boldsymbol{y}) \cdot e^{\kappa_{n,\tau}(\boldsymbol{x},\boldsymbol{y})}. \tag{17}$$

The tilting parameter τ is chosen as $\tau = \hat{\tau}_n(\boldsymbol{x},\boldsymbol{y})$, where

$$\hat{\tau}_n(\boldsymbol{x},\boldsymbol{y}) = \arg\min_{\tau\geq 0} \kappa_{n,\tau}(\boldsymbol{x},\boldsymbol{y}). \tag{18}$$

Note that τ used in the function $\kappa_{n,\tau}(\boldsymbol{x},\boldsymbol{y})$ depends on both $\boldsymbol{x}$ and $\boldsymbol{y}$. Yet, we drop the dependence on $\boldsymbol{x},\boldsymbol{y}$ in $\hat{\tau}_n$ to lighten the notation. Basic results in large-deviation theory imply that for memoryless channels the pairwise error probability (2) behaves exponentially as

$$\lim_{n\to\infty} \frac{\log \mathrm{pep}_n(\boldsymbol{x},\boldsymbol{y})}{\kappa_{n,\hat{\tau}_n}(\boldsymbol{x},\boldsymbol{y})} = 1. \tag{19}$$

We now address the importance-sampling estimate of the random-coding union bound in (1), an expectation with respect to the integration variables $\boldsymbol{x}$ and $\boldsymbol{y}$. In this case, we select the random variable

$$g_n(\boldsymbol{x},\boldsymbol{y}) = \log(M-1) + \kappa_{n,\frac{1}{1+\rho}}(\boldsymbol{x},\boldsymbol{y}) \tag{20}$$

because its cumulant generating function, given by

$$\chi_n(\rho) = \log \mathrm{E}\left[(M-1)^\rho \left(\frac{\mathrm{E}\big[W^n(\boldsymbol{Y}|\overline{\boldsymbol{X}})^{\frac{1}{1+\rho}}|\boldsymbol{Y}\big]}{W^n(\boldsymbol{Y}|\boldsymbol{X})^{\frac{1}{1+\rho}}}\right)^\rho\right], \tag{21}$$

gives the random-coding exponent [9, Sec. 5.6]. As a result, we will restrict the parameter ρ in the $[0,1]$ interval. Using (6), every pair of samples $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ is drawn from

$$\bar{P}_\rho^n(\boldsymbol{x},\boldsymbol{y}) = Q^n(\boldsymbol{x})\bar{W}_\rho^n(\boldsymbol{y}|\boldsymbol{x}), \tag{22}$$

where $\bar{W}_\rho^n(\boldsymbol{y}|\boldsymbol{x})$ is the tilted channel density given by

$$\bar{W}_\rho^n(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{\mu_n} W^n(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{1+\rho}} \Big(\mathrm{E}[W^n(\boldsymbol{y}|\overline{\boldsymbol{X}})^{\frac{1}{1+\rho}}]\Big)^\rho \tag{23}$$

with normalizing factor μ_n. Inspecting (22), we observe that the transmitted codewords $\boldsymbol{x}_i$ are generated with the original random coding distribution $Q^n(\boldsymbol{x})$, whereas the received sequences $\boldsymbol{y}_i$ are drawn from the modified channel transition probability (23).

The importance-sampling estimator for the RCU bound (1) based on the independently generated pairs of samples $\boldsymbol{x}_i,\boldsymbol{y}_i$ from the probability distribution (22) is given by

$$\hat{\mathrm{rcu}}_{n,N_1,N_2} = \hat{\alpha}_{n,N_1,N_2}(\rho) \cdot e^{\chi_n(\rho)}, \tag{24}$$

where the pre-factor estimate reads

$$\hat{\alpha}_{n,N_1,N_2}(\rho) = \frac{1}{N_2}\sum_{i=1}^{N_2} e^{-\rho\cdot g_n(\boldsymbol{x}_i,\boldsymbol{y}_i)} f_{\mathrm{rcu}}(\boldsymbol{x}_i,\boldsymbol{y}_i) \tag{25}$$

with $f_{\mathrm{rcu}}(\boldsymbol{x}_i,\boldsymbol{y}_i)$ a function that depends on the pairwise error probability estimate (17) as

$$f_{\mathrm{rcu}}(\boldsymbol{x},\boldsymbol{y}) = \min\{1,(M-1)\hat{\mathrm{pep}}_{n,N_1}(\boldsymbol{x},\boldsymbol{y})\}. \tag{26}$$

For choice of

$$\hat{\rho}_n = \arg\min_{0\leq\rho\leq 1} \chi_n(\rho), \tag{27}$$

it follows from basic results in large-deviation theory that

$$\lim_{n\to\infty} \frac{\log \mathrm{rcu}_n}{\chi_n(\hat{\rho}_n)} = 1. \tag{28}$$

In summary, we proposed an importance-sampling estimator for the RCU bound (1) built from two nested estimators. Transmitted codewords $\boldsymbol{x}$ are drawn from the original random-coding distribution and received sequences $\boldsymbol{y}$ are generated from the modified channel transition probability (23) with optimal tilting parameter $\hat{\rho}_n$ related to the random-coding error exponent (28). For a given transmitted codeword and received sequence, the pairwise codewords $\overline{\boldsymbol{x}}$ are generated independently from $\boldsymbol{x}$ but conditioned on $\boldsymbol{y}$ from the conditional distribution (14) with optimal tilting parameter $\hat{\tau}_n$ related to the exponential decay of the pairwise error probability (19).

We remark that (14) and (23) might not be standard probability distributions. Yet, samples can be efficiently generated using, e. g., the rejection method described in [10, Ch. II.3].

We finally briefly discuss the performance analysis of the proposed importance-sampling estimator. We observe that $\hat{\mathrm{rcu}}_{n,N_1,N_2}$ is the sum of N_2 independent terms, each of them a nonlinear function of the inner estimator $\hat{\mathrm{pep}}_{n,N_1}(\boldsymbol{x}_i,\boldsymbol{y}_i)$ that is also the sum of N_1 independent terms. Using refined central-limits theorems and Taylor expansions in inverse powers of N_1 and N_2, we show in [11] that for memoryless channels and sufficiently large code length n, as both N_1 and N_2 tend to infinity the importance-sampling estimator (24) converges in probability[1] to the exact RCU bound rcu_n according to

$$\hat{\mathrm{rcu}}_{n,N_1,N_2} \xrightarrow[N_1,N_2\to\infty]{\mathrm{p}} \mathrm{rcu}_n\left(1-\frac{k_{1,n}}{N_1}+\sqrt{\frac{k_{2,n}}{N_2}}\Theta\right), \quad (30)$$

where $k_{1,n}$ and $k_{2,n}$ are positive numbers growing with n as $O(\sqrt{n})$, and Θ is the standard normal random variable.

Since $k_{1,n}$ in (30) is a positive term and Θ is a zero-mean random variable, it implies a negative bias in the estimation of the RCU bound. Yet, the estimator is consistent, as the bias vanishes as N_1 goes to infinity, although the bias might be significant for small values of N_1. The variance term $k_{2,n}$ in (30) grows as the squared root of n, implying a significant reduction in the variance with the importance-sampling estimator, as the number of samples needed to accurately estimate the RCU bound for a given confidence level grows as $N_2 \propto \sqrt{n}$, rather than the typical growth $N_2 \propto \mathrm{rcu}_n{}^{-1}$ in standard Monte Carlo [6, Sec. 4.1], which would be exponential in the code length n in our setting of a memoryless channel.

IV. Numerical Example

We illustrate the above importance-sampling estimator of the RCU bound for the binary phase-shift keying (BPSK) modulation. We denote the symbol set $\mathcal{X}=\{-\sqrt{P},+\sqrt{P}\}$, where P is a positive number describing an average power constraint. A codeword $\boldsymbol{x}=(x_1,\ldots,x_n)$ is transmitted over the i.i.d. Rayleigh fading channel described by

$$y_i = h_i x_i + w_i, \quad (31)$$

where $\boldsymbol{y}=(y_1,\ldots,y_n)$ is the received sequence, $\boldsymbol{w}=(w_1,\ldots,w_n)$ is an i.i.d. real-valued zero-mean Gaussian noise with variance σ^2. Since the phase of the fading coefficients is irrelevant, we assume that $\boldsymbol{h}=(h_1,\ldots,h_n)$ is a real-valued i.i.d. Rayleigh distributed with density

$$p^n(\boldsymbol{h}) = \prod_{i=1}^{n} 2h_i e^{-h_i^2} \mathbb{1}\{h_i \geq 0\}. \quad (32)$$

The symmetry of BPSK implies that the input distribution $Q^n(\boldsymbol{x})$ that optimizes both the exponential decay (28) and the channel capacity, denoted as C_b, is the uniform distribution

$$Q^n(\boldsymbol{x}) = \frac{1}{2^n}. \quad (33)$$

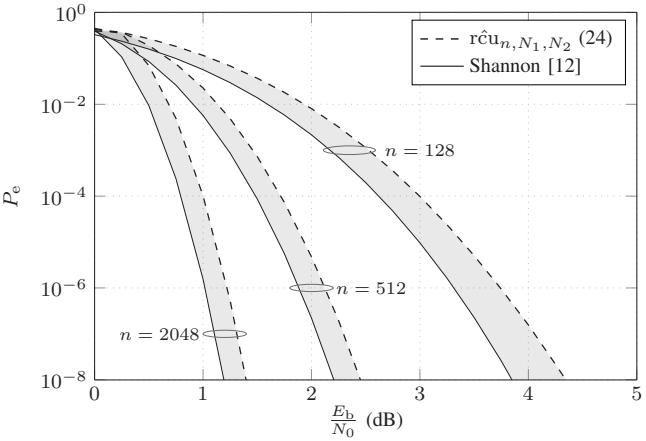


Fig. 1. Error probability versus E_b/N_0 over the AWGN channel, for code rate $R_\mathrm{b}=0.5$, $N_1=N_2=500$ samples, and several code lengths n.

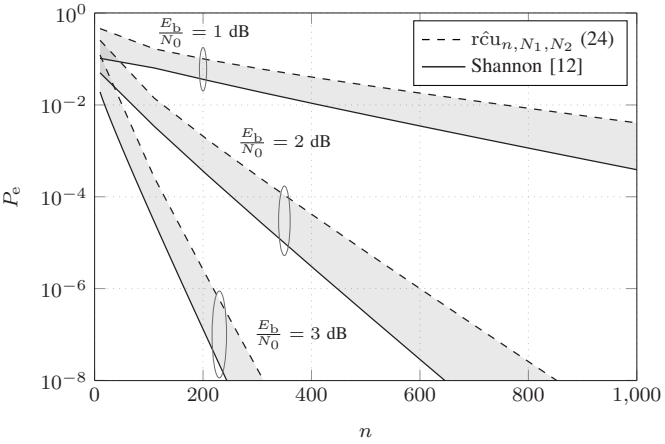


Fig. 2. Error probability versus n over the AWGN channel, for code rate $R_\mathrm{b}=0.5$, $N_1=N_2=500$ samples, and several values of E_b/N_0.

[1] Two sequences of random variables A_N and B_N indexed by N are said to converge in probability if for all $\varepsilon>0$, it holds

$$\lim_{N\to\infty} \Pr[|A_N - B_N| > \varepsilon] = 0. \quad (29)$$

We denote the convergence in probability by $A_N \xrightarrow[N\to\infty]{\mathrm{p}} B_N$.

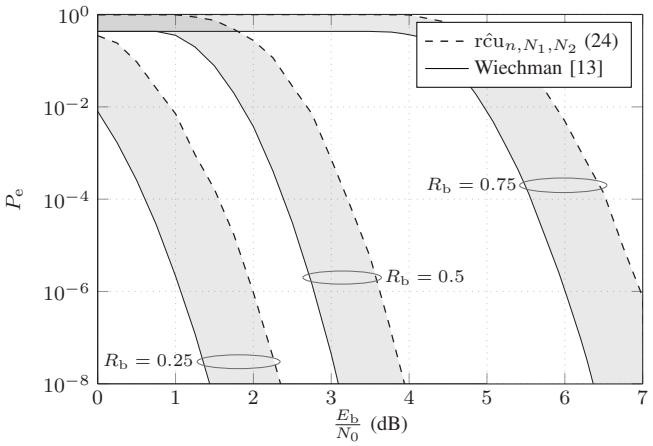


Fig. 3. Error probability versus E_b/N_0 over the i.i.d. Rayleigh channel, for code length $n = 1024$, $N_1 = N_2 = 500$ samples, and several code rates R_b.

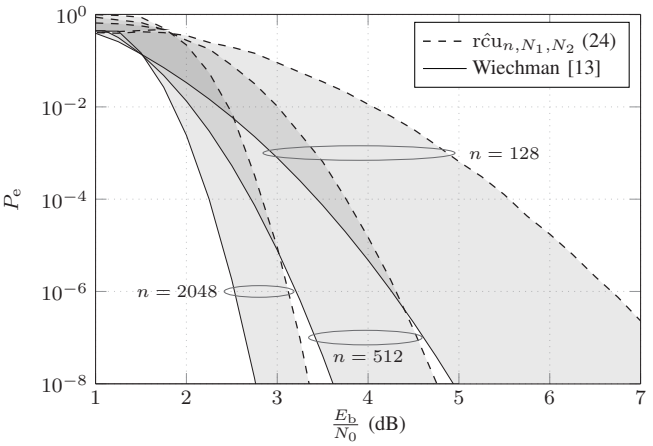


Fig. 4. Error probability versus E_b/N_0 over the i.i.d. Rayleigh channel, for code rate $R_\mathrm{b} = 0.5$, $N_1 = N_2 = 500$ samples, and several code lengths n.

The input distribution $Q^n(\boldsymbol{x})$, together with the channel conditional density given by

$$W^n(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{h}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - h_i x_i)^2}{2\sigma^2}}, \tag{34}$$

determine the required parameters for the importance-sampling estimator (24), namely the the cumulant-generating functions (13) and (21), the optimal tilting parameters $\hat{\tau}_n$ and $\hat{\rho}_n$ respectively in (18) and (27), and the tilted distributions (14) and (23). The additive white Gaussian noise (AWGN) channel can be recovered from (34) by setting

$$p^n(\boldsymbol{h}) = \delta_n(\boldsymbol{h} - \mathbf{1}) \tag{35}$$

where $\delta_n(\cdot)$ is the n-dimensional Dirac delta, and $\mathbf{1}$ is the all-ones length-n vector. As usual, we define the code rate as

$$R_\mathrm{b} = \frac{1}{n} \log_2 M, \tag{36}$$

and the coded average E_b/N_0 ratio as

$$\frac{E_\mathrm{b}}{N_0} = \frac{P}{\sigma^2} \cdot \frac{1}{2R_\mathrm{b}}. \tag{37}$$

We set $N_1 = N_2 = 500$ to estimate the achievable error probability by means of the RCU, and include Shannon's sphere-packing bound [12, Eq. (15)] for the AWGN channel or an improved sphere-packing bound [13, Th. 3.1] for the i.i.d. Rayleigh fading channel. The error probability of good binary codes must lie between the RCU and the sphere-packing bounds, as shown in Figs. 1–4 in gray-shaded regions for several configurations of codeword length n, code rate R_b and coded E_b/N_0 ratio. In the presence of fading, we observe a larger gap between achievability and converse bounds compared to the AWGN case, especially for small values of n. As another example, a performance loss of approximately 2 dB in E_b/N_0 is noticed at $n = 2048$ in Fig. 4 for the fading case when compared to the AWGN case in Fig. 1.

V. Conclusion

In this paper, we proposed an importance-sampling technique to estimate the random-coding union (RCU) bound to the achievable error probability for the transmission of coded data over a continuous-output channel. We derived closed-form expressions for the optimal tilted distributions needed to generate the samples of the two nested estimators involved, and illustrated the transmission of the coded BPSK modulation over the AWGN and i.i.d. Rayleigh fading channels.

References

[1] S. Ahn, K. Yang, and D. Har, "Evaluation of the low error-rate performance of LDPC codes over Rayleigh fading channels using importance sampling," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2166–2177, 2013.

[2] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[3] Y. Altuğ and A. B. Wagner, "Refinement of the random coding bound," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6005–6023, 2014.

[4] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, 2014.

[5] J. Honda, "Exact asymptotics for the random coding error probability," in *IEEE Int. Symp. on Inf. Theory (ISIT)*, 2015, pp. 91–95.

[6] J. Bucklew, *Introduction to Rare Event Simulation*. New York: Springer-Verlag, 2013.

[7] K. Shanmugam and P. Balaban, "A modified Monte-Carlo simulation technique for the evaluation of error rate in digital communication systems," *IEEE Trans. Commun.*, vol. 28, no. 11, pp. 1916–1924, Nov. 1980.

[8] R. Durrett, *Probability: Theory and Examples*. Belmont, CA: Duxbury, 1996.

[9] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[10] L. Devroye, *Non-Uniform Random Variate Generation*. New York: Springer-Verlag, 1986.

[11] J. Font-Segura, A. Martinez, and A. Guillén i Fàbregas, "Performance analysis of the RCU importance-sampling estimator," 2019. [Online]. Available: `itc.upf.edu/biblio/1114`

[12] C. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. Journal*, vol. 38, no. 3, pp. 611–656, May 1959.

[13] G. Wiechman and I. Sason, "An improved sphere-packing bound for finite-length codes over symmetric memoryless channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1962–1990, 2008.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

On the Error Probability of Optimal Codes in Gaussian Channels under Average Power Constraint

Gonzalo Vazquez-Vilar
Universidad Carlos III de Madrid, Spain
Email: gvazquez@ieee.org

***Abstract*—This paper studies the performance of block coding on an additive white Gaussian noise channel under different power limitations at the transmitter. Lower bounds are presented for the minimum error probability of codes satisfying an average power constraint. These bounds are tighter than previous results in the literature, and yield a better understanding on the structure of good codes under an average power limitation.**

I. Introduction

We consider the problem of transmitting equiprobable messages over several uses of an additive white Gaussian noise (AWGN) channel. We consider different power constraints: *equal power constraint* (all the codewords in the transmission code have equal energy); *maximal power constraint* (the energy of all the codewords is below a certain threshold); and *average power constraint* (while some codewords may violate the threshold, the energy budget is satisfied in average).

In his 1959 paper, Shannon derived a lower bound to the error probability of any equal power constrained codebook via geometrical arguments [1, Eq. (20)]. Following a different approach, Polyanskiy, Poor and Verdú applied a particular instance of a binary hypothesis test to lower bound the same error probability [2, Th. 41]. While [2, Th. 41] was derived originally under an equal power constraint, it was recently shown to also hold under a maximal power constraint [3, Th. 3]. Other connections among the system performance under the three power constraints are studied in [1, Sec. XIII] (see also [2, Lem. 39]).

In this work, we establish direct lower bounds for codes satisfying an average power limitation at the transmitter. Our analysis is based on the meta-converse bound [2, Th. 27] evaluated for auxiliary Gaussian distributions. We characterize the error probability of the binary hypothesis test appearing in this bound for the AWGN channel, and use its properties to avoid the optimization over input distributions. Our results show that, if the cardinality of the codebook is below a certain threshold, [2, Th. 41] and [3, Th. 3] hold under an average power limitation without any modifications. The resulting bound is tighter than previous results in the literature for the same power constraint and provide an accurate characterization of the error probability for a wide range of system parameters.

G. Vazquez-Vilar is also with the Gregorio Marañón Health Research Institute, Madrid, Spain. This work has been funded in part by the European Research Council (ERC) under grant 714161, and by the Spanish Ministry of Economy and Competitiveness under grant TEC2016-78434-C3 (AEI/FEDER, EU).

II. System Model

We consider the problem of transmitting M equiprobable messages over n uses of an AWGN channel with noise power σ^2. Specifically, we consider the channel with law $W = P_{\boldsymbol{Y}|\boldsymbol{X}}$ which, for an input $\boldsymbol{x} = (x_1, \ldots, x_n)$ and output $\boldsymbol{y} = (y_1, \ldots, y_n)$, has a probability density function (pdf)

$$w(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} \varphi_{x_i,\sigma}(y_i), \tag{1}$$

where $\varphi_{\mu,\sigma}(\cdot)$ denotes the pdf of the Gaussian distribution,

$$\varphi_{\mu,\sigma}(x) \triangleq \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2}$$

In our communications system, a source produces a certain message $v \in \{1, \ldots, M\}$ randomly with equal probability. This message is mapped by the encoder to a codeword $\boldsymbol{c}_v$ according to a codebook $\mathcal{C} \triangleq \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}$, and the sequence $\boldsymbol{x} = \boldsymbol{c}_v$ is transmitted over the channel. Then, based on the channel output $\boldsymbol{y}$, the decoder guesses the transmitted message $\hat{v} \in \{1, \ldots, M\}$. We define the average error probability

$$P_{\mathrm{e}}(\mathcal{C}) \triangleq \Pr\{\hat{V} \neq V\}, \tag{3}$$

where the underlying probability is induced by the chain of source, encoder, channel and decoder.

We consider codebooks satisfying the following constraints:

- Equal power constraint Υ:

$$\mathcal{L}_{\mathrm{e}}(n, M, \Upsilon) \triangleq \left\{\mathcal{C} \;\middle|\; \|\boldsymbol{c}_i\|^2 = n\Upsilon,\ i = 1, \ldots, M\right\} \tag{4}$$

- Maximal power constraint Υ:

$$\mathcal{L}_{\mathrm{m}}(n, M, \Upsilon) \triangleq \left\{\mathcal{C} \;\middle|\; \|\boldsymbol{c}_i\|^2 \leq n\Upsilon,\ i = 1, \ldots, M\right\} \tag{5}$$

- Average power constraint Υ:

$$\mathcal{L}_{\mathrm{a}}(n, M, \Upsilon) \triangleq \left\{\mathcal{C} \;\middle|\; \tfrac{1}{M}\sum\nolimits_{i=1}^{M} \|\boldsymbol{c}_i\|^2 \leq n\Upsilon\right\} \tag{6}$$

Clearly, $\mathcal{L}_{\mathrm{e}} \subseteq \mathcal{L}_{\mathrm{m}} \subseteq \mathcal{L}_{\mathrm{a}}$. In the following, we study lower bounds on the error probability $P_{\mathrm{e}}(\mathcal{C})$ under equal, maximal and average power constraints. While derivation of converse bounds is easier under an equal power constraint, the maximal power and average power constraints are more relevant for practical applications.

III. Meta-converse Bound for Equal and Maximal Power Constraints

In [2], Polyanskiy *et al.* proved that the error probability of a binary hypothesis test with certain parameters lower bounds the error probability $P_{\mathrm{e}}(\mathcal{C})$ for a certain channel W. In particular, [2, Th. 27] shows that

$$P_{\mathrm{e}}(\mathcal{C}) \geq \inf_{P\in\mathcal{P}} \sup_{Q} \left\{ \alpha_{\frac{1}{M}}\left(PW, P\times Q\right)\right\}, \tag{7}$$

where $\mathcal{P}$ is the set of distributions over the input alphabet $\mathcal{X}^n$ satisfying a certain constraint and Q is an auxiliary distribution over the output alphabet $\mathcal{Y}^n$ which is not allowed to depend on the input $\boldsymbol{x}$. Here $\alpha_\beta\left(A,B\right)$ denotes the minimum type-I error for a maximum type-II error $\beta\in[0,1]$ in a binary hypothesis test between the distributions A and B. Specifically, for two distributions A and B defined over an alphabet $\mathcal{Z}$, the function $\alpha_\beta\left(A,B\right)$ is given by

$$\alpha_\beta(A,B) \triangleq \inf_{\substack{0\leq T\leq 1:\\ \mathrm{E}_B[T(Z)]\leq\beta}} \Big\{1-\mathrm{E}_A[T(Z)]\Big\}, \tag{8}$$

where $T:\mathcal{Z}\to[0,1]$ and $\mathrm{E}_P[\cdot]$ is the expectation operator with respect to the random variable $Z\sim P$.

The bound (7) is usually referred to as the *meta-converse bound* since several converse results in the literature can be recovered from it via relaxation. While it is possible to restrict the set of distributions Q over which the bound is maximized and still obtain a lower bound, the minimization over P needs to be carried out over all the n-dimensional probability distributions (not necessarily product) satisfying $\mathcal{P}$.

For the Gaussian channel W, Polyanskiy *et al.* fixed Q to be zero-mean Gaussian distributed with variance θ^2 and independent entries, i.e., with pdf

$$q(\boldsymbol{y}) = \prod_{i=1}^{n} \varphi_{0,\theta}(y_i). \tag{9}$$

For this choice of Q, $\alpha_{\frac{1}{M}}(\cdot,\cdot)$ presents spherical symmetry. Then, restricting the input distribution to lie on the surface of a n-dimensional hyper-sphere of squared radius $n\Upsilon$ and setting $\theta^2=\Upsilon+\sigma^2$, they obtained the following result.

Theorem 1 (Converse, equal power constraint [2, Th. 41]): Let $\mathcal{C}\in\mathcal{L}_{\mathrm{e}}(n,M,\Upsilon)$ be a length-n code of cardinality M satisfying an equal power constraint. Then, for $\theta^2=\Upsilon+\sigma^2$,

$$P_{\mathrm{e}}(\mathcal{C}) \geq \alpha_{\frac{1}{M}}\big(\varphi^n_{\sqrt{\Upsilon},\sigma}, \varphi^n_{0,\theta}\big). \tag{10}$$

The bound in Theorem 1 can be extended to maximal and average power constraints using, e.g., [2, Lem. 39]. A direct lower bound under maximal power constraint is given next.

Theorem 2 (Converse, maximal power constraint [3, Th. 3]): Let $\mathcal{C}\in\mathcal{L}_{\mathrm{m}}(n,M,\Upsilon)$ be a length-n code of cardinality M satisfying a maximal power constraint. For any $\theta\geq\sigma$, $n\geq 1$, the lower bound (10) holds for this code.

The bounds in Theorems 1 and 2 coincide for equal and maximal power constraints. Then, one may wonder if this is also the case for codes satisfying an average power constraint. In Section IV, we will show that the lower bound (10) holds in this setting under certain conditions (but not in general).

A. Computation of $\alpha_\beta\big(\varphi^n_{\sqrt{\gamma},\sigma}, \varphi^n_{0,\theta}\big)$

Computation of Theorems 1 and 2 require to evaluate

$$f(\beta,\gamma) \triangleq \alpha_\beta\big(\varphi^n_{\sqrt{\gamma},\sigma}, \varphi^n_{0,\theta}\big). \tag{11}$$

We next provide a parametric formulation of this function.

Proposition 1: Let $\sigma,\theta>0$ and $n\geq 1$, be fixed parameters, and define $\delta \triangleq \theta^2-\sigma^2$. The trade-off between α and β admits the following parametric formulation as a function of the auxiliary parameter $t\geq 0$,

$$\alpha(\gamma,t) = Q_{\frac{n}{2}}\left(\sqrt{n\gamma}\frac{\sigma}{\delta}, \frac{t}{\sigma}\right), \tag{12}$$

$$\beta(\gamma,t) = 1-Q_{\frac{n}{2}}\left(\sqrt{n\gamma}\frac{\theta}{\delta}, \frac{t}{\theta}\right), \tag{13}$$

where $Q_m(x,y)$ denotes the generalized Marcum Q-function. Let t_b satisfy $\beta(\gamma,t_b)=b$ according to (13). Then, the function (11) is given by $f(b,\gamma)=\alpha(\gamma,t_b)$ according to (12).

Proof outline: Following the lines of the proof of [2, Th. 41], we obtain a parametric formulation in terms of two non-central χ^2 distributions. Then, to recover (12)-(13), we write the cumulative density function $F_{n,\nu}(x)$ of a non-central χ^2 distribution with n degrees of freedom and non-centrality parameter ν in terms of the generalized Marcum Q-function $Q_m(a,b)$ as $F_{n,\nu}(x)=1-Q_{\frac{n}{2}}\big(\sqrt{\nu},\sqrt{x}\big)$. ∎

In Proposition 1, we need to invert the marcum-Q function in (13) to evaluate $f(\beta,\gamma)$. The following alternative expression is more adequate for implementation purposes, as it only requires to solve a one dimensional optimization problem.

Corollary 1: Let $\sigma,\theta>0$ and $n\geq 1$, be fixed parameters. The function $f(\beta,\gamma)=\alpha_\beta\big(\varphi^n_{\sqrt{\gamma},\sigma}, \varphi^n_{0,\theta}\big)$ is given by

$$f(\beta,\gamma) = \max_{t\geq 0}\Bigg\{ Q_{\frac{n}{2}}\left(\sqrt{n\gamma}\frac{\sigma}{\delta}, \frac{t}{\sigma}\right) + \frac{\theta^n}{\sigma^n} e^{\frac{1}{2}\left(\frac{n\gamma}{\delta}-\frac{\delta t^2}{\sigma^2\theta^2}\right)} \times \left(1-\beta-Q_{\frac{n}{2}}\left(\sqrt{n\gamma}\frac{\theta}{\delta}, \frac{t}{\theta}\right)\right)\Bigg\}. \tag{14}$$

Proof outline: We define

$$j(\boldsymbol{y}) \triangleq \log\frac{\varphi^n_{\sqrt{\gamma},\sigma}(\boldsymbol{y})}{\varphi^n_{0,\theta}(\boldsymbol{y})} \tag{15}$$

$$= n\log\frac{\theta}{\sigma} - \frac{1}{2}\sum_{i=1}^{n}\frac{\theta^2(y_i-\sqrt{\gamma})^2-\sigma^2 y_i^2}{\sigma^2\theta^2}. \tag{16}$$

According to the Neyman-Pearson lemma, the trade-off $\alpha_\beta\big(\varphi^n_{\sqrt{\gamma},\sigma}, \varphi^n_{0,\theta}\big)$ admits the parametric form

$$\alpha(t') = \Pr\big[j(\boldsymbol{Y}_0)\leq t'\big], \tag{17}$$

$$\beta(t') = \Pr\big[j(\boldsymbol{Y}_1)> t'\big], \tag{18}$$

in terms of $t'\in\mathbb{R}$ and where $\boldsymbol{Y}_0\sim\varphi^n_{\sqrt{\gamma},\sigma}$, $\boldsymbol{Y}_1\sim\varphi^n_{0,\theta}$.

We apply [4, Lem. 1] to the tail probabilities (17)-(18) and consider the change of variables $t'\leftrightarrow t$, which are related as $t^2=2\sigma^2\theta^2\frac{1}{\delta}\left(n\log\frac{\theta}{\sigma}+\frac{n}{2}\frac{\gamma}{\delta}-t'\right)$. Then, to obtain the desired result, we proceed as in the proof of Proposition 1 and use that $e^{t'}=\frac{\theta^n}{\sigma^n}\exp\big\{\frac{1}{2}\big(\frac{n\gamma}{\delta}-\frac{\delta t^2}{\sigma^2\theta^2}\big)\big\}$. ∎

IV. Lower Bounds for Average-Power Constraint

The Legendre-Fenchel (LF) transform of a function g is

$$g^*(b) = \max_{a\in\mathcal{A}}\big\{\langle a,b\rangle - g(a)\big\}, \tag{19}$$

where $\mathcal{A}$ is the domain of the function g and $\langle a,b\rangle$ denotes the interior product between a and b.

The function g^* is usually referred to as Fenchel's conjugate (or convex conjugate) of g. If g is a convex function with closed domain, applying the LF transform twice recovers the original function, *i.e.*, $g^{**} = g$. If g is not convex, applying the LF transform twice returns the lower convex envelope of g, which is the largest lower semi-continuous convex function function majorized by g. For our problem, for $f(\beta,\gamma)$ in (11), we define

$$\underline{f}(\beta,\gamma) \triangleq f^{**}(\beta,\gamma), \tag{20}$$

and note that $\underline{f}(\beta,\gamma) \le f(\beta,\gamma)$ for any $\beta\in[0,1]$ and $\gamma\ge 0$.

The lower convex envelope (20) is a lower bound to the error probability in the average power constraint setting.

Theorem 3 (Converse, average power constraint): Let $\mathcal{C}\in\mathcal{L}_{\mathrm{a}}(n,M,\Upsilon)$ be a length-n code of cardinality M satisfying the average power constraint Υ. Then, for any $\theta\ge\sigma$, $n\ge 1$,

$$P_{\mathrm{e}}(\mathcal{C}) \ge \underline{f}\big(\tfrac{1}{M},\Upsilon\big), \tag{21}$$

where $\underline{f}(\beta,\gamma)$ is the lower convex envelope (20) of $f(\beta,\gamma)$ defined in (11).

Proof: We start by considering the general meta-converse bound in (7) with $\mathcal{P}=\mathcal{P}_{\mathrm{a}}(\Upsilon)$ corresponding to the set of distributions satisfying an average power constraint, i.e.,

$$\mathcal{P}_{\mathrm{a}}(\Upsilon) \triangleq \Big\{\boldsymbol{X}\sim P_{\boldsymbol{X}} \;\Big|\; \mathrm{E}\big[\|\boldsymbol{X}\|^2\big]\le n\Upsilon\Big\}. \tag{22}$$

To solve the minimization over P in (7) we shall use the following decomposition. For any $\gamma\ge 0$, we define the set $\mathcal{S}_\gamma \triangleq \big\{\boldsymbol{x}\,|\,\|\boldsymbol{x}\|^2 = n\gamma\big\}$. Then, any input distribution $P_{\boldsymbol{X}}$ induces a distribution over the parameter γ, $P_\gamma \triangleq \Pr\{\boldsymbol{X}\in\mathcal{S}_\gamma\}$, and a conditional distribution

$$\mathrm{d}P_{\boldsymbol{X}|\gamma}(\boldsymbol{x}) = \begin{cases} \frac{\mathrm{d}P_{\boldsymbol{X}}(\boldsymbol{x})}{\mathrm{d}P_\gamma}, & \boldsymbol{x}\in\mathcal{S}_\gamma,\\ 0, & \text{otherwise.}\end{cases} \tag{23}$$

It follows that $P_{\boldsymbol{X}}(\boldsymbol{x}) = \int P_{\boldsymbol{X}|\gamma}(\boldsymbol{x})\,\mathrm{d}P_\gamma$. Here, $\mathrm{d}P_\gamma\ge 0$ and $\int \mathrm{d}P_\gamma = 1$. Furthermore, the conditional distributions $P_{\boldsymbol{X}|\gamma}$ have disjoint support. Then, we apply [5, Lem. 25] to write

$$\begin{aligned}
&\inf_{P\in\mathcal{P}_{\mathrm{a}}(\Upsilon)}\Big\{\alpha_{\frac{1}{M}}\big(PW, P\times Q\big)\Big\}\\
&\quad= \inf_{\substack{\{P_\gamma,\beta_\gamma\}:\\ \int\gamma\,\mathrm{d}P_\gamma=\Upsilon\\ \int\beta_\gamma\,\mathrm{d}P_\gamma=\frac{1}{M}}}\Big\{\int \alpha_{\beta_\gamma}\big(P_\gamma W, P_\gamma\times Q\big)\,\mathrm{d}P_\gamma\Big\} &(24)\\
&\quad= \inf_{\substack{\{P_\gamma,\beta_\gamma\}:\\ \int\gamma\,\mathrm{d}P_\gamma=\Upsilon\\ \int\beta_\gamma\,\mathrm{d}P_\gamma=\frac{1}{M}}}\Big\{\int \alpha_{\beta_\gamma}\big(\varphi^n_{\sqrt{\gamma},\sigma},\varphi^n_{0,\theta}\big)\,\mathrm{d}P_\gamma\Big\}, &(25)
\end{aligned}$$

where the last step follows from the spherical symmetry of each of the tests in (24), using that $\boldsymbol{x}=(\sqrt{\gamma},\ldots,\sqrt{\gamma})\in\mathcal{S}_\gamma$.

Using that $\underline{f}(\beta,\gamma)\le f(\beta,\gamma) = \alpha_\beta\big(\varphi^n_{\sqrt{\gamma},\sigma},\varphi^n_{0,\theta}\big)$, we lower-bound the right-hand side of (25) as

$$\begin{aligned}
&\inf_{\substack{\{P_\gamma,\beta_\gamma\}:\\ \int\gamma\,\mathrm{d}P_\gamma=\Upsilon\\ \int\beta_\gamma\,\mathrm{d}P_\gamma=\frac{1}{M}}}\Big\{\int f\big(\beta_\gamma,\gamma\big)\,\mathrm{d}P_\gamma\Big\}\\
&\qquad\ge \inf_{\substack{\{P_\gamma,\beta_\gamma\}:\\ \int\gamma\,\mathrm{d}P_\gamma=\Upsilon\\ \int\beta_\gamma\,\mathrm{d}P_\gamma=\frac{1}{M}}}\Big\{\int \underline{f}\big(\beta_\gamma,\gamma\big)\,\mathrm{d}P_\gamma\Big\} &(26)\\
&\qquad\ge \inf_{\substack{\{P_\gamma,\beta_\gamma\}:\\ \int\gamma\,\mathrm{d}P_\gamma=\Upsilon\\ \int\beta_\gamma\,\mathrm{d}P_\gamma=\frac{1}{M}}}\Big\{\underline{f}\big(\tfrac{1}{M},\Upsilon\big)\Big\} &(27)\\
&\qquad= \underline{f}\big(\tfrac{1}{M},\Upsilon\big), &(28)
\end{aligned}$$

where (27) follows by applying Jensen's inequality since $\underline{f}(\beta,\gamma)$ is jointly convex in both parameters and by using the constraints; and (28) holds since the objective of the optimization in (27) does not depend on $\{P_\gamma,\beta_\gamma\}$.

The lower bound (21) then follows from combining (7), (24)-(25) and the inequalities (26)-(28). ∎

The function $\underline{f}(\beta,\gamma)$ can be evaluated numerically by considering a 2-dimensional grid of the parameters (β,γ), using (14) to compute $f(\beta,\gamma)$ over this grid, and obtaining the corresponding convex envelope. Nevertheless, sometimes $\underline{f}\big(\frac{1}{M},\Upsilon\big) = f\big(\frac{1}{M},\Upsilon\big) = \alpha_{\frac{1}{M}}\big(\varphi^n_{\sqrt{\Upsilon},\sigma},\varphi^n_{0,\theta}\big)$ and these steps can be avoided, as the next result shows.

Corollary 2: Let $\sigma,\theta>0$ and $n\ge 1$, be fixed parameters, and define $\delta\triangleq\theta^2-\sigma^2$. For $t\ge 0$, we define

$$\xi_1(t) \triangleq Q_{\frac{n}{2}}\left(\sqrt{n\Upsilon}\frac{\sigma}{\delta},\frac{t}{\sigma}\right) - Q_{\frac{n}{2}}\left(0,\sqrt{\big(\tfrac{t^2}{\sigma^2}-n\Upsilon\tfrac{\theta^2}{\delta^2}\big)_+}\right), \tag{29}$$

$$\xi_2(t) \triangleq \frac{\theta^n}{\sigma^n}e^{-\frac{1}{2}\left(\frac{t^2}{\sigma^2\theta^2}-\frac{n\Upsilon}{\delta}\right)}\left(Q_{\frac{n}{2}}\left(0,\sqrt{\big(\tfrac{t^2}{\theta^2}-n\Upsilon\tfrac{\sigma^2}{\delta^2}\big)_+}\right) - Q_{\frac{n}{2}}\left(\sqrt{n\Upsilon}\frac{\theta}{\delta},\frac{t}{\theta}\right)\right), \tag{30}$$

$$\xi_3(t) = \frac{n\Upsilon}{2\delta}\left(\frac{t\delta}{\sigma^2\sqrt{n\Upsilon}}\right)^{\frac{n}{2}}e^{-\frac{1}{2}\left(\frac{t^2}{\sigma^2}+n\Upsilon\frac{\sigma^2}{\delta^2}\right)}I_{\frac{n}{2}}\left(\sqrt{n\Upsilon}\frac{t}{\delta}\right), \tag{31}$$

where $(a)_+=\max(0,a)$, $Q_m(a,b)$ is the Marcum Q-function and $I_m(\cdot)$ denotes the m-th order modified Bessel function of the first kind. Let t_0 be the solution to the implicit equation

$$\xi_1(t_0)+\xi_2(t_0)+\xi_3(t_0)=0, \tag{32}$$

and let

$$\bar{M} \triangleq \big(1-Q_{\frac{n}{2}}\big(\sqrt{n\Upsilon}\theta/\delta,\, t_0/\theta\big)\big)^{-1}. \tag{33}$$

Then, for any code $\mathcal{C}\in\mathcal{L}_{\mathrm{a}}(n,M,\Upsilon)$ with cardinality $M\le\bar{M}$,

$$P_{\mathrm{e}}(\mathcal{C}) \ge \alpha_{\frac{1}{M}}\big(\varphi^n_{\sqrt{\Upsilon},\sigma},\varphi^n_{0,\theta}\big). \tag{34}$$

Proof: See the Appendix. ∎

Corollary 2 implies that the bound from Theorems 1 and 2 holds in the average power constraint setting if the cardinality of the codebook is sufficiently small. Indeed, it follows that this condition is satisfied for typical communication systems. For transmission rates very close to capacity or above capacity, the bound (21) is needed instead (see the example in Fig. 2).

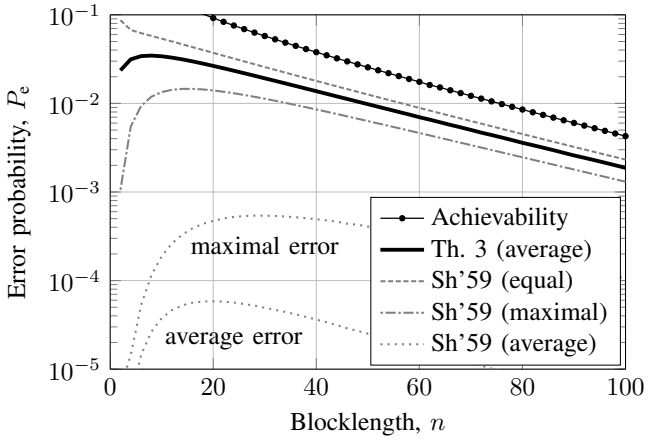


Fig. 1: Upper and lower bounds to the channel coding error probability over an AWGN channel with SNR $= 10$ dB and rate $R = 1.5$ bits/channel use.

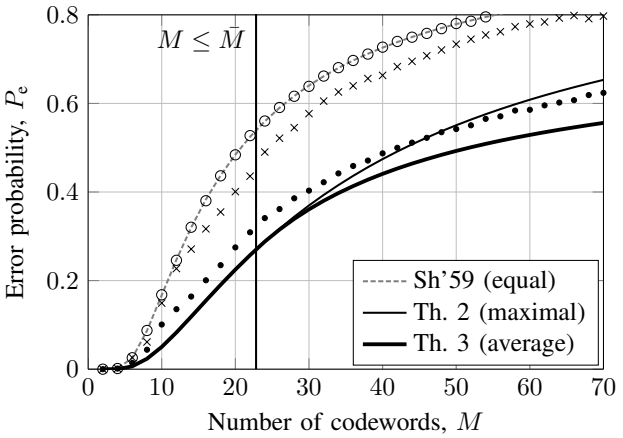


Fig. 2: Lower bounds to the channel coding error probability over an AWGN channel with $n = 2$ and SNR $= 10$ dB. Markers show the simulated error probability of a sequence of codes satisfying an equal ($\circ$), maximal ($\times$) and average ($\bullet$) power constraints. Vertical line corresponds to the boundary $M \leq \bar{M} \approx 22.8$ from Corollary 2.

V. Numerical Examples

A. Comparison with previous results

We consider the transmission of $M = 2^{nR}$ codewords over n uses of an AWGN channel with $R = 1.5$ bits/channel use and SNR $= 10 \log_{10} \frac{\Upsilon}{\sigma^2} = 10$ dB. The channel capacity is $C = \frac{1}{2} \log_2\big(1 + \frac{\Upsilon}{\sigma^2}\big) \approx 1.8$ bits/channel use.

Figure 1 compares the lower bound from Theorem 3 with previous results in the literature. In particular, we consider Shannon'59 achievability and converse bounds for equal power constraint [1, Eq. (20)], Shannon'59 converse bound for maximal power constraint [1, Eqs. (20) and (83)], and the lower bound for average power constraint that follows from combining [1, Eq. (20)] and [6, Lem. 65]. While the bounds in Figure 1 hold under the average probability of error formalism, for reference we also include the curve *Sh'59 (average)* for maximal error probability, which is tighter than that for average error probability (see [6, Lem. 65] for details).

As the transmission rate R is close to capacity C, the optimizing θ^2 in Theorem 3 is close to the variance of the capacity achieving output distribution. Then, for simplicity, we fix $\theta^2 = \Upsilon + \sigma^2$. For the system parameters considered, the condition $M \leq \bar{M}$ from Corollary 2 is satisfied for all n and Theorem 3 can be evaluated using (34). It thus follows that the bounds from Theorems 1, 2 and 3 coincide.

The results in Figure 1 show that that Shannon'59 lower bound is the tightest bound in the equal power constraint setting. However, under both maximal and average power constraints, Theorem 3 yields a tighter lower bound and presents a small constant gap to the achievability bound from [1, Eq. (20)].[1] Indeed, for an average power constraint and under the average probability of error formalism the advantage of Theorem 3 over previous results is significant in the finite blocklength regime, as shown in Figure 1.

[1]The rate considered here is above the critical rate of the channel, and therefore the error exponents of the achievability and converse bounds in Figure 1 coincide. This is not longer true for rates below the critical rate.

B. Constellation design under power constraints

We consider the problem of transmitting M codewords with $n = 2$ uses of an AWGN channel with SNR $= 10$ dB. This problem is analogous to determining the best 2-dimensional constellation for an uncoded communication system.

Figure 2 depicts Shannon'59 lower bound [1, Eq. (20)], and the bounds from Theorems 2 and 3, both with $\theta^2 = \Upsilon + \sigma^2$. The vertical line shows the boundary of the region $M \leq \bar{M}$ from Corollary 2, where the bounds from Theorems 2 and 3 coincide. With markers, we show the simulated ML decoding error probability of a sequence of M-PSK (phase-shift keying) constellations satisfying an equal power constraint ($\circ$), of a sequence of M-APSK (amplitude-phase-shift keying) constellations satisfying a maximal ($\times$) and average ($\bullet$) power constraints (both optimized using an stochastic algorithm).

As 2-dimensional cones coincide with the ML decoding regions of an M-PSK constellation, Shannon'59 curve is on top of the corresponding simulated probability ($\circ$). However, Shannon'59 lower bound does not apply to M-APSK constellations satisfying maximal ($\times$) and average ($\bullet$) power constraints. We can see that while Theorem 3 applies in both of these settings, this is not the case for Theorem 2, that in general only applies under maximal power constraint. As stated in Corollary 2, the bounds from Theorems 2 and 3 coincide for $M \leq \bar{M} \approx 22.8$.

An analysis of the average power constrained codes ($\bullet$) that violate Theorem 2 shows that they present several constellation points concentrated at the origin $(0, 0)$. As these symbols coincide, it is not possible to distinguish between them and they will often yield a decoding error. However, since the symbol $(0, 0)$ does not require any energy for its transmission, the average power for the remaining symbols is increased and this code yields an overall smaller probability of error.

ACKNOWLEDGMENT

The author gratefully acknowledges the insightful comments by the reviewers and fruitful discussions with Tobias Koch and David Morales-Jimenez.

APPENDIX

We characterize the region where $f(\beta,\gamma)$ and its convex envelope $\underline{f}(\beta,\gamma)$ coincide. We shall use the following result.

Proposition 2: Suppose g is differentiable with gradient ∇g. Let $\mathcal{A}$ denote the domain of g, and let $a \in \mathcal{A}$. If the inequality

$$g(\bar{a}) \geq g(a) + \nabla g(a)^T(\bar{a} - a), \tag{35}$$

is satisfied for all $\bar{a} \in \mathcal{A}$, then, it holds that $g(a) = g^{**}(a)$.

Proof: As g^{**} is the lower convex envelope of g, then $g(a) \geq g^{**}(a)$ trivially. It remains to show that (35) implies $g(a) \leq g^{**}(a)$. Fenchel's inequality [7, Sec. 3.3.2] yields

$$g^{**}(a) \geq \langle a, b\rangle - g^*(b), \tag{36}$$

for any b in the domain of g^*.

Setting $b = \nabla g(a)$ and using (19) in (36), we obtain

$$\begin{aligned} g^{**}(a) &\geq \nabla g(a)^T a - \max_{\bar{a}\in\mathcal{A}}\big\{\nabla g(a)^T\bar{a} - g(\bar{a})\big\} && (37)\\ &= \min_{\bar{a}\in\mathcal{A}}\big\{\nabla g(a)^T(a-\bar{a}) + g(\bar{a})\big\} && (38)\\ &\geq \min_{\bar{a}\in\mathcal{A}}\big\{g(a)\big\}, && (39)\end{aligned}$$

where in the last step we used (35) to lower bound $g(\bar{a})$. Since the objective of (39) does not depend on $\bar{a}$, we conclude from (37)-(39) that $g(a) \leq g^{**}(a)$ and the result follows. ∎

We apply Proposition 2 to the function $f(\beta,\gamma)$. We recall that $f(\beta,\gamma)$ is differentiable for $\beta \in [0,1]$ and $\gamma \geq 0$ with derivatives given in [8, App. A]. We define the gradients

$$\nabla_\beta f(b,g) \triangleq \frac{\partial f(\beta,\gamma)}{\partial\beta}\Big|_{\beta=b,\gamma=g}, \tag{40}$$

$$\nabla_\gamma f(b,g) \triangleq \frac{\partial f(\beta,\gamma)}{\partial\gamma}\Big|_{\beta=b,\gamma=g}. \tag{41}$$

According to Proposition 2, the function $f(\beta_0,\gamma_0)$ and its convex envelope $\underline{f}(\beta_0,\gamma_0)$ coincide if

$$\begin{aligned} f(\beta,\gamma) \;\geq\; & f(\beta_0,\gamma_0) + (\beta-\beta_0)\nabla_\beta f(\beta_0,\gamma_0)\\ & + (\gamma-\gamma_0)\nabla_\gamma f(\beta_0,\gamma_0). \end{aligned} \tag{42}$$

is satisfied for all $\beta \in [0,1]$ and $\gamma \geq 0$. This condition implies that the first-order Taylor approximation of f at (β_0,γ_0) is a global under-estimator of the function $f(\beta,\gamma)$.

The derivatives of $f(\beta,\gamma)$, given in [8, App. A], show that the function is decreasing in both parameters, convex with respect to β for all $\beta \in [0,1]$, and jointly convex with respect to (β,γ) except for the neighborhood near the axis $\gamma = 0$. Using these properties, it can be shown that the the condition (42) only needs to be verified along the axis $\gamma = 0$.

Then, we conclude that $f(\beta_0,\gamma_0) = \underline{f}(\beta_0,\gamma_0)$ if (42) holds for every $\beta \in [0,1]$ and $\gamma = 0$, i.e., if

$$\begin{aligned} f(\beta_0,\gamma_0) - f(\beta,0) \geq\; & (\beta_0-\beta)\nabla_\beta f(\beta_0,\gamma_0)\\ & + \gamma_0\nabla_\gamma f(\beta_0,\gamma_0). \end{aligned} \tag{43}$$

Let $\theta \geq \sigma > 0$, $n \geq 1$. Let t_0 be the value such that $\beta(\gamma_0,t_0) = \beta_0$ and let $\bar{t}$ satisfy $\beta(0,\bar{t}) = \beta$, for $\beta(\gamma,t)$ defined in (13). Using (12) and the derivatives in [8, App. A], yields

$$f(\beta_0,\gamma_0) - f(\beta,0) = Q_{\frac{n}{2}}\left(\sqrt{n\gamma_0}\frac{\sigma}{\delta}, \frac{t_0}{\sigma}\right) - Q_{\frac{n}{2}}\left(0, \frac{\bar{t}}{\sigma}\right), \tag{44}$$

$$\nabla_\beta f(\beta_0,\gamma_0) = -\frac{\theta^n}{\sigma^n} e^{\frac{1}{2}\left(\frac{n\gamma_0}{\delta} - t_0^2\left(\frac{1}{\sigma^2}-\frac{1}{\theta^2}\right)\right)}, \tag{45}$$

$$\begin{aligned} \nabla_\gamma f(\beta_0,\gamma_0) = & -\frac{n}{2\delta}\left(\frac{t_0\delta}{\sigma^2\sqrt{n\gamma_0}}\right)^{\frac{n}{2}} I_{\frac{n}{2}}\left(\frac{t_0\sqrt{n\gamma_0}}{\delta}\right)\\ & \times e^{-\frac{1}{2}\left(\frac{n\gamma_0\sigma^2}{\delta^2} + \frac{t_0^2}{\sigma^2}\right)}. \end{aligned} \tag{46}$$

As $\beta(\gamma_0,t_0) = \beta_0$ and $\beta(0,\bar{t}) = \beta$, using (13), it follows that

$$\beta_0 - \beta = Q_{\frac{n}{2}}\left(0, \frac{\bar{t}}{\theta}\right) - Q_{\frac{n}{2}}\left(\sqrt{n\gamma_0}\frac{\theta}{\delta}, \frac{t_0}{\theta}\right). \tag{47}$$

Substituting (44) and (47) in (43), reorganizing terms, yields

$$\begin{aligned} & Q_{\frac{n}{2}}\left(\sqrt{n\gamma_0}\frac{\sigma}{\delta}, \frac{t_0}{\sigma}\right) + \nabla_\beta f(\beta_0,\gamma_0) Q_{\frac{n}{2}}\left(\sqrt{n\gamma_0}\frac{\theta}{\delta}, \frac{t_0}{\theta}\right)\\ & - \gamma_0\nabla_\gamma f(\beta_0,\gamma_0) \;\geq\; Q_{\frac{n}{2}}\left(0, \frac{\bar{t}}{\sigma}\right) + \nabla_\beta f(\beta_0,\gamma_0) Q_{\frac{n}{2}}\left(0, \frac{\bar{t}}{\theta}\right). \end{aligned} \tag{48}$$

The interval $\beta \in [0,1]$ corresponds to $\bar{t} \geq 0$. We maximize the right-hand side of (48) over $\bar{t} \geq 0$ and we only verify the condition (48) for this maximum value. To this end, we find the derivative of the right-hand side of (48) with respect to $\bar{t}$, we identify the resulting expression with zero, and we use (45). We conclude that the right-hand side of (48) is maximized for

$$\bar{t}_\star = \sqrt{\left(t_0^2 - n\gamma\sigma^2\theta^2/\delta^2\right)_+} \tag{49}$$

where the threshold $(a)_+ = \max(0,a)$ follows from the constraint $\bar{t} \geq 0$. By evaluating the second derivative of (48), it can be verified that $\bar{t}_\star$ in (49) is indeed a maximum.

Using (45), (46) and (49) in (48) we obtain the desired characterization for the region of interest. For the statement of the result in Corollary 2, we select the smallest t_0 that fulfills (48) (which satisfies the condition with equality) and invert the transformation $\beta(\gamma_0,t_0) = \beta_0$ for $\gamma_0 = \Upsilon$ and $\beta_0 = \frac{1}{M}$.

REFERENCES

[1] C. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell System Technical Journal*, vol. 38, p. 611656, 1959.

[2] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[3] G. Vazquez-Vilar, "On the error probability of optimal codes in Gaussian channels under maximal power constraint," in *2019 IEEE Int. Symp. on Inf. Theory*, Paris, France, July 2019.

[4] G. Vazquez-Vilar, A. Guillén i Fàbregas, T. Koch, and A. Lancho, "Saddlepoint approximation of the error probability of binary hypothesis testing," in *2018 IEEE Int. Symp. on Inf. Theory*, June 2018, pp. 2306–2310.

[5] Y. Polyanskiy, "Saddle point in the minimax converse for channel coding," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2576–2595, May 2013.

[6] ——, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, Sep. 2010.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, US: Cambridge University Press, 2004.

[8] G. Vazquez-Vilar, "Error probability bounds for Gaussian channels under maximal and average power constraints," *preprint arXiv:1907.03163*, 2019.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

On the Broadcast Approach over Parallel MIMO Two-state Fading Channel

Kfir M. Cohen, Avi Steiner, Shlomo Shamai (Shitz)

Technion - Israel Institute of Technology. {kfir.cohen.m, steiner.avi}@gmail.com, sshlomo@ee.technion.ac.il

***Abstract*—The single user parallel multiple input multiple output (MIMO) slow (block) flat fading channel, subject to a two-state fading per channel with additive white Gaussian noise (AWGN) is examined. The fading in each of the parallel channels is interpreted as state, which takes on two values with prescribed probabilities. We focus here on the variable to fixed channel rate (the broadcast approach) where a novel view of extension of El-Gamal's capacity of degraded broadcast product channels is examined. The optimized average rate is analytically derived over the parameters of the proposed scheme, and comparison to the simple scheme that employs the broadcast approach per each of the parallel channels separately. The achievable improvement in rates under the latency demand (transmission in a single fading block) is reflected.**

I. Introduction

Recent growth in bandwidth requirements of the 5G wireless communications networks, under stringent low latency requirements lead to vast contributions of innovations. This work focuses on the slow (block) fading parallel MIMO channel [1], where channel state is known at the receiver only. Under this channel model the transmitter may adopt a broadcast approach [2], which can optimize the expected transmission rate under no transmission channel state information (CSI), which is essentially characterized by the *variable-to-fixed coding* [3].

The *broadcast approach* [2] for slow flat-fading channels [4] uses the degradedness nature of the fading channel and applies multi-layer coding, to deliver variable-to-fixed coding over block fading channels. The amount of successfully decoded layers depends on the channel realization. For deeply fading channels few layers are decoded, while for high fading gains, more layers can be decoded. Rate and power allocation per layer are optimized to maximize the expected rate. The broadcast approach can be compared to the ergodic bound [5], achievable given transmit CSI, and other contributions such as [6]–[14].

El-Gamal [15] composed two degraded broadcast channels [16], [17] into a three-user setup: an encoder with two outputs, each driving a dual-output broadcast channel; two decoders, each is input by one less-noisy broadcast channel output and one more-noisy output of the other channel (called 'unmatched'). This was coined *degraded broadcast product channel*. For the AWGN case, the capacity region (private and common rates) was derived.

In this paper, the MIMO setup for the broadcast approach is revisited, with new tools that differ from those in [2], [18]. This is by analyzing the finite state parallel MIMO channel, where El-Gamal's capacity region [15] is used to address the multi-layering optimization problem for maximizing the expected rate of a two-state fading [19]–[21] parallel MIMO channel.

II. Channel Model

Consider a single user parallel MIMO channel setting, where a message w is to be block-encoded and sent through a diagonal matrix two-input two-output flat fading channel depicted in Fig. 1. The channel is given by

$$\begin{aligned}\mathbf{Y}_1 &= H_1\mathbf{X}_1 + \mathbf{N}_1,\\ \mathbf{Y}_2 &= H_2\mathbf{X}_2 + \mathbf{N}_2,\end{aligned} \tag{1}$$

where $\mathbf{Y}_i \in \mathbb{C}^n$ is the received n-length symbols vector on channel $i \in 1,2$, $\mathbf{X}_i \in \mathbb{C}^n$ is the transmitted vector over channel i which satisfies the power constraint $\mathbb{E}[|X_i|^2] \leq P$, $i \in 1,2$. The additive noise vector is denoted $\mathbf{N}_i \in \mathbb{C}^n$ and its elements are complex normal i.i.d with zero mean and unit variance $\mathcal{CN}(0,1)$. The i-th sub-channel fading coefficient is denoted $H_i \in \mathbb{R}_+$, is drawn by some probability function $P_H(\cdot)$ and its value remains fixed during a block transmission, changes along blocks independently, and H_1 and H_2 are statistically independent. These channel states are known only to the receiver side and are not fed back to the transmitter. With no loss of generality, the channel fading H_i is assumed to be real and positive.

For a given realization set of channel states $\{H_1, H_2\}$ known to both the transmitter and receiver, the per-block Shannon capacity is well known [1]. Since H_1 and H_2 are unknown to the transmitter, setting the rates to withstand the worst (lowest) possible H_i may occur a great deal of rate loss. Variable-to-fixed coding allows to deliver higher throughput, at the expense that only parts of the message are decodable, according the channel conditions. Clearly, the expected achievable rate can be higher than the worst-case classical capacity. The recovered message $\hat{w}$ has different cardinality upon the realization set.

In this work, the channel model is limited to a two-state symmetric case. Each channel $i = 1,2$ can have independent fading gain realizations $S_i \in \{A, B\}$, state A denotes a fading coefficient $H_i = H_A$ with probability P_A; whereas state B refers to the sub-channel $H_i = H_B$, and $|H_A| < |H_B|$, and is with probability $P_B = 1 - P_A$. This is reflected by the condition $P_H(h) = P_A\delta(h - H_A) + P_B\delta(h - H_B)$ where $\delta(\cdot)$ is the kronecker delta.For brevity, denote the fading gains by $\nu = |H|^2$, $\nu_a = |H_A|^2$ and $\nu_b = |H_B|^2$ and by definition $\nu_b > \nu_a$. The common power constraint is given by $\mathbb{E}[|X_i|^2] \leq P$, $i = 1,2$. The ergodic capacity of the two state fading parallel MIMO channel is specified by $C_{\text{erg}} = 2(P_A \log(1 + P\nu_a) + P_B \log(1 + P\nu_b))$.

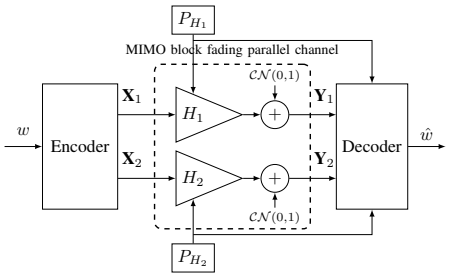


Fig. 1: The parallel MIMO block fading channel with channel state information at receiver. All codewords are of length n.

III. Preliminary: Capacity of Degraded Gaussian Broadcast Product Channels

Consider the model introduced in [15]: two receiver discrete memoryless degraded product broadcast channels. The Gaussian case was addressed as a special case. A single transmitter codes two n-length codewords consisting of a common message $w_0 \in \{1, ..., 2^{nR_0}\}$ to be decoded by both users, and two private messages $w_{BA} \in \{1, ..., 2^{nR_{BA}}\}$ and $w_{AB} \in \{1, ..., 2^{nR_{AB}}\}$, one for each of the two decoding users. A single function codes these 3 messages into two codewords; each undergoes parallel degraded broadcast subchannels

$$\begin{cases} Y_1 = X_1 + N_{11} \\ Z_1 = Y_1 + N_{12} \end{cases} \qquad \begin{cases} Z_2 = X_2 + N_{21} \\ Y_2 = Z_2 + N_{22}, \end{cases}$$

and $N_{11}, N_{21} \sim \mathcal{CN}(0, \nu_b^{-1})$, $N_{21}, N_{22} \sim \mathcal{CN}(0, \nu_a^{-1} - \nu_b^{-1})$.

As depicted in the bold and red parts of Fig. 2, two users (namely AB and BA) receive both common and private messages from the transmitter independently decode the messages. This is an unmatched setting, as Y_1 is less noisy than Z_1, alas Z_2 is less noisy than Y_2. Hence, each of the users has one less-noisy channel output alongside another which is the noisier output of the other sub-channel.

Following Theorem 2 of [15] which shows this case, and exploiting symmetry for equal power allocation to both sub-channels, optimal allocation is expected to be achieved by equal common rate allocation to every user (state). Denoting $\bar{\alpha} = 1 - \alpha$, the capacity region (R_0, R_{BA}, R_{AB}) is

$$\begin{aligned} R_0 &\le \log\left(1 + \tfrac{\nu_a \alpha P}{1+\nu_a \bar{\alpha} P}\right) + \log\left(1 + \tfrac{\nu_b \alpha P}{1+\nu_b \bar{\alpha} P}\right) \\ R_0 + R_{BA} = R_0 + R_{AB} &\le \log\left(1 + \tfrac{\nu_a \alpha P}{1+\nu_a \bar{\alpha} P}\right) + \log(1 + \nu_b P) \\ R_0 + R_{BA} + R_{AB} &\le \log\left(1 + \nu_b P\right) + \log\left(1 + \tfrac{\nu_a \alpha P}{1+\nu_a \bar{\alpha} P}\right) \\ &\quad + \log\left(1 + \nu_b \bar{\alpha} P\right). \end{aligned} \tag{2}$$

IV. Main Contribution

A. Extended Degraded Gaussian Broadcast Product Channels

The classical product channel is extended by introducing two dual-input receivers in addition to the original two. The first has the two more noisy channel outputs (Z_1, Y_2), whereas the second gets the two less noisy outputs (Z_2, Y_1). To support this, two messages w_{AA} and w_{BB} are added. The total two n-length codewords are the superposition of three codewords by independent encoders as follows $(\mathbf{X}_1, \mathbf{X}_2) = f_{AA}(w_{AA}) + f_{cr}(w_0, w_{BA}, w_{AB}) + f_{BB}(w_{BB})$, where subscript cr stands for "crossed" states ((A, B) or (B, A)). See Fig. 2 for an illustration.

Stream AA is decoded first, regardless of whether the others can be decoded (this is done by treating all the other streams as interference). Then, both streams AB and BA including their common stream subscripted 0 can be decoded after removing the AA impact from their decoder inputs (treating the BB stream as interference). Finally, removing all above decoded streams allows decoding stream BB. From (2), we have

$$\begin{aligned} &R_{AA} \le 2\log\left(1 + \tfrac{\alpha_{AA} P}{\nu_a^{-1} + \bar{\alpha}_{AA} P}\right); \\ &R_{AA} + R_0 \le 2\log\left(1 + \tfrac{\alpha_{AA} P}{\nu_a^{-1} + \bar{\alpha}_{AA} P}\right) \\ &\quad + \log\left(1 + \tfrac{\alpha\alpha_{cr} P}{\nu_b^{-1} + (\bar{\alpha}\alpha_{cr} + \alpha_{BB})P}\right) + \log\left(1 + \tfrac{\alpha\alpha_{cr} P}{\nu_a^{-1} + (\bar{\alpha}\alpha_{cr} + \alpha_{BB})P}\right); \\ &R_{AA} + R_0 + R_{BA} = R_{AA} + R_0 + R_{AB} \\ &\le 2\log\left(1 + \tfrac{\alpha_{AA} P}{\nu_a^{-1} + \bar{\alpha}_{AA} P}\right) + \log\left(1 + \tfrac{\alpha\alpha_{cr} P}{\nu_a^{-1} + (\bar{\alpha}\alpha_{cr} + \alpha_{BB})P}\right) \\ &\quad + \log\left(1 + \tfrac{\alpha_{cr} P}{\nu_b^{-1} + \alpha_{BB} P}\right); \\ &R_{AA} + R_0 + R_{BA} + R_{AB} \\ &\le 2\log\left(1 + \tfrac{\alpha_{AA} P}{\nu_a^{-1} + \bar{\alpha}_{AA} P}\right) + \log\left(1 + \tfrac{\alpha_{cr} P}{\nu_b^{-1} + \alpha_{BB} P}\right) \\ &\quad + \log\left(1 + \tfrac{\alpha\alpha_{cr} P}{\nu_a^{-1} + (\bar{\alpha}\alpha_{cr} + \alpha_{BB})P}\right) + \log\left(1 + \tfrac{\bar{\alpha}\alpha_{cr} P}{\nu_b^{-1} + \alpha_{BB} P}\right); \\ &R_{AA} + R_0 + R_{BA} + R_{AB} + R_{BB} \\ &\le 2\log\left(1 + \tfrac{\alpha_{AA} P}{\nu_a^{-1} + \bar{\alpha}_{AA} P}\right) + \log\left(1 + \tfrac{\alpha_{cr} P}{\nu_b^{-1} + \alpha_{BB} P}\right) \\ &\quad + \log\left(1 + \tfrac{\alpha\alpha_{cr} P}{\nu_a^{-1} + (\bar{\alpha}\alpha_{cr} + \alpha_{BB})P}\right) + \log\left(1 + \tfrac{\bar{\alpha}\alpha_{cr} P}{\nu_b^{-1} + \alpha_{BB} P}\right) \\ &\quad + 2\log\left(1 + \tfrac{\alpha_{BB} P}{\nu_b^{-1}}\right); \end{aligned} \tag{3}$$

where $\alpha_{AA}, \alpha_{cr}, \alpha_{BB} \in [0, 1]$ are the relative power allocations for the subscripted letters $\alpha_{AA} + \alpha_{cr} + \alpha_{BB} = 1$, and $\alpha \in [0, 1]$ is the single user private power allocation within the unmatched channel.

B. Suggested Encoding and Decoding Scheme

Wrapping the extended model of Section IV-A with a message splitter at the transmitter and channel state dependent message multiplexer at the receiver enriches the domain. Fig. 3 illustrates the encoding and decoding schemes in full.

During decoding, the 4 possible channel states $\mathbf{S} = (S_1, S_2)$ impose different decoding capabilities. If $\mathbf{S} = (A, A)$, then $g_{AA}(\cdot)$ can reconstruct w_{AA} to achieve a total rate of R_{AA}. For $\mathbf{S} = (B, A)$, $g_{BA}(\cdot)$ is capable of reconstructing three messages (w_{AA}, w_0, w_{BA}) with sum rate of $R_{AA} + R_0 + R_{BA}$. Similarly for $\mathbf{S} = (A, B)$, $g_{AB}(\cdot)$ reconstructs (w_{AA}, w_0, w_{AB}) with sum rate $R_{AA} + R_0 + R_{AB}$. When both channels are permissive $\mathbf{S} = (B, B)$, all 5 messages $(w_{AA}, w_0, w_{BA}, w_{AB}, w_{BB})$ are reconstructed at $g_{BB}(\cdot)$ under the rate $R_{AA} + R_0 + R_{BA} + R_{AB} + R_{BB}$.

C. Average Sum Rate

Stitching up all cases with their probabilities, gives rise to the average rate of the parallel channel of

$$\begin{aligned} R_{\text{avg}} = {} & P_A^2 R_{AA} + P_A P_B (R_{AA} + R_0 + R_{AB}) \\ & + P_B P_A (R_{AA} + R_0 + R_{BA}) \\ & + P_B^2 (R_{AA} + R_0 + R_{BA} + R_{AB} + R_{BB}). \end{aligned} \tag{4}$$

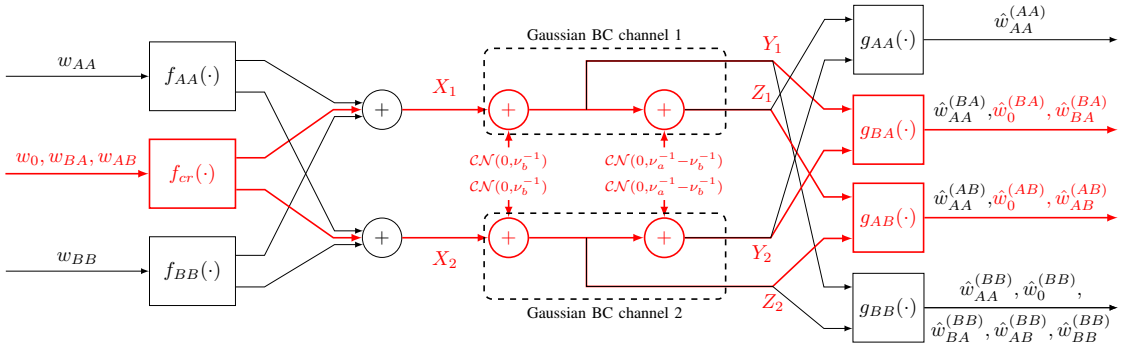


Fig. 2: Encoding-decoding scheme of the 2 receiver Gaussian degraded product broadcast channel with users: AA,AB,BA,BB

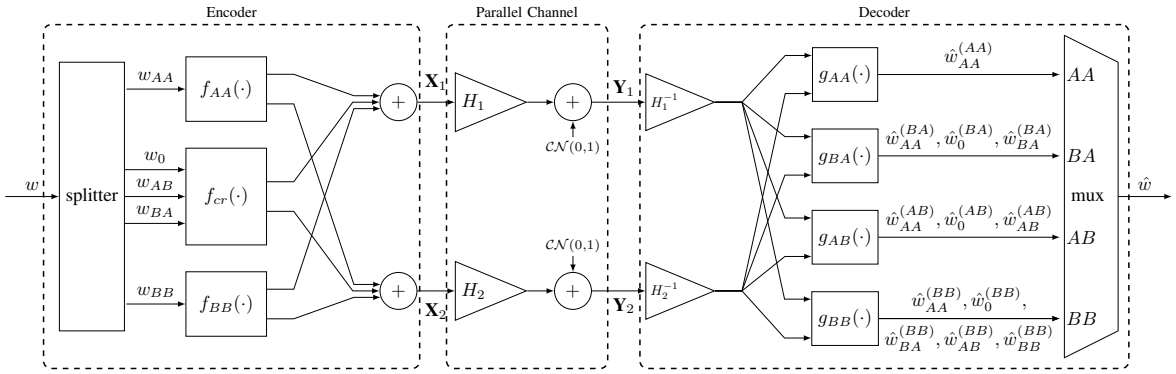


Fig. 3: Encoding and decoding scheme of the two receiver Gaussian degraded product broadcast channel broadcast approach

Using (3), and since both channels have identical statistics lead to $R_{AB} = R_{BA}$, and the achievable average rate is

$$R_{\text{avg}} = 2(P_A+P_B)^2 \log\left(1+\nu_a P\right) + R_0(1-\alpha_{AA}) + R_1(1-\alpha_{AA}-\alpha\alpha_{cr}) + R_2(1-\alpha_{AA}-\alpha_{cr}), \quad (5)$$

where the new notations are

$$R_0(\alpha_0) = [(P_A+P_B)^2 - P_A^2]\log(1+\nu_b\alpha_0 P) - [(P_A+P_B)^2 + P_A^2]\log(1+\nu_a\alpha_0 P), \quad (6)$$

$$R_1(\alpha_1) = P_B^2 \log(1+\nu_b\alpha_1 P) - [(P_A+P_B)^2 - P_A^2]\log(1+\nu_a\alpha_1 P), \quad (7)$$

$$R_2(\alpha_2) = -2P_A P_B \log(1+\nu_b\alpha_2 P). \quad (8)$$

and the arguements $\alpha_0 = 1-\alpha_{AA}$, $\alpha_1 = 1-\alpha_{AA}-\alpha\alpha_{cr}$ and $\alpha_2 = 1-\alpha_{AA}-\alpha_{cr} = \alpha_{BB}$. Note that $R_0(\alpha_0)$ and $R_1(\alpha_1)$ are not obliged to be positive, as they can be negative for some scenarios, and $R_2(\alpha_2)$ is non-positive by definition.

Denoting the domain D' of valid power allocations vector $\boldsymbol{\alpha}' = [\alpha, \alpha_{AA}, \alpha_{cr}, \alpha_{BB}]^T \in [0,1]^4$ and the operator $[x]_+ = \max\{0,x\}$ yield the following.

Proposition 1. *The maximal sum rate of the symmetric two parallel two state channel over all power allocations is*

$$\max_{\boldsymbol{\alpha}'\in D'} R_{\text{avg}}(\boldsymbol{\alpha}') = 2(P_A+P_B)^2\log(1+\nu_a P) + \max_{0\le\alpha_{AA}\le 1}\left\{R_0(1-\alpha_{AA}) + R_1(\alpha_1^{\text{opt}}(\alpha_{AA}))\right\},$$

where

$$\alpha_1^{\text{opt}}(\alpha_{AA}) = \max\{0, \min\{1-\alpha_{AA}, \alpha_1^*\}\}, \quad (9)$$

$$\alpha_1^* = \frac{P_B^2\nu_b - [(P_A+P_B)^2 - P_A^2]\nu_a}{[(P_A+P_B)^2 - P_A^2 - P_B^2]\nu_a\nu_b P}, \quad (10)$$

where the latter solves $\frac{\partial}{\partial\alpha_1} R_1(\alpha_1^*) = 0$.

Proof. Consider the transform $t' : D' \to D$ defined by $[\alpha_0,\alpha_1,\alpha_2]^T = \boldsymbol{\alpha} = t'(\boldsymbol{\alpha}'^T) = t'\left([\alpha,\alpha_{AA},\alpha_{cr},\alpha_{BB}]^T\right) = [1-\alpha_{AA}, 1-\alpha_{AA}-\alpha\alpha_{cr}, \alpha_{BB}]^T$, which is bijective, with inverse transform $t : D \to D'$ defined by $[\alpha,\alpha_{AA},\alpha_{cr},\alpha_{BB}]^T = \boldsymbol{\alpha}' = t'(\boldsymbol{\alpha}'^T) = t'\left([\alpha_0,\alpha_1,\alpha_2]^T\right) = [\frac{\alpha_0-\alpha_1}{\alpha_0-\alpha_2}, 1-\alpha_0, \alpha_0-\alpha_2, \alpha_2]^T$. Bijectiveness leads to

$$\max_{\boldsymbol{\alpha}'\in D'} R_{\text{avg}}(\boldsymbol{\alpha}') = \max_{\boldsymbol{\alpha}\in D}\left\{2(P_A+P_B)^2\log(1+\nu_a P) + \sum_{i=0}^{2} R_i(\alpha_i)\right\}$$
$$= 2(P_A+P_B)^2\log\left(1+\nu_a P\right) + \max_{\substack{\alpha_0,\alpha_1:\\ 0\le\alpha_1\le\alpha_0\le 1}}\{R_0(\alpha_0)+R_1(\alpha_1)\}$$

The maximization of $R_2(\alpha_2)$ yields $\alpha_2^{\text{opt}} = 0$, as $R_2(\alpha_2)$ is a decreasing function. Further simplification gives,

$$\max_{\boldsymbol{\alpha}'\in D'} R_{\text{avg}}(\boldsymbol{\alpha}') = 2(P_A+P_B)^2\log\left(1+\nu_a P\right) + \max_{0\le\alpha_{AA}\le 1}\left\{R_0(1-\alpha_{AA}) + \max_{\substack{\alpha_1:\\ 0\le\alpha_1\le 1-\alpha_{AA}}} R_1(\alpha_1)\right\}.$$

The inner maximization is done over α_1 while α_{AA} is fixed prior to the maximization. By taking the first derivative w.r.t α_1 and some calculus, optimality is achieved for (9). ■

Corollary 2. *The optimal power allocation for the state* (B,B) *is* $\alpha_{BB}^{\text{opt}} = 0$.

This is true for any set of parameters ν_a, ν_b, P_A, P_B, even if $P_B \to 1$ and $\nu_b \gg \nu_a$. Inherently, a penalty occurs when trying to exploit the double permissive state.

Corollary 3. *Under the optimal power allocation,* $\alpha^{\text{opt}}(\alpha_{AA}) = 1 - \alpha_1^{\text{opt}}(\alpha_{AA})/(1-\alpha_{AA})$.

This removes a degree of freedom in the optimization problem. Using these corollaries, and the notation $\boldsymbol{\alpha}' = [\alpha, \alpha_{AA}, \alpha_{cr}, \alpha_{BB}]^T$ instead of $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]^T$, we have:

Theorem 4. *The maximal sum rate of the symmetric two-parallel two-state channel over all allocations $\boldsymbol{\alpha}' \in D'$ is*

$$R_{\text{avg}}^{\text{opt}} = 2(P_A + P_B)^2 \log(1 + \nu_a P) + \max_{0 \le \alpha_{AA} \le 1} \left\{ R_0(1-\alpha_{AA}) + R_1((1-\alpha_{AA}) \cdot (1-\alpha^{\text{opt}}(\alpha_{AA}))) \right\}$$

where

$$\alpha^{\text{opt}}(\alpha_{AA}) = \left[\min \left\{ 1, 1 - \frac{P_B^2 \nu_b - [(P_A+P_B)^2 - P_A^2]\nu_a}{2P_A \cdot P_B \cdot \nu_a \nu_b P (1-\alpha_{AA})} \right\} \right]_+ . \quad (11)$$

Denoting the argument of the maximization as α_{AA}^{opt}, the optimal power allocation vector is

$$\boldsymbol{\alpha}'^{\text{opt}} = [\alpha^{\text{opt}}(\alpha_{AA}), \alpha_{AA}^{\text{opt}}, 1 - \alpha_{AA}^{\text{opt}}, 0]^T .$$

Proof. Use Prop. 1 and note that $\alpha_1 = 1 - \alpha_{AA} - \alpha\alpha_{cr} = (1 - \alpha_{AA})(1 - \alpha)$ for the optimal alocation $\alpha_{BB} = 0$. ∎

D. Sub Optimal Schemes

For evaluation of the advantage of the joint α_{AA} and α, the following sub optimal schemes are introduced: a) independent broadcasting; b) privately broadcasting; and c) only common broadcasting.

Definition 5. *A scheme for which the encoder disjointly encodes different messages into each single channel of the parallel channel using the broadcast approach over the fading channel is denoted* ***independent broadcasting****.*

The broadcast approach for fading SISO channel (introduced in [8], elaborated in [2]) relies on two main operations: superposition coding by layering at the transmitter; and successive interference cancellation at the receiver. The maximal average sum rate of the symmetric two parallel two state channel under **independent broadcasting** is

$$R_{\text{avg}}^{\text{ind-bc,opt}} = 2(P_A + P_B) \log \left(\frac{1+\nu_a P}{1+\nu_a (1-\alpha^{\text{ind-bc,opt}}) P} \right) + 2P_B \log \left(1 + \nu_b (1 - \alpha^{\text{ind-bc,opt}}) P \right),$$

$$\alpha^{\text{bc,opt}} = \left[\min \left\{ 1, 1 - \frac{P_B \nu_b - (P_A+P_B)\nu_a}{P_A \nu_a \nu_b P} \right\} \right]_+ . \quad (12)$$

Definition 6. *A scheme for which no power is allocated for the common stream in the (B, A) and (A, B) states (message w_0) is denoted* ***privately broadcasting****.*

This scheme is equivalent to setting $\alpha = 0$ in Theorem 4, thus allocating encoding power from the common stream ($R_0 = 0$) to the other streams R_{AA}, R_{AB}, R_{BA} and R_{BB} which achieves optimality for

$$\alpha_{AA}^{\text{prv-bc,opt}} = \left[\min \left\{ 1, 1 - \frac{[P_B - P_A]\nu_b - [P_B + P_A]\nu_a}{2P_A \nu_a \nu_b P} \right\} \right]_+ .$$

Definition 7. *A scheme for which all of the crossed state power is allocated for common stream only (message w_0) and no power is allocated privately (no allocation for messages w_{AB} and w_{BA}) is denoted* ***only common broadcasting****.*

This scheme is equivalent to setting $\alpha = 1$ in Theorem 4, thus allocating encoding power from the private streams ($R_{AB} = R_{BA} = 0$) to the other streams R_{AA}, R_0 and R_{BB} which achieves optimality for

$$\alpha_{AA}^{\text{cmn-bc,opt}} = \left[\min \left\{ 1, 1 - \frac{[(P_A+P_B)^2 - P_A^2]\nu_b - [(P_A+P_B)^2 + P_A^2]\nu_a}{2P_A^2 \nu_a \nu_b P} \right\} \right]_+ .$$

E. Numerical Results

Fig. 4 demonstrates the optimality of the proposed scheme (Theorem 4). The selected metric is the part of each scheme as a fraction of ergodic capacity. It is always superior in comparison to the other sub-optimal schemes, and captures a large portion of the ergodic capacity which stands as the upper bound. The sub-optimal methods inferior or superior to other sub-optimal methods, dependent on the parameters set. Some parameters sets can make them coincide for all SNR values. The gap to ergodic capacity does not change much, indicating that most coding gain is achieved via one of the classical broadcasting, and the specific one is parameters-set dependent.

V. SISO Block Fading

A. SISO consecutive block encoding model

Consider a block fading channel, as depicted at Fig. 5. Each n discrete time samples, a message w is to be encoded into the sequence $\mathbf{X} \in \mathbb{C}^n$, which enters the single input single output block fading channel satisfying the power constraint $\mathbb{E}|X|^2 \le P$ where X is the single letter random variable representation of the vector $\mathbf{X}$ and P is the power constraint $\mathbf{Y} = H\mathbf{X} + \mathbf{N}$. The channel gain $H \in \mathbb{C}$ is fixed within the n length block, and changes in-between blocks according to a priori known statistics P_H in a memoryless fashion. A complex normal noise is added, i.i.d. per channel output sample. The decoder is fully aware of the block gain (by channel sounding using pilot symbols) and reconstructs the message $\hat{w}$. The encoder has no way to know the channel realizations, yet has knowledge regarding its statistics P_H.

This setting, when allowing consecutive blocks *variable-to-fixed coding* [3] joint encoding, is actually a variant of the parallel MIMO single user case, where the diversity is over time blocks. Any development done so far can be applied on this special case. By allowing coding over two blocks at a time, the parallel channel model described till this section holds for this channel as well. The drawback is additional latency, yet only in the length of a single block, which in some use cases can be justified for the boos of achievable average rate.

B. A comment on Whiting [20]

The result in Theorem 4 differs from the one presented in [20] for the two-parallel two state channel. In [20] it is chosen to transmit only common information to the pairs (A, B) and (B, A). [20, eq. (39)] clearly states that for the crossed states (A, B) and (B, A) only common rate is used without justification. It is further claimed that this is an expected rate upper bound for some power allocation. Our result fully coincides with [20, eq. (39)] for $\alpha = 1$ rather than as in (9). However, this work proves that $\alpha = 1$ is suboptimal, and does not yield the maximal average rate. Furthermore, [20] does not notice that $\alpha_{BB} = 0$, whereas in this paper it is shown analytically to be optimal in Corollary 2.

VI. Conclusion

The broadcast approach for the parallel MIMO two state block fading channel is studied. The optimal scheme based on the concept of El-Gamal's degraded broadcast product channel, requires transmission of both private and common streams on two states (A, B) or (B, A). The expected rate is maximized

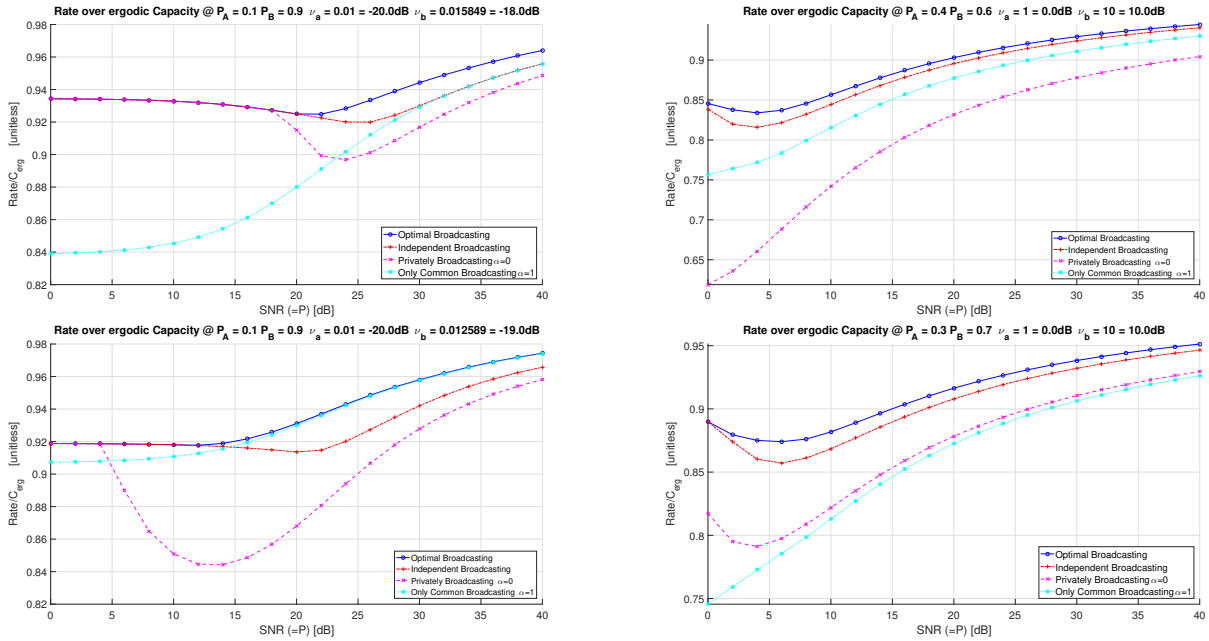


Fig. 4: Average sum rate as portion of the ergodic capacity of different schemes for several parameters-sets.

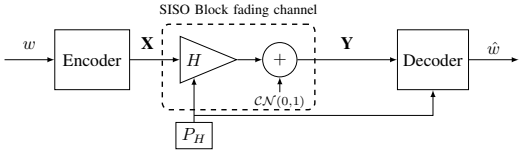


Fig. 5: The SISO n-length block fading channel and system.

analytically for layered transmission over the parallel channel. We demonstrate that the simple broadcast approach operating on each of the parallel channels separately achieves a significant portion of the optimal average rate. While the simple two-state parallel channel is considered here, the results apply directly to reduced latency constraints, that permit decoding over two fading blocks of a single two state fading channel. Evidently, extensions to a richer state spaces are called for, which may motivate new broadcast approach concepts of direct interest to future latency limited wireless systems. The framework considered motivates extensions where also the number of parallel channels received is random (adding thus a zero state), and this model may give rise to examine also secrecy constraints [22].

VII. Acknowledgment

This work has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630.

References

[1] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. on Telecomm.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.
[2] S. Shamai (Shitz) and A. Steiner, "A broadcast approach for a single user slowly fading MIMO channel," *IEEE Trans. on Inf. The.*, vol. 49, no. 10, pp. 2617–2635, Oct. 2003.
[3] S. Verdu and S. Shamai, "Variable-rate channel capacity," *IEEE Trans. on Inf. The.*, vol. 56, no. 6, pp. 2651–2667, June 2010.
[4] D. Tse and P. Viswanath, "Fundamentals of wireless communication," 2005.
[5] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading channels: Information theoretic and communication aspects," *IEEE Trans. on Inf. The.*, vol. 44, no. 6, pp. 2619–2692, October 1998.
[6] K. M. Cohen, A. Steiner, and S. Shamai, "The broadcast approach under mixed delay constraints," in *2012 IEEE Int. Symp. on Inf. The. Proceedings*, July 2012, pp. 209–213.
[7] G. Cocco, D. Gunduz, and C. Ibars, "Streaming transmission over block fading channels with delay constraint," *IEEE Trans. on Wir. Comm.*, vol. 12, no. 9, pp. 4315–4327, Sep. 2013.
[8] S. Shamai (Shitz), "A broadcast approach for the multiple-access slow fading channel," *IEEE Int. Symp. Info. The.*, p. 128, June 25-30 2000.
[9] M. Zamani and A. K. Khandani, "Broadcast approaches to the diamond channel," *IEEE Trans. on Inf. The.*, vol. 60, no. 1, pp. 623–642, Jan 2014.
[10] M. Shaqfeh, W. Mesbah, and H. Alnuweiri, "Utility maximization for layered transmission using the broadcast approach," *IEEE Trans. on Wir. Comm.*, vol. 11, no. 3, pp. 1228–1238, March 2012.
[11] J. Wang, J. Liang, and S. Muhaidat, "On the distortion exponents of layered broadcast transmission in multi-relay cooperative networks," *IEEE Trans. on Sig. Proc.*, vol. 58, no. 10, pp. 5340–5352, Oct 2010.
[12] Y. Yao and G. B. Giannakis, "Rate-maximizing power allocation in OFDM based on partial channel knowledge," *IEEE Trans. on Wir. Comm.*, vol. 4, no. 3, pp. 1073–1083, May 2005.
[13] C. Tian, A. Steiner, S. Shamai, and S. Diggavi, "Expected distortion for gaussian source with a broadcast transmission strategy over a fading channel," in *Workshop on Inf. The. for Wireless Networks*, July 2007.
[14] J. W. Yoo, T. Liu, S. S. (Shitz), and C. Tian, "Worst-case expected-capacity loss of slow-fading channels," *IEEE Trans. on Inf. The.*, vol. 59, no. 6, pp. 3764–3779, June 2013.
[15] A. El Gamal, "Capacity of the product and sum of two unmatched broadcast channels," *Prob. Pere. Inf.*, vol. 16, no. 1, pp. 1–16, Jan.-March 1980 (English Translation).
[16] T. Cover, "Broadcast channels," *IEEE Trans. on Inf. The.*, vol. 18, no. 1, pp. 2–14, Jan. 1972.
[17] T. M. Cover, "Comments on broadcast channels," *IEEE Trans. on Inf. The.*, vol. 44, no. 6, pp. 2524–2530, Oct 1998.
[18] A. Steiner and S. Shamai, "Multi-layer broadcasting over a block fading mimo channel," *IEEE Trans. on Wir. Comm.*, vol. 6, no. 11, pp. 3937–3945, November 2007.
[19] A. T. M. Zohdy and S. S. (Shitz), "Broadcast approach to multiple access with local CSIT," *submitted to the IEEE Trans. on. Comm.*
[20] P. A. Whiting and E. M. Yeh, "Broadcasting over uncertain channels with decoding delay constraints," *IEEE Trans. Info. The.*, vol. 52, no. 3, pp. 904–921, March 2006.
[21] S. Kazemi and A. Tajer, "Multiaccess communication via a broadcast approach adapted to the multiuser channel," *IEEE Trans. on. Comm.*, vol. 66, no. 8, pp. 3341–3353, Aug 2018.
[22] Y. Liang, L. Lai, H. V. Poor, and S. Shamai, "A broadcast approach for fading wiretap channels," *IEEE Trans. on Inf. The.*, vol. 60, no. 2, pp. 842–858, Feb 2014.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

On the Per-User Probability of Error in Gaussian Many-Access Channels

Jithin Ravi[†‡] and Tobias Koch[†‡]

[†]Signal Theory and Communications Department, Universidad Carlos III de Madrid, 28911, Leganés, Spain
[‡]Gregorio Marañón Health Research Institute, 28007, Madrid, Spain.
Emails: {rjithin,koch}@tsc.uc3m.es

***Abstract*—We consider a Gaussian multiple-access channel where the number of users grows with the blocklength n. For this setup, the maximum number of bits per unit-energy that can be transmitted reliably as a function of the order of growth of the users is analyzed. For the *per-user probability of error*, we show that if the number of users grows sublinearly with the blocklength, then each user can achieve the capacity per unit-energy of the Gaussian single-user channel. Conversely, if the number of users grows at least linearly with the blocklength, then the capacity per unit-energy is zero. Thus, there is a sharp transition between orders of growth where interference-free communication is feasible and orders of growth where reliable communication at a positive rate per unit-energy is infeasible. The same observation was made by Ravi and Koch (*Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2019) when the *per-user probability of error* is replaced by the *joint probability of error*, with the difference that the transition threshold is located at $n/\log n$ rather than at n. We further discuss the rates per unit-energy that can be achieved if one allows for a non-vanishing error probability.**

I. Introduction

Recently, Chen *et al.* [1] introduced the many-access channel (MnAC) as a multiple-access channel (MAC) where the number of users grows with the blocklength. The MnAC model is motivated by systems consisting of a single receiver and many transmitters, the number of which is comparable or even larger than the blocklength. This situation may occur, *e.g.*, in a machine-to-machine communication system with many thousands of devices in a given cell. In [1], Chen *et al.* considered a Gaussian MnAC with k_n users and determined the number of messages M_n each user can transmit reliably with a codebook of average power not exceeding P. Since then, MnACs have been studied in various papers under different settings. For example, Polyanskiy [2] considered a Gaussian MnAC where the number of active users grows linearly in the blocklength and each user's payload is fixed. Zadik *et al.* [3] presented improved bounds on the tradeoff between user density and energy-per-bit of this channel. Generalizations to quasi-static fading MnACs can be found in [4]–[7]. Shahi *et al.* [8] studied the capacity region of strongly asynchronous MnACs. Ravi and Koch [9], [10] characterized the capacity per unit-energy of Gaussian MnACs as a function of the order of growth of the number of users.

Roughly, papers on the MnAC can be divided into two groups: The first group, including [1], [8]–[10], considers a classical information-theoretic setting where the number of messages M_n transmitted by each user grows with n and the probability of a decoding error is defined as

$$P_{e,J}^{(n)} \triangleq \Pr\{(\hat{W}_1,\ldots,\hat{W}_{k_n}) \neq (W_1,\ldots,W_{k_n})\}. \quad (1)$$

Here, W_i denotes the message transmitted by user i and $\hat{W}_i$ denotes the decoder's estimate of this message. The second group, including [2]–[7], assumes that M_n is fixed and defines the probability of a decoding error as

$$P_{e,A}^{(n)} \triangleq \frac{1}{k_n}\sum_{i=1}^{k_n}\Pr\{\hat{W}_i \neq W_i\}. \quad (2)$$

The error probability $P_{e,A}^{(n)}$ is sometimes referred to as *per-user probability of error*. In this paper, we shall refer to it as *average probability of error (APE)*. In contrast, we shall refer to $P_{e,J}^{(n)}$ as *joint probability of error (JPE)*.

This paper aims at a better understanding of the implications of the above assumptions on the *capacity per unit-energy*, defined as the largest number of bits per unit-energy that can be transmitted with vanishing error probability [11]. To this end, we consider the APE and study the behavior of the capacity per unit-energy of Gaussian MnACs as a function of the order of growth of the number of users k_n. We demonstrate that, if the order of growth of k_n is sublinear, then each user can achieve the capacity per unit-energy $\frac{\log e}{N_0}$ of the single-user Gaussian channel (where $N_0/2$ is the noise power). Conversely, if the growth of k_n is linear or superlinear, then the capacity per unit-energy is zero. Thus, there is a sharp transition between orders of growth where interference-free communication is feasible and orders of growth where reliable communication at a positive rate is infeasible. The same behavior has been observed for the JPE, but with the transition threshold located at $n/\log n$ [9], [10]. Consequently, relaxing the error probability from JPE to APE merely shifts the transition threshold from $n/\log n$ to n.

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). T. Koch has further received funding from the Spanish Ministerio de Economía y Competitividad under Grants RYC-2014-16332 and TEC2016-78434-C3-3-R (AEI/FEDER, EU).

Our results imply that, when the number of users grows linearly in n, as assumed, *e.g.*, in [2]–[7], the capacity per unit-energy is zero, irrespective of whether one considers the APE or the JPE. We further show that, for the JPE, this holds true even if we allow for a non-vanishing error probability. We thus conclude that, when the number of users of the Gaussian MnAC grows linearly in n, a positive rate per unit-energy can be achieved only if one considers the APE and one allows for a non-vanishing error probability.

The rest of the paper is organized as follows. In Section II, we introduce the system model. In Section III, we characterize the capacity per unit-energy of the Gaussian MnAC with APE and compare it to the capacity per unit-energy of the Gaussian MnAC with JPE obtained in [9], [10]. Section IV discusses the rates per unit-energy that can be achieved if one allows for a non-vanishing error probability. Section V concludes the paper with a discussion of the obtained results.

II. Problem Formulation and Definitions

A. Model and Definitions

Suppose there are k users that wish to transmit their messages $W_i, i = 1, \ldots, k$, which are assumed to be independent and uniformly distributed on $\{1, \ldots, M_n^{(i)}\}$, to one common receiver. To achieve this, they send a codeword of n symbols over the channel, where n is referred to as the *blocklength*. We consider a many-access scenario where the number of users k grows with n, hence, we denote it as k_n. We further consider a Gaussian channel model where, for k_n users and blocklength n, the received vector $\mathbf{Y}$ is given by

$$\mathbf{Y} = \sum_{i=1}^{k_n} \mathbf{X}_i(W_i) + \mathbf{Z}.$$

Here, $\mathbf{X}_i(W_i)$ is the length-n transmitted codeword by user i for message W_i and $\mathbf{Z}$ is a vector of n i.i.d. Gaussian components $Z_j \sim \mathcal{N}(0, N_0/2)$ independent of $\mathbf{X}_i$.

We next introduce the notion of an $\left(n, \left\{M_n^{(\cdot)}\right\}, \left\{E_n^{(\cdot)}\right\}, \epsilon\right)$ code. We use the subscripts "J" and "A" to indicate whether the JPE or the APE is considered.

Definition 1: For $0 \leq \epsilon < 1$, an $\left(n, \left\{M_n^{(\cdot)}\right\}, \left\{E_n^{(\cdot)}\right\}, \epsilon\right)_J$ code for the Gaussian MnAC consists of:

1) k_n encoding functions $f_i : \{1, \ldots, M_n^{(i)}\} \to \mathcal{X}^n$, which map user i's message to the codeword $\mathbf{X}_i(W_i)$, satisfying the energy constraint

$$\sum_{j=1}^{n} x_{ij}^2(w_i) \leq E_n^{(i)}. \tag{3}$$

Here, x_{ij} is the jth symbol of the transmitted codeword.

2) A decoding function $g : \mathcal{Y}^n \to \{M_n^{(\cdot)}\}$, which maps the received vector $\mathbf{Y}$ to the messages of all users and whose JPE, defined in (1), satisfies $P_{e,J}^{(n)} \leq \epsilon$.

An $\left(n, \left\{M_n^{(\cdot)}\right\}, \left\{E_n^{(\cdot)}\right\}, \epsilon\right)_A$ code for the Gaussian MnAC consists of the same encoding functions f_i, $i = 1, \ldots, k_n$ and a decoding function $g : \mathcal{Y}^n \to \{M_n^{(\cdot)}\}$ whose APE, defined in (2), satisfies $P_{e,A}^{(n)} \leq \epsilon$.

We shall say that the $(n, \{M_n^{(\cdot)}\}, \{E_n^{(\cdot)}\}, \epsilon)_\xi$ code ($\xi \in \{J, A\}$) is *symmetric* if $M_n^{(i)} = M_n$ and $E_n^{(i)} = E_n$ for all $i = 1, \ldots, k_n$. For compactness, we denote a symmetric code by $(n, M_n, E_n, \epsilon)_\xi$, $\xi \in \{J, A\}$. In this paper, we restrict ourselves to symmetric codes.

Definition 2: Let $\xi \in \{J, A\}$. For a symmetric code, the rate per unit-energy $\dot{R}^\xi$ is said to be ϵ-achievable if for every $\alpha > 0$ there exists an n_0 such that if $n \geq n_0$, then an $(n, M_n, E_n, \epsilon)_\xi$ code can be found whose rate per unit-energy satisfies $\frac{\log M_n}{E_n} > \dot{R}^\xi - \alpha$. Furthermore, $\dot{R}^\xi$ is said to be achievable if it is ϵ-achievable for all $0 < \epsilon < 1$. The ϵ-capacity per unit-energy $\dot{C}_\epsilon^\xi$ is the supremum of all ϵ-achievable rates per unit-energy. Similarly, the capacity per unit-energy $\dot{C}^\xi$ is the supremum of all achievable rates per unit-energy.

Remark 1: In [11, Def. 2], a rate per unit-energy $\dot{R}$ is said to be ϵ-achievable if for every $\alpha > 0$ there exists an E_0 such that if $E \geq E_0$, then an (n, M, E, ϵ) code can be found whose rate per unit-energy satisfies $\frac{\log M}{E} > \dot{R} - \alpha$. Thus, the energy E is supposed to be large rather than the blocklength n, as required in Definition 2. For the MnAC, where the number of users grows with the blocklength, we believe it is more natural to impose that $n \to \infty$. Definition 2 is also consistent with the definition of energy-per-bit in [2], [3]. Further note that, for the capacity per unit-energy, where a vanishing error probability is required, our definition is actually equivalent to [11, Def. 2]. Indeed, as observed in [9, Lemma 1] for the JPE, and as we argue below for the APE, a vanishing error probability can only be achieved if $E_n \to \infty$ as $n \to \infty$.

B. Order Notations

Let $\{a_n\}$ and $\{b_n\}$ be two sequences of nonnegative real numbers. We write $a_n = o(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. We further write $a_n = \Omega(b_n)$ if $\liminf_{n\to\infty} \frac{a_n}{b_n} > 0$ and $a_n = \omega(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = \infty$.

III. Capacity Per Unit-Energy of Gaussian Many-Access Channels

In this section, we discuss the behavior of the capacity per unit-energy as a function of the growth of k_n. Specifically, in Subsection III-A we review the results for the case of JPE that we originally presented in [9], [10]. In Subsection III-B, we then present one of the main results of this paper, a characterization of the capacity per unit-energy as a function of the growth of the number of users for APE (Theorem 2). The proof of Theorem 2 is given in Subsection III-C.

A. Joint Probability of Error

Theorem 1: The capacity per unit-energy $\dot{C}^J$ for JPE has the following behavior:

1) If $k_n = o(n/\log n)$, then $\dot{C}^J = \frac{\log e}{N_0}$.
2) If $k_n = \omega(n/\log n)$, then $\dot{C}^J = 0$.

Proof: Part 1) is [9, Th. 2]. Part 2) is [9, Th. 1]. ∎

In words, if the order of growth is below $n/\log n$, then each user can achieve the single-user capacity per unit-energy.

Conversely, for any order of growth above $n/\log n$, no positive rate per unit-energy is achievable. Thus, there is a sharp transition between orders of growth where interference-free communication is feasible and orders of growth where reliable communication at a positive rate per unit-energy is infeasible.

B. Average Probability of Error

Theorem 2: The capacity per unit-energy $\dot{C}^A$ for APE has the following behavior:

1) If $k_n = o(n)$, then $\dot{C}^A = \frac{\log e}{N_0}$.
2) If $k_n = \Omega(n)$, then $\dot{C}^A = 0$.

Proof: See Section III-C. ∎

We observe a similar behavior as for JPE. Again, there is a sharp transition between orders of growth where interference-free communication is feasible and orders of growth where reliable communication at a positive rate per unit-energy is infeasible. The main difference is that the transition threshold is shifted from $n/\log n$ to n.

C. Proof of Theorem 2

Part 1): We first argue that $P_{e,A}^{(n)} \to 0$ only if $E_n \to \infty$, and that in this case $\dot{C}^A \leq \frac{\log e}{N_0}$. Indeed, let $P_i \triangleq \Pr\{\hat{W}_i \neq W_i\}$ denote the probability that message W_i is decoded erroneously. We then have that $P_{e,A}^{(n)} \geq \min_i P_i$. Furthermore, P_i is lower-bounded by the error probability of the Gaussian single-user channel, since a single-user channel can be obtained from the MnAC if a genie informs the receiver about the codewords transmitted by users $j \neq i$. By applying the lower bound [12, eq. (30)] on the error probability of the Gaussian single-user channel, we thus obtain

$$P_{e,A}^{(n)} \geq Q\left(\sqrt{\frac{2E_n}{N_0}}\right), \quad M_n \geq 2. \tag{4}$$

Hence $P_{e,A}^{(n)} \to 0$ only if $E_n \to \infty$. As mentioned in Remark 1, when E_n tends to infinity as $n \to \infty$, the capacity per unit-energy $\dot{C}^A$ coincides with the capacity per unit-energy defined in [11], which for the Gaussian single-user channel is given by $\frac{\log e}{N_0}$ [11, Ex. 3]. Furthermore, if $P_{e,A}^{(n)} \to 0$ as $n \to \infty$, then there exists at least one user i for which $P_i \to 0$ as $n \to \infty$. By the above genie argument, this user's rate per unit-energy is upper-bounded by the capacity per unit-energy of the Gaussian single-user channel. Since for the class of symmetric codes considered in this paper each user transmits at the same rate per unit-energy, we conclude that $\dot{C}^A \leq \frac{\log e}{N_0}$.

We next show that any rate per unit-energy $\dot{R}^A < \frac{\log e}{N_0}$ is achievable. For a given $0 < \epsilon < 1$, let $0 < \epsilon' < \epsilon$, and define

$$A_n \triangleq \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbf{1}(\hat{W}_i \neq W_i)$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. Further define $\mathcal{A}_n \triangleq \{0, 1/k_n, \ldots, 1\}$ and $\mathcal{A}_n^{\epsilon'} \triangleq \{a \in \mathcal{A}_n \colon a \geq \epsilon'\}$. Noting that $P_{e,A}^{(n)} = \mathsf{E}[A_n]$, we then obtain that

$$\begin{aligned} P_{e,A}^{(n)} &= \sum_{a \in \mathcal{A}_n} a \Pr\{A_n = a\} \\ &= \sum_{a \in \mathcal{A}_n \setminus \mathcal{A}_n^{\epsilon'}} a \Pr\{A_n = a\} + \sum_{a \in \mathcal{A}_n^{\epsilon'}} a \Pr\{A_n = a\} \\ &\leq \epsilon' + \sum_{a \in \mathcal{A}_n^{\epsilon'}} \Pr\{A_n = a\} \end{aligned} \tag{5}$$

where we used that $a \leq \epsilon'$ for $a \in \mathcal{A}_n \setminus \mathcal{A}_n^{\epsilon'}$ and $a \leq 1$ for $a \in \mathcal{A}_n^{\epsilon'}$. Next we show that if $\dot{R}^A < \frac{\log e}{N_0}$, then

$$\lim_{n \to \infty} \sum_{a \in \mathcal{A}_n^{\epsilon'}} \Pr\{A_n = a\} = 0. \tag{6}$$

It then follows from (5) that $P_{e,A}^{(n)} \leq \epsilon$ for sufficiently large n and all $0 < \epsilon < 1$. Thus, any rate per unit-energy $\dot{R}^A < \frac{\log e}{N_0}$ is achievable which proves Part 1) of Theorem 2.

To prove (6), we need the following lemma.

Lemma 1: For any arbitrary $0 < \rho \leq 1$, we have

$$\Pr\{A_n = a\} \leq \binom{k_n}{ak_n} M_n^{ak_n\rho} e^{-nE_0(a,\rho)}, \quad a \in \mathcal{A}_n \setminus \{0\}$$

where

$$E_0(a,\rho) \triangleq \frac{\rho}{2} \ln\left(1 + \frac{a2k_nE_n}{n(\rho+1)N_0}\right).$$

Proof: See [13, Th. 2]. ∎

Using Lemma 1, we can upper-bound the second term on the right-hand side (RHS) of (5) as

$$\begin{aligned} &\sum_{a \in \mathcal{A}_n^{\epsilon'}} \Pr\{A_n = a\} \\ &\leq \left(\max_{a \in \mathcal{A}_n^{\epsilon'}} \exp\left[-nE_0(a,\rho) + \ln M_n^{a\rho k_n}\right]\right) \sum_{a \in \mathcal{A}_n^{\epsilon'}} \binom{k_n}{ak_n} \\ &\leq \max_{a \in \mathcal{A}_n^{\epsilon'}} \exp\left[-E_n f_n(a,\rho)\right] \end{aligned} \tag{7}$$

where

$$f_n(a,\rho) \triangleq \frac{nE_0(a,\rho)}{E_n} - \frac{a\rho k_n \ln M_n}{E_n} - \frac{k_n \ln 2}{E_n}.$$

We next choose $E_n = (\ln(n/k_n)k_n/n)^{-1}$. This implies that $E_n \to \infty$ and $E_nk_n/n \to 0$ as $n \to \infty$ since, by the theorem's assumption, $k_n = o(n)$. We then show that, for this choice of E_n and $\dot{R}^A = \frac{\log e}{(1+\rho)N_0} - \delta$ (for some arbitrary $0 < \delta < \frac{\log e}{(1+\rho)N_0}$), we have

$$\liminf_{n \to \infty} \min_{a \in \mathcal{A}_n^{\epsilon'}} f_n(a,\rho) > 0. \tag{8}$$

Thus, for $\dot{R}^A = \frac{\log e}{(1+\rho)N_0} - \delta$, the RHS of (7) vanishes as $n \to \infty$. Since $0 < \rho < 1$ and $\delta > 0$ are arbitrary, (6) follows.

To obtain (8), we first show that, for any fixed value of ρ and our choices of E_n and $\dot{R}^A$,

$$\liminf_{n \to \infty} \frac{df_n(a,\rho)}{da} > 0, \quad \epsilon' \leq a \leq 1. \tag{9}$$

Hence

$$\liminf_{n\to\infty} \min_{a\in\mathcal{A}_n^{\epsilon'}} f_n(a,\rho) \geq \liminf_{n\to\infty} f_n(\epsilon',\rho). \tag{10}$$

Indeed, basic algebraic manipulations yield for $\epsilon' \leq a \leq 1$

$$\frac{df_n(a,\rho)}{da} \geq \rho k_n \left[\frac{1}{1+\frac{2k_nE_n}{n(\rho+1)N_0}} \frac{1}{(1+\rho)N_0} - \frac{\dot{R^A}}{\log e}\right]. \tag{11}$$

Recall that, for the given choice of E_n, we have $\frac{k_nE_n}{n} \to 0$ as $n \to \infty$. It follows that the bracketed term in (11) tends to $\frac{\delta}{\log e}$ as $n \to \infty$. This proves (9).

We next show that the RHS of (10) is positive for every $0 < \rho < 1$. Let

$$i_n(\epsilon',\rho) \triangleq \frac{nE_0(\epsilon',\rho)}{E_n}$$
$$j_n(\epsilon',\rho) \triangleq \frac{\epsilon'\rho k_n \dot{R^A}}{\log e}$$
$$h_n \triangleq \frac{k_n \ln 2}{E_n}.$$

For our choices of E_n and $\dot{R^A}$, we have that $h_n/j_n(\epsilon',\rho) \to 0$ as $n \to \infty$. Consequently,

$$\begin{aligned}\liminf_{n\to\infty} f_n(\epsilon',\rho) &\geq \liminf_{n\to\infty} j_n(\epsilon',\rho) \liminf_{n\to\infty} \frac{f_n(\epsilon',\rho)}{j_n(\epsilon',\rho)} \\ &= \liminf_{n\to\infty} j_n(\epsilon',\rho)\left\{\liminf_{n\to\infty} \frac{i_n(\epsilon',\rho)}{j_n(\epsilon',\rho)} - 1\right\}.\end{aligned}$$

Note that $j_n(\epsilon',\rho) \geq \epsilon'\rho\dot{R^A}/\log e$, which is bounded away from zero for our choice of $\dot{R^A}$ and $\delta < \frac{\log e}{(1+\rho)N_0}$. The RHS of (10) is thus positive if $\liminf_{n\to\infty} i_n(\epsilon',\rho)/j_n(\epsilon',\rho) > 1$, which is what we show next. Indeed, we have for our choice of E_n and $k_n = o(n)$ that

$$\lim_{n\to\infty} \frac{i_n(\epsilon',\rho)}{j_n(\epsilon',\rho)} = \frac{\log e}{(1+\rho)N_0\dot{R^A}}.$$

For our choice of $\dot{R^A}$, this is strictly larger than 1. We thus conclude that the RHS of (10) is positive, from which (8), and hence also (6), follows. This proves Part 1) of Theorem 2.

Part 2): Fano's inequality yields that

$$\log M_n \leq 1 + P_i \log M_n + I(W_i;\hat{W}_i)$$

for $i = 1,\ldots,k_n$. Averaging over all i's then gives

$$\begin{aligned}\log M_n &\leq 1 + \frac{1}{k_n}\sum_{i=1}^{k_n} P_i \log M_n + \frac{1}{k_n} I(\mathbf{W};\hat{\mathbf{W}}) \\ &\leq 1 + P_{e,A}^{(n)} \log M_n + \frac{1}{k_n} I(\mathbf{X};\mathbf{Y}) \\ &\leq 1 + P_{e,A}^{(n)} \log M_n + \frac{n}{2k_n}\log\left(1+\frac{2k_nE_n}{nN_0}\right) \end{aligned} \tag{12}$$

where $\mathbf{X} \triangleq (\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_{k_n})$. Here, the first inequality follows because the messages $W_i, i = 1,\ldots,k_n$ are independent and because conditioning reduces entropy, the second inequality follows from the definition of $P_{e,A}^{(n)}$ and the data processing inequality, and the third inequality follows by upper-bounding $I(\mathbf{X};\mathbf{Y})$ by $\frac{n}{2}\log\big(1+\frac{2k_nE_n}{nN_0}\big)$.

Dividing both sides of (12) by E_n, and solving the inequality for $\dot{R^A}$, we obtain the upper bound

$$\dot{R^A} \leq \frac{\frac{1}{E_n} + \frac{n}{2k_nE_n}\log(1+\frac{2k_nE_n}{nN_0})}{1-P_{e,A}^{(n)}}. \tag{13}$$

As argued at the beginning of the proof of Part 1), we have $P_{e,A}^{(n)} \to 0$ only if $E_n \to \infty$. If $k_n = \Omega(n)$, then this implies that $k_nE_n/n \to \infty$ as $n \to \infty$. It thus follows from (13) that, if $k_n = \Omega(n)$, then $\dot{C}^A = 0$, which is Part 2) of Theorem 1.

IV. Non-Vanishing Error Probability

In this section, we briefly discuss how the largest achievable rate per unit-energy changes if we allow for a non-vanishing error probability. With the help of the following example, we first argue that when the number of users is bounded in n, then a simple orthogonal-access scheme achieves an ϵ-achievable rate per unit-energy that can be strictly larger than the single-user capacity per unit-energy $\frac{\log e}{N_0}$.

Example 1: Consider a k-user Gaussian MAC with normalized noise variance $N_0/2 = 1$ and where the number of users is independent of n. Suppose that each user transmits one out of two messages ($M_n = 2$) with energy $E_n = 1$ by following an orthogonal-access scheme where each user gets one channel use and remains silent in the remaining channel uses. In this channel use, each user transmits either $+1$ or -1 to convey its message. Since the access scheme is orthogonal, the receiver can perform independent decoding for each user, which yields $P_i = Q(1)$. Consequently, we can achieve the rate per unit-energy $\frac{\log M_n}{E_n} = 1$ at APE $P_{e,A}^{(n)} = Q(1)$ and at JPE $P_{e,J}^{(n)} = 1-(1-Q(1))^k$ [9, eq. (6)]. Thus, for some $0 < \epsilon < 1$, we have that $\dot{C}_\epsilon^\xi > \frac{\log e}{N_0}$, $\xi \in \{J,A\}$.

Remark 2: A crucial ingredient in the above scheme is that the energy E_n is bounded in n. Indeed, it follows from [12, Th. 3] that if $E_n \to \infty$ as $n \to \infty$, then the ϵ-capacity per unit-energy of the Gaussian single-user channel is equal to $\frac{\log e}{N_0}$, irrespective of $0 < \epsilon < 1$. The genie argument provided at the beginning of Section III-C then yields that the same is true for the Gaussian MnAC.

In the rest of this section, we discuss the ϵ-capacity per unit-energy when the number of users k_n tends to infinity as n tends to infinity. Specifically, in Subsection IV-A we discuss the ϵ-capacity per unit-energy for JPE as a function of the order of growth of the number of users. In Subsection IV-B, we briefly discuss the ϵ-capacity per unit-energy for APE when k_n grows linearly in n.

A. Non-Vanishing JPE

Theorem 3: The ϵ-capacity per unit-energy $\dot{C}_\epsilon^J$ for JPE has the following behavior:

1) If $k_n = \omega(1)$ and $k_n = o(n/\log n)$, then $\dot{C}_\epsilon^J = \frac{\log e}{N_0}$ for every $0 < \epsilon < 1$.
2) If $k_n = \omega(n/\log n)$, then $\dot{C}_\epsilon^J = 0$ for every $0 < \epsilon < 1$.

Proof: We first prove Part 1). It follows from [9, eq. (20)] that, for $M_n \geq 2$,[1]

$$P_{e,J}^{(n)} \geq 1 - \frac{64E_n/N_0 + \log 2}{\log k_n}. \tag{14}$$

This implies that $P_{e,J}^{(n)}$ tends to one unless $E_n = \Omega(\log k_n)$. Since by the theorem's assumption $k_n = \omega(1)$, it follows that $E_n \to \infty$ is necessary to achieve a JPE strictly smaller than one. As argued in Remark 2, if $E_n \to \infty$ as $n \to \infty$, then the ϵ-capacity per unit-energy of the Gaussian MnAC cannot exceed the single-user capacity per unit-energy $\frac{\log e}{N_0}$. Furthermore, by Theorem 1, if $k_n = o(n/\log n)$ then any rate per unit-energy satisfying $\dot{R}^J < \frac{\log e}{N_0}$ is achievable, hence it is also ϵ-achievable. We thus conclude that, if $k_n = \omega(1)$ and $k_n = o(n/\log n)$, then $\dot{C}_\epsilon^J = \frac{\log e}{N_0}$ for every $0 < \epsilon < 1$.

To prove Part 2), we use that, by Fano's inequality, we can upper-bound $\dot{R}^J$ as [9, eq. (2)]

$$\dot{R}^J \leq \frac{\frac{1}{k_n E_n} + \frac{n}{2k_n E_n}\log(1 + \frac{2k_n E_n}{nN_0})}{1 - P_{e,J}^{(n)}}. \tag{15}$$

By (14), $P_{e,J}^{(n)}$ tends to one unless $E_n = \Omega(\log k_n)$. For $k_n = \omega(n/\log n)$, this implies that $k_n E_n/n \to \infty$ as $n \to \infty$, so the RHS of (15) vanishes as n tends to infinity. We thus conclude that, if $k_n = \omega(n/\log n)$, then $\dot{C}_\epsilon^J = 0$ for every $0 < \epsilon < 1$. ∎

B. Non-Vanishing APE

For the APE, we restrict ourselves to the case where $k_n = \mu n$ for some $\mu > 0$, since it is a common assumption in the analysis of MnACs; see, *e.g.*, [2]–[7]. By inspecting the proof of Part 1) of Theorem 2, one can show that, for every $\mu > 0$ and $0 < \epsilon' < \epsilon < 1$, there exists an E independent of n and a $0 < \rho \leq 1$ such that the RHS of (7) vanishes with n for some positive $\dot{R}^A$. By (5), it then follows that $P_{e,A}^{(n)} \leq \epsilon$ for sufficiently large n, hence, there exists a positive rate per unit-energy $\dot{R}^A$ that is ϵ-achievable.

While (5) and (7) yield an upper bound on $P_{e,A}^{(n)}$ that is sufficient to demonstrate the qualitative behavior of $\dot{C}_\epsilon^A$, this bound is looser than the bounds obtained in [2], [3]. Specifically, [2], [3] derived bounds on the minimum energy-per-bit $\mathcal{E}^*(M, \mu, \epsilon)$ required to send M messages at an APE not exceeding ϵ when the number of users is given by $k_n = \mu n$. Since the rate per unit-energy is the inverse of the energy-per-bit, these bounds also apply to $\dot{C}_\epsilon^A$. The achievability and converse bounds presented in [3] further suggest that there exists a critical user density μ below which interference-free communication is feasible. This conjectured effect can be confirmed when each user sends only one bit ($M = 2$), since in this case $\mathcal{E}^*(M, \mu, \epsilon)$ can be evaluated in closed form for $\mu \leq 1$. For simplicity, assume that $N_0/2 = 1$. Then,

$$\mathcal{E}^*(2, \mu, \epsilon) = \left(\max\{0, Q^{-1}(\epsilon)\}\right)^2, \quad 0 \leq \mu \leq 1. \tag{16}$$

Indeed, that $\mathcal{E}^*(2, \mu, \epsilon) \geq (\max\{0, Q^{-1}(\epsilon)\})^2$ follows from (4). Furthermore, when $\mu \leq 1$, applying the orthogonal-access scheme presented in Example 1 with energy $(\max\{0, Q^{-1}(\epsilon)\})^2$ achieves $P_{e,A}^{(n)} = \epsilon$. Observe that the RHS of (16) does not depend on μ and agrees with the minimum energy-per-bit required to send one bit over the Gaussian single-user channel with error probability ϵ. Thus, when $\mu \leq 1$, we can send one bit free of interference.

[1] A similar bound was presented in [14, p. 84] for the case where $M_n = 2$.

V. Conclusion

A common assumption in the analysis of MnACs is that the number of users grows linearly with the blocklength. Theorems 1 and 2 imply that in this case the capacity per unit-energy is zero, irrespective of whether one considers the APE or the JPE. Theorem 3 further demonstrates that, for the JPE, this holds true even if we allow for a non-vanishing error probability. The situation changes for the APE. Here a positive rate per unit-energy can be achieved if one allows for a non-vanishing error probability. Another crucial assumption is that the energy E_n and payload $\log M_n$ are bounded in n. Indeed, for $k_n = \mu n$, the RHS of (13) vanishes as E_n tends to infinity, so when $E_n \to \infty$ no positive rate per unit-energy is ϵ-achievable. Moreover, for $k_n = \mu n$ and a bounded E_n, (12) implies that the payload $\log M_n$ is bounded, too. We conclude that the arguably most common assumptions in the literature on MnACs—linear growth of the number of users, a non-vanishing APE, and a fixed payload—are the only set of assumptions under which a positive rate per unit-energy is achievable, unless we consider nonlinear growths of k_n.

References

[1] X. Chen, T. Y. Chen, and D. Guo, "Capacity of Gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3516–3539, Jun. 2017.

[2] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory*, Aachen, Germany, Jun. 2017.

[3] I. Zadik, Y. Polyanskiy, and C. Thrampoulidis, "Improved bounds on Gaussian MAC and sparse regression via Gaussian inequalities," in *Proc. IEEE Int. Symp. Inf. Theory*, Paris, France, Jul. 2019.

[4] S. S. Kowshik and Y. Polyanskiy, "Quasi-static fading MAC with many users and finite payload," in *Proc. IEEE Int. Symp. Inf. Theory*, Paris, France, Jul. 2019.

[5] ——, "Fundamental limits of many-user MAC with finite payloads and fading," `arXiv: 1901.06732 [cs.IT]`, Jan. 2019.

[6] S. S. Kowshik, K. V. Andreev, A. Frolov, and Y. Polyanskiy, "Energy efficient random access for the quasi-static fading MAC," in *Proc. IEEE Int. Symp. Inf. Theory*, Paris, France, Jul. 2019.

[7] ——, "Energy efficient coded random access for the wireless uplink," `arXiv: 1907.09448 [cs.IT]`, Jul. 2019.

[8] S. Shahi, D. Tuninetti, and N. Devroye, "The strongly asynchronous massive access channel," `arXiv: 1807.09934 [cs.IT]`, Jul. 2018.

[9] J. Ravi and T. Koch, "Capacity per unit-energy of Gaussian many-access channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Paris, France, Jul. 2019.

[10] ——, "Capacity per unit-energy of Gaussian many-access channels," `arXiv:1904.11742 [cs.IT]`, Apr. 2019.

[11] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1019–1030, Sep. 1990.

[12] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Minimum energy to send k bits through the Gaussian channel with and without feedback," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4880–4902, Aug. 2011.

[13] R. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inf. Theory*, vol. 31, no. 2, pp. 124–142, Mar. 1985.

[14] Y. Polyanskiy, "Information theoretic perspective on massive multiple-access," *in Short Course (slides) Skoltech Inst. of Tech., Moscow, Russia*, Jul. 2018.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Approximate Bit-wise MAP Detection For Greedy Sparse Signal Recovery Algorithms

Jeongmin Chae and Song-Nam Hong
Ajou University, Suwon, Korea,
email: {jmchae92 and snhong}@ajou.ac.kr

***Abstract*—A greedy algorithm is a fascinating choice in support recovery problem due to its easy implementation and lower complexity compared with other optimization-based algorithms. In this paper, we present a novel greedy algorithm, referred to as *bit-wise* maximum a posteriori (MAP) detector. In the proposed method, for each iteration, one includes the best index to a target support in the sense of maximizing a posteriori probability given an observation, support indices previously chosen, and a priori information on a sparse vector. In other words, the proposed method employs statistical information on a given sparse recovery system while the other greedy-based algorithms (e.g., orthogonal matching pursuit (OMP)) uses the correlation values in magnitude. We remark that the proposed method has much lower complexity than the (vector-wise) MAP, where the complexity of the former is linear with a sparsity level but the latter is exponential. We further reduce the complexity of the proposed method by efficiently computing a posteriori probability for each iteration. Via simulations, we demonstrate that the proposed method can outperform the other greedy algorithms based on correlations, by exploiting statistical information properly.**

***Index Terms*—Sparse vector recovery, compressed sensing, MAP detector, greedy algorithm.**

I. Introduction

An inverse problem is widely studied in which a vector signal $\mathbf{x} \in \mathbb{R}^N$ is recovered from a set of linear noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$, with an $M \times N$ measurement matrix $\mathbf{A}$. In particular when $M < N$ (i.e., under-determined system), the above problem has infinite solutions and thus it can be solved if some additional a priori information on $\mathbf{x}$ is available. In [1], [2], it has been proved that $\mathbf{x}$ can be exactly reconstructed with the a priori knowledge on the sparsity of $\mathbf{x}$ (i.e., $\|\mathbf{x}\|_0 = K$ with $K \ll N$), where K is referred to as the sparsity level. Also, the optimal sparse vector can be obtained by solving ℓ_0-minimization problem such as

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \eta, \tag{1}$$

where $\|\mathbf{x}\|_0$ is introduced to ensure the sparsity of $\mathbf{x}$. In general, the above ℓ_0-minimization is known to be NP-hard. Leveraging the idea of convex optimization, a well-established method, called LASSO, was proposed in which ℓ_1-norm is used as a convex-relaxation of ℓ_0-norm [3], [4]. LASSO can solve the sparse signal recovery problem with stability while it has polynomial bounded computational complexity.

A greedy approach seems to be attractive due to its lower complexity than convex-based algorithms in sparse signal reconstruction. The key idea of greedy-based algorithms is to estimate the support of a sparse signal vector in a sequential fashion, where for each iteration, one index is added to a target support by solving a sub-optimization problem. Since the sub-optimization problem has much lower complexity than the overall sparse signal recovery problem, the greedy approach can significantly reduce the computational complexity. Orthogonal matching pursuit (OMP) [5]–[7] is the most famous greedy approach where for each iteration, it identifies the best support index in the sense of correlations between column vectors in the measurement matrix and the residual vector. In addition, to overcome the downside of OMP, numerous advanced greedy algorithms have been proposed such as Compressive Sampling Matching Pursuit (CoSaMP) [9], Subspace Pursuit (SP) [10] and generalized OMP [11]. The fundamental concept of such advanced algorithms lies in the selection of multiple support indices for each iteration, which can decrease the probability for estimating incorrect support indices. In the above greedy algorithms, they only rely on the order statistics of the correlation values in magnitude to estimate support. However, it may not be optimal in the sense of support detection in probability depending on the statistical distributions of sparse signal vector and noise. Inspired by this, a greedy algorithm, named Bayesian matching pursuit (BMP), has been proposed in [12].

Our contributions: We propose a novel greedy algorithm, named *bit-wise* MAP detector, for sparse signal recovery problem. The key idea of the proposed algorithm is that for each iteration, one adds the best index to a target support in the sense of maximizing a posteriori probability given an observation, support indices previously chosen, and a priori information on a sparse signal vector. Namely, the proposed method needs to solve bit-wise MAP detection for each iteration, which has much lower complexity than the (vector-wise) MAP detection. This is because the complexity of the former is linear with the sparsity level while the latter is exponential. Unfortunately, the complexity to solve bit-wise MAP detection problem is still expensive since it requires the marginalization of joint probability mass function (PMF) with a large-size random vector. We address this problem by presenting an efficient way to compute a good proxy (i.e., lower-bound) of a posteriori probability. Via simulations, we demonstrate that the proposed method can outperform the other greedy algorithms based on correlations, by exploiting statistical information properly.

II. Preliminaries

In this section we will provide some useful notations and state the sparse signal recovery problem.

A. Notations

We provide some notations which will be used throughout the paper. Let $[N] \triangleq \{1, ..., N\}$. We use $\bar{X}$ and $\mathbf{x}$ to denote a random vector and its values, respectively. Also, for a vector $\mathbf{x} \in \mathbb{R}^N$, x_i denotes the i-th component of $\mathbf{x}$ for $\in [N]$. Similarly for a matrix $\mathbf{B} \in \mathbb{R}^{M \times N}$ the (i,j)-th component of $\mathbf{B}$ is denoted as $\mathbf{B}_{i,j}$. The diagonal approximation of a square matrix $\mathbf{S}$ is denoted by $\mathrm{diag}(\mathbf{S})$, where $\mathrm{diag}(\mathbf{S})$ denotes the diagonal matrix whose i-th diagonal component is $\mathbf{S}_{i,i}$. For any positive $K \leq N$, we let Ω denote the set of all length-N binary vectors with the sparsity level K, i.e.,

$$\Omega \triangleq \{\mathbf{x} \in \{0,1\}^N : \|\mathbf{x}\|_0 = K\}. \tag{2}$$

Given an index subset $\mathcal{I} \subseteq [N]$, we define the subset of Ω as

$$\Omega_{\mathcal{I}} \triangleq \{\mathbf{x} \in \{0,1\}^N : \|\mathbf{x}\|_0 = K, x_i = 1 \text{ for } i \in \mathcal{I}\}, \tag{3}$$

where $|\Omega_{\mathcal{I}}| = \binom{(N-|\mathcal{I}|)}{(K-|\mathcal{I}|)}$. Also, given a vector $\mathbf{x} \in \mathbb{R}^N$, $\mathcal{S}(\mathbf{x})$ represents its support containing the indices of non-zero locations of $\mathbf{x}$ such as

$$\mathcal{S}(\mathbf{x}) \triangleq \{i | x_i \neq 0, i \in [N]\}. \tag{4}$$

As an extension, we also define $\mathcal{S}(\Omega_{\mathcal{I}}) \triangleq \{\mathcal{S}(\mathbf{x}) | \mathbf{x} \in \Omega_{\mathcal{I}}\}$. Given two PMFs $p(x)$ and $q(x)$, the Kullback-Leibler (KL) divergence is denoted as $\mathcal{D}_{\mathrm{KL}}(p||q)$ For two probability distributions $p(x)$ and $q(x)$. Also, for $0 \leq a \leq 1$, $\mathrm{Bern}(a)$ represents a Bernoulli distribution with $\mathbb{P}(X = 1) = a$. Finally, to simplify the expressions, we introduce the notation λ_t given by $\lambda_t \triangleq \frac{K-t}{N-t}$ for $t \leq K$.

B. Problem Formulation

We consider a N-dimensional binary sparse signal recovery problem from a noisy observation. Let $\mathbf{x} \in \{0,1\}^N$ denote a K-sparse binary signal vector (i.e., $\|\mathbf{x}\|_0 = K$). Then, the measurement vector $\mathbf{y} \in \mathbb{R}^M$ is obtained as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}, \tag{5}$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N] \in \mathbb{R}^{M \times N}$ represents a *fixed* measurement matrix and $\mathbf{z} \in \mathbb{R}^M$ follows the zero-mean white Gaussian distribution, namely, $\bar{Z} \sim \mathcal{N}(\mathbf{0}_N, \sigma^2 \mathbf{I})$. Throughout the paper, it is assumed that the sparsity level K is given as a priori information and additionally, the marginal PMFs of the sparse signal vector $\bar{X} = (X_1, ..., X_n)^T$ (denoted by $p_i(a)$) are given as

$$p_i(a) \triangleq \mathbb{P}(X_i = a) \text{ for } i \in [N] \text{ and } a \in \{0,1\}. \tag{6}$$

It is noticeable that in the case of no priori knowledge on the distribution of $\bar{X}$, the marginal PMFs can be assigned as uniform distribution (i.e., $p_i(1) = 0.5$ for $i \in [N]$).

Algorithm 1 Approximate Bit-wise MAP Detector

Input: Measurement matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, noisy observation $\mathbf{y} \in \mathbb{R}^M$, sparsity level K, and noise level σ^2.
Output: Support $\hat{\mathcal{I}}^{(K)} = \{\hat{i}_1, ..., \hat{i}_K\}$.
1: **Initialization** $\hat{\mathcal{I}}^{(0)} = \phi$
2: **for** $k = 1 : K$ **do**
3: Find the k-th support index $\hat{i}_k$ by taking the solution of

$$\hat{i}_k = \underset{i_k \in [N] \setminus \hat{\mathcal{I}}^{(k-1)}}{\mathrm{argmax}} \Lambda\left(i_k | \hat{\mathcal{I}}^{(k-1)}\right),$$

where the objective function is defined in (10).
4: Update the support $\hat{\mathcal{I}}^{(k)} = \hat{\mathcal{I}}^{(k-1)} \cup \{\hat{i}_k\}$.
5: **end for**

From the above model, we will investigate the maximum a posteriori (MAP) support recovery problem, which can be mathematically formulated as

$$\hat{\mathcal{I}} = \underset{\mathcal{I} \in \mathcal{S}(\Omega)}{\mathrm{argmax}} \ \log \mathbb{P}(\mathcal{S}(\bar{X}) = \mathcal{I} \big| \mathbf{y}). \tag{7}$$

Unfortunately, it is too complex to solve the above problem due to its combinatorial nature. Specifically, we need to check the objective function (a posteriori probability) with the $\binom{N}{K}$ plausible candidates, which requires an exponential complexity with the sparsity level K. In the following sections, we will address the above complexity problem by introducing a novel greedy approach.

III. The Proposed Bit-wise MAP Support Detector

In this section, we propose a novel greedy approach to solve the support recovery problem in (7) efficiently. In the proposed method, K support indices (i.e., non-zero components of a sparse signal vector $\mathbf{x}$) are derived in a sequential way via bit-wise MAP detection. Specifically, from the chain rule, the objective function (7) can be factorized as

$$\begin{aligned} &\log \mathbb{P}(\mathcal{S}(\bar{X}) = \mathcal{I}^{(K)} \big| \mathbf{y}) \\ &\quad = \sum_{k=1}^{K} \underbrace{\log \mathbb{P}(i_k \in \mathcal{S}(\bar{X}) | \mathcal{I}^{(k-1)} \subset \mathcal{S}(\bar{X}), \mathbf{y})}_{\triangleq \Phi(i_k | \mathcal{I}^{(k-1)})}, \end{aligned} \tag{8}$$

where the above index sets are defined as $\mathcal{I}^{(K)} = \{i_1, ..., i_K\}$ and $\mathcal{I}^{(k)} = \{i_1, ..., i_k\} \subset \mathcal{I}^{(K)}$ for $k = 1, ..., K-1$, with $\mathcal{I}^{(0)} = \phi$. In the proposed greedy approach, we find a support $\hat{\mathcal{I}}^{(K)} = \{\hat{i}_1, ..., \hat{i}_K\}$ sequentially, by finding a local optimal solution based on the previously chosen solutions. This is mathematically formulated as

$$\hat{i}_k = \underset{i_k \in [N] \setminus \hat{\mathcal{I}}^{(k-1)}}{\mathrm{argmax}} \ \Phi(i_k | \hat{\mathcal{I}}^{(k-1)}). \tag{9}$$

This problem is referred to as bit-wise MAP detection, which has much lower complexity than vector-wise MAP detection in (7) since the complexity of the former is linear with the sparsity level K while the that of the latter is exponential. Although the proposed greedy approach significantly reduces the computation complexity, it still suffers from the expensive complexity for computing a posteriori probability (i.e.,

$\Phi(i_k|\hat{\mathcal{I}}^{(k-1)})$). This is due to the marginalization of a large-scale random vector, which requires the summations of all possible K sparse vector signals $\mathbf{x} \in \Omega_{\hat{\mathcal{I}}^{(k)}}$.

To address the complexity problem, we will derive a good proxy (which is simply computable) of the objective function in (9), which is given as

$$\Lambda\left(i_k|\hat{\mathcal{I}}^{(k-1)}\right) \triangleq \underbrace{\sum_{i=1}^{N} -\mathcal{D}_{\mathrm{KL}}(\mathrm{Bern}(\mu_i^{(k)}), p_i)}_{\text{A priori}} + \underbrace{\frac{1}{\sigma^2}\mathbf{y}^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu}^{(k)} - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{R}^{(k)})}_{\text{Likelihood}}, \quad (10)$$

where

$$\mu_i^{(k)} = \begin{cases} 1, & i \in \hat{\mathcal{I}}^{(k-1)} \cup \{i_k\} \\ \lambda_k, & \text{else} \end{cases} \quad (11)$$

and

$$\mathbf{R}_{i,j}^{(k)} = \begin{cases} 1, & i,j \in \hat{\mathcal{I}}^{(k-1)} \cup \{i_k\} \\ \lambda_k\lambda_{k+1}, & i,j \notin \hat{\mathcal{I}}^{(k-1)} \cup \{i_k\} \\ \lambda_k, & \text{else.} \end{cases} \quad (12)$$

This is in fact a lower bound on the objective function $\Phi(i_k|\hat{\mathcal{I}}^{(k-1)})$ which is obtained by using the concavity of $\log$ function and Jensen's inequality (see Section IV for details). As expected, the objective function in (10) will be further simplified when a priori distribution on a sparse signal vector is unknown, since a priori term is removed. Based on this, the proposed greedy algorithm is described in **Algorithm 1**. In Section V, it will be demonstrated that the proposed proxy function performs very well.

IV. Good Proxy of A Posterior Probability

In this section, we will explain how to derive the proxy function in (10) from the objective function in (9). Note that we will use the notations C_0, C_1, C_2, and C_3 in the below in order to indicate the constant terms which does not impact on the bit-wise MAP optimization in (9). Then, our goal is to efficiently compute the following a posteriori probability for a given index set $\mathcal{I} = \hat{\mathcal{I}}^{(k-1)} \cup \{i_k\} = \{\hat{i}_1, ..., \hat{i}_{k-1}, i_k\}$:

$$\begin{aligned}\Phi(i_k|\hat{\mathcal{I}}^{(k-1)}) - C_0 &= \log \mathbb{P}(\mathcal{I} \subset \mathcal{S}(\bar{X})|\mathbf{y}) \\ &= \log \mathbb{P}(X_{\hat{i}_1} = 1, ..., X_{\hat{i}_{j-1}} = 1, X_{i_j} = 1|\mathbf{y}) \\ &= \log \sum_{\hat{\mathbf{x}} \in \Omega_{\mathcal{I}}} \mathbb{P}(\bar{X} = \hat{\mathbf{x}}|\mathbf{y}) \\ &= \log \sum_{\hat{\mathbf{x}} \in \Omega_{\mathcal{I}}} \frac{p_{\bar{X}}(\hat{\mathbf{x}}) f_{\bar{Y}|\bar{X}}(\mathbf{y}|\hat{\mathbf{x}})}{f_{\bar{Y}}(\mathbf{y})} \\ &= \log\left(\frac{|\Omega_{\mathcal{I}}|}{f_{\bar{Y}}(\mathbf{y})}\right) \frac{1}{|\Omega_{\mathcal{I}}|} \sum_{\hat{\mathbf{x}} \in \Omega_{\mathcal{I}}} p_{\bar{X}}(\hat{\mathbf{x}}) f_{\bar{Y}|\bar{X}}(\mathbf{y}|\hat{\mathbf{x}}), \quad (13)\end{aligned}$$

where $p_{\bar{X}}$ and $f_{\bar{X}|\bar{Y}}$ denote the joint PMF and conditional PDF, respectively, and $\Omega_{\mathcal{I}}$ is defined in Section I. We first provide some definitions which will be used throughout this section.

Definition 1. *We define a length-N auxiliary random vector $\bar{U}$ which takes the values in the set $\Omega_{\mathcal{I}}$ uniformly. Its joint PMF is denoted by $\mathbf{q}(\Omega_{\mathcal{I}})$ where each element in $\Omega_{\mathcal{I}}$ can occur with probability $1/|\Omega_{\mathcal{I}}| = 1/\binom{N-k}{K-k}$ since $|\mathcal{I}| = k$. Then, its marginal PMF can be easily obtained as $U_j \sim \mathrm{Bern}(\lambda_k), j \notin \mathcal{I}$ and $U_j \sim \mathrm{Bern}(1), j \in \mathcal{I}$.*

From Definition 1, (13) can be written as

$$\begin{aligned}&\log \mathbb{P}(\mathcal{I} \subset \mathcal{S}(\bar{X})|\mathbf{y}) - C_1 \\ &= \log \frac{1}{|\Omega_{\mathcal{I}}|} \sum_{\hat{\mathbf{x}} \in \Omega_{\mathcal{I}}} p_{\bar{X}}(\hat{\mathbf{x}}) f_{\bar{Y}|\bar{X}}(\mathbf{y}|\hat{\mathbf{x}}) \\ &= \log \mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[p_{\bar{X}}(\bar{U}) f_{\bar{Y}|\bar{X}}(\mathbf{y}|\bar{U})\right] \\ &\geq \underbrace{\mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log p_{\bar{X}}(\bar{U})\right]}_{\text{A priori}} + \underbrace{\mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log f_{\bar{Y}|\bar{X}}(\mathbf{y}|\bar{U})\right]}_{\text{Likelihood}}, \quad (14)\end{aligned}$$

where the last inequality follows the Jensen's inequality due to the concavity of $\log$ function.

A. The computation of a priori part

In this subsection, we will compute the a priori part in (14). From the a priori probability p_j for $j \in [N]$, we first approximate the joint PMF of $\bar{X}$ as $p_{\bar{X}}(\mathbf{x}) \approx \prod_{i=1}^{N} p_j(x_i)$. Then, we have:

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log p_{\bar{X}}(\bar{U})\right] \\ &= \mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\sum_{i=1}^{N} \log p_j(U_j)\right] \\ &= \sum_{j \in \mathcal{I}} \log p_j(1) + \sum_{j \in [N]\setminus\mathcal{I}} \mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}[\log p_j(U_j)]. \quad (15)\end{aligned}$$

Leveraging the marginal PMFs of U_j's in (1), we have:

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}[\log p_j(U_j)] \\ &= \lambda_k \log p_j(1) + (1-\lambda_k)\log p_j(0) \\ &= -\mathcal{H}_2\left(\mathrm{Bern}(\lambda_k)\right) - \mathcal{D}_{\mathrm{KL}}\left(\mathrm{Bern}(\lambda_k))\middle|\middle|p_j\right), \quad (16)\end{aligned}$$

where $\mathcal{H}_2$ and $\mathcal{D}_{\mathrm{KL}}(\cdot||\cdot)$ denote the binary entropy function and KL divergence, respectively. By plugging (16) into (15), we have:

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log p_{\bar{X}}(\bar{U})\right] - C_2 \\ &= \sum_{j \in \mathcal{I}} \log p_j(1) - \sum_{j \in [N]\setminus\mathcal{I}} \mathcal{D}_{\mathrm{KL}}\left(\mathrm{Bern}(\lambda_k))\middle|\middle|p_j\right). \quad (17)\end{aligned}$$

B. The computation of likelihood part

In this subsection, we will compute the likelihood part in (14). We first introduce a binary random vector $\bar{V}$ for the ease of exposition, which is defined as $\bar{V} = \mathbf{A}\bar{U}$. Using this, we have:

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log f_{\bar{Y}|\bar{X}}(\mathbf{y}|\bar{U})\right] \\ &= \mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log \prod_{j=1}^{M} f_{Y_j|\bar{X}}(y_j|\bar{U})\right] \\ &= \sum_{j=1}^{M} \mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_j - V_j)^2}{2\sigma^2}\right)\right)\right].\end{aligned}$$

Focusing on the interesting terms depending on i_k, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{q}(\Omega_{\mathcal{I}})}\left[\log f_{\bar{Y}|\bar{X}}(\mathbf{y}|\bar{U})\right] - C_3 \\
&\quad = \frac{1}{\sigma^2}\left(\sum_{j=1}^{M} y_j \mathbb{E}[V_j] - \frac{1}{2}\mathbb{E}[V_j^2]\right) \\
&\quad = \frac{1}{\sigma^2}\mathbf{y}^{\mathsf{T}}\mathbf{A}\mathbb{E}[\bar{U}] - \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{A}\mathbb{E}[\bar{U}\bar{U}^{\mathsf{T}}]\mathbf{A}^{\mathsf{T}}\right) \\
&\quad = \frac{1}{\sigma^2}\mathbf{y}^{\mathsf{T}}\mathbf{A}\mathbb{E}[\bar{U}] - \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbb{E}[\bar{U}\bar{U}^{\mathsf{T}}]\right), \qquad (18)
\end{aligned}$$

where

$$\mathbf{A}\mathbb{E}[\bar{U}] = \sum_{j=1}^{n}\mathbf{a}_j\mathbb{E}[U_j] = \sum_{j\in\mathcal{I}}\mathbf{a}_j + \lambda_k\sum_{j\in[N]\setminus\mathcal{I}}\mathbf{a}_j, \qquad (19)$$

and the (i,j)-element of the matrix $\mathbb{E}[\bar{U}\bar{U}^T]$ is computed as

$$\mathbb{E}[\bar{U}\bar{U}^{\mathsf{T}}]_{i,j} = \begin{cases} 1, & i,j \in \mathcal{I} \\ \lambda_k\lambda_{k+1} & i,j \notin \mathcal{I} \\ \lambda_k & \text{else.} \end{cases} \qquad (20)$$

From (17), (18), (19), and (20), and eliminating the constant terms C_1, C_2, and C_3, we can easily derive our objective function in (10) for the bit-wise MAP detection problem.

Remark 1. *In this paper, we only considered a* binary *sparse signal vector for support recovery problem. Yet, we would like to emphasize that the proposed method can be straightforwardly extended to a more general case in which X_i follows a given probability distribution when i belongs to support. In this case, we only need to modify the computations of expectations in (19) and (20) where they should be performed by taking into account the probability distribution of X_i.*

V. Numerical Results

In this section we provide numerical results to show the superiority of the proposed bit-wise MAP detector. We used the reconstruction probability as a performance metric and considered OMP as benchmark method (see Remark 2 for the comparisons with the other greedy algorithms).

No knowledge on a priori distribution: We consider the case that a priori information on a sparse vector signal $\mathbf{x}$ is unknown (i.e., each component of $\mathbf{x}$ can be 1 with equal probability under the constraints of K sparsity). It is remarkable that in this case, a priori term in (10) of the proxy objective function is not used. Fig. 1 shows the reconstruction probabilities of the proposed bit-wise MAP detector and OMP as a function of SNRs. For the simulations, we considered the 50×120 measurement matrix $\mathbf{A}$ (i.e., $M = 50$ and $N = 120$) whose elements are drawn from I.I.D. Gaussian distribution with zero mean and unit variance. The proposed method shows the 4∼5 times better reconstruction performances than OMP in the relative lower SNR regimes (e.g., $0 \sim 10$ dBs). For the range of higher than 20dB, the proposed method achieves the two times higher reconstruction performance than OMP. Not surprisingly, the proposed method performs better in the relatively lower SNR regimes since in this case, the use of statistical information on noise does matter. We next evaluated the reconstruction probabilities of the proposed method and OMP as a function of sparsity levels (see Fig. 2). In this case, we considered a little bit larger measurement matrix (e.g., 80×150 measurement matrix $\mathbf{A}$) to see the performances with a larger sparsity level (e.g., $K = 30$). It was shown that the proposed method can successfully recover the sparse signal vector with 0.85 reconstruction probability even in high sparsity condition (e.g., $K = 14$). Whereas, the reconstruction probability of OMP is lower than 0.4 after $K = 14$. These results demonstrated that the proposed algorithm can identify supports better than OMP even in high SNR regime, nonetheless, statistical information on noise gives smaller effect compared with relatively lower SNR regimes (as shown in Fig. 1).

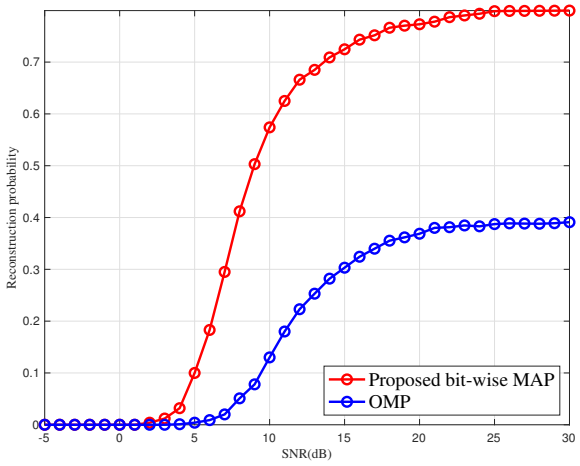


Fig. 1. $K = 10$. Performance comparison of the proposed method in terms of SNRs when a priori distribution on a sparse signal vector is unknown.

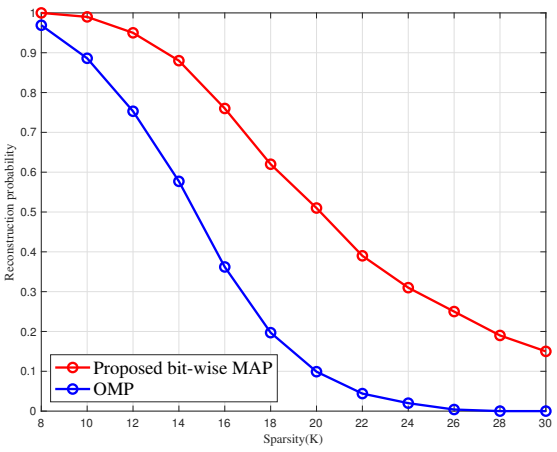


Fig. 2. SNR $= 30$dB. Performance comparison of the proposed method in terms of sparsity levels when a priori distribution on a sparse signal vector is unknown.

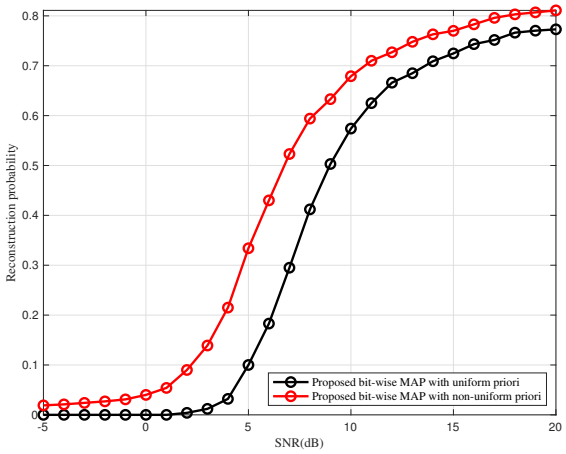


Fig. 3. The impact of a priori distribution on the proposed bit-wise MAP detector.

Non-uniform a priori distribution: Lastly we will verify that the proposed method can indeed exploit a priori information on a sparse signal vector (i.e, KL divergence terms in (10) really works). For the simulation, we set $p_i(1) = 0.7$ for $i \in \mathcal{S}(\mathbf{x})$, and $p_i(1) = 0.5$ for $i \notin \mathcal{S}(\mathbf{x})$. Fig. 3 shows that the reconstruction probability is considerably improved due to the use of a priori information. As expected, in relatively lower SNR regimes, the use of a priori information performs better since in this case, the information from a noisy observation is not sufficient. Namely, the proposed method can identify true support with higher probability, when we know more accurate priori information on a sparse signal vector to make up for the uncertainty of a noisy observation.

From the above results, we can conclude that the proposed algorithm can achieve better reconstruction performances than the conventional greedy algorithm not only in various SNR regimes but in different sparsity levels. In addition, if we have some knowledge on a priori distribution of a sparse signal vector, the proposed method can further improve the performance while the conventional greedy algorithms cannot.

Remark 2. *In our simulations, we only considered OMP as benchmark method. Yet, there exist the advanced greedy algorithms by incorporating the idea of multiple indices selection into the underlying OMP [12]. We would like to highlight that the proposed method can be straightforwardly combined with the advanced algorithms by simply replacing OMP with the proposed method. Thus, for fair comparisons with the advanced algorithms, the proposed method should be also enhanced with the multiple indices selection, which is left for a future work. Given our simulation results, it is expected that the proposed method together with the idea of advanced greedy algorithms would outperform the conventional ones based on OMP.*

VI. Conclusion

In this paper, we proposed a novel greedy algorithm where for each iteration, it finds the best support index by solving bit-wise maximum a posteriori (MAP) detection. Namely, the proposed method exploited the statistical distributions of a sparse signal vector and noise, differently from the existing greedy-based algorithms which rely on the correlation values in magnitude. Our major contribution is to introduce a good proxy function (which is simply evaluated) for the objective function of a bit-wise MAP detection problem (i.e., a selection function in the greedy algorithm), which enables the proposed method practical. Via simulation results, we demonstrated that the proposed method improves the reconstruction probability in all SNR regimes compared with the representative greedy algorithm, named OMP. Moreover, we showed that KL-divergence term, depending on a priori distribution on a spare signal vector, performs quite well. Our ongoing work is to extend the proposed method for the case of general sparse signal vector with a certain probability distribution. Another interesting research direction is to consider sparse support recovery problems with multiple or quantized measurements.

Acknowledgment

This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT1902-00.

References

[1] E.Candès, "Compressive sampling," presented at the Int. CongrMath., Madrid, Spain, Aug 2006
[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory.,* vol. 52, no. 4, pp. 1289-1306, Apr. 2006.
[3] E. J. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, pp. 969, Apr. 2007.
[4] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589-592, Feb. 2008.
[5] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory.,* vol. 53, no. 12, pp. 4655-4666, Dec. 2007.
[6] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4680-4688, July 2011.
[7] T. Zhang, "Sparse recovery with orthogonal matching pursuit under RIP," *IEEE Trans. Inf. Theory*., vol. 57, no. 9, pp. 6215-6221, Sept. 2011.
[8] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265-274, Nov. 2009.
[9] D. Needell and J. A. Tropp, "CoSaMP: iterative signal recovery from incomplete and inaccurate samples," *Commun. ACM*, vol. 53, no. 12, pp. 93-100, Dec. 2010.
[10] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230-2249, May 2009.
[11] J. Wang, S. Kwon, and B. Shim, "Generalized orthogonal matching pursuit," *IEEE Trans. Signal Process*., vol. 60, no. 12, pp. 6202-6216, Dec. 2012
[12] N. Lee, "MAP Support Detection for Greedy Sparse Signal Recovery Algorithms in Compressive Sensing," *IEEE Trans. Signal Processing*, vol 64, no. 19, pp. 4987-4999, Oct 2016.
[13] R. Ward, "Compressed sensing with cross validation," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5773-5782, Dec. 2009.

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

Multilevel Codes in Lattice-Reduction-Aided Decision-Feedback Equalization

Robert F.H. Fischer[1], Sebastian Stern[1], Johannes B. Huber[2]

[1]Institut für Nachrichtentechnik, Universität Ulm, Ulm, Germany, Email: {robert.fischer,sebastian.stern}@uni-ulm.de
[2]Lehrstuhl für digitale Übertragung, Universität Erlangen-Nürnberg, Erlangen, Germany, Email: johannes.huber@fau.de

***Abstract*—The application of multilevel codes in lattice-reduction-aided (LRA) decision-feedback equalization (DFE) is discussed. There, integer linear combinations of the codewords in signal space have to be decoded. Since multilevel codes do not generate lattices in general and non-integer interference of not yet decoded users is present, straightforward decoding is not possible. A generalized version of multistage decoding adapted to LRA DFE is proposed. Thereby, multilevel constructions using state-of-the-art binary channel codes can be used, which makes coded LRA DFE schemes applicable in practice. The performance of the proposed structure is covered via numerical simulations.**

I. Introduction

Lattice-reduction-aided (LRA) schemes [20], [19] and the tightly related *integer-forcing* (IF) schemes [13], [22] are low-complexity but well-performing approaches for the equalization in *multiple-input/multiple-output (MIMO)* multiuser uplink scenarios. They share the concept of decoding *integer linear combinations* of the user's signals; they differ in the way how the integer interference is handled, cf. [4].

In IF schemes a strong coupling between integer equalization and decoding/code constraints is present. In LRA schemes the code has to be linear in signal space, i.e., lattice codes can be used. In [5], and independently in [1], it has been shown that for *LRA linear equalization (LE)* this linearity—integer linear combinations of codewords are valid codewords—can be relaxed and *multilevel codes (MLC)* can be employed together with a generalized version of *multistage decoding (MSD)* incorporating "*carry correction*".

In this paper,[1] we generalize this result to *LRA decision-feedback equalization (DFE)*. Using DFE, the noise prediction gain over linear equalization can be utilized leading to improved performance [2]. However, the successive decoding in DFE and the carry correction procedure in [5] cannot be combined straightforwardly. To solve this problem, we introduce a new version of generalized MSD which employs tentative decisions. Via this approach, which requires only marginal additional complexity compared to independent MSD, multilevel constructions using state-of-the-art binary channel codes can be used in LRA DFE schemes, which simplifies implementation significantly.

The paper is organized as follows: The system model is introduced in Sec. II and LRA DFE is discussed. Sec. III reviews multilevel codes, multistage decoding with carry correction, and introduces the new decoding scheme. Results from numerical simulations are presented in Sec. IV. The paper is briefly summarized in Sec. V.

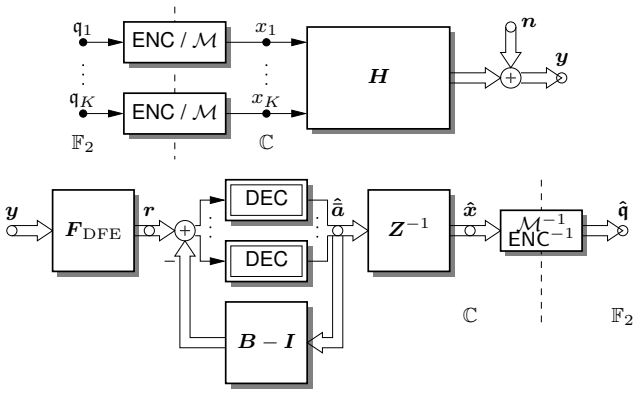


Fig. 1. System model. Top: channel model; Bottom: decision-feedback equalization structure.

II. System Model

In Fig. 1 (top), the considered system model is depicted. We assume K non-cooperating (single-antenna) users k, $k = 1, \ldots, K$, communicating their binary source symbols[2] $\mathfrak{q}_k \in \mathbb{F}_2$ to a central receiver with $N_{\mathrm{R}} \geq K$ receive antennas. At the transmitters, the symbols are encoded and mapped to complex-valued transmit symbols x_k, drawn from the signal constellation $\mathcal{A}$ with variance σ_x^2.

The input/output relation in vector/matrix notation is given as usual by

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}\,, \tag{1}$$

where $\boldsymbol{x}$ denotes the K-dimensional transmit vector, $\boldsymbol{H}$ the $N_{\mathrm{R}} \times K$ channel matrix with flat-fading coefficients, $\boldsymbol{n}$ the N_{R}-dimensional noise vector (we assume zero-mean spatially white Gaussian noise components with variance σ_n^2 per dimension), and $\boldsymbol{y}$ the N_{R}-dimensional receive vector. Joint processing of all components of $\boldsymbol{y}$ is performed at the receiver.

Lattice-reduction-aided and integer-forcing equalization are low-complexity, well-performing approaches. In both strategies, employing a successive equalization strategy improves

[1]A more comprehensive version can be found in [6].

[2]We clearly distinguish quantities over the complex numbers (typeset as $\boldsymbol{x}$, $\boldsymbol{H}$, $\boldsymbol{Z}$, ...), and over finite fields (typeset in Fraktur font; $\mathfrak{q}$, $\mathfrak{c}$, $\mathfrak{Z}$, ...). Vectors over the complex numbers are column vectors, row vectors are signified by underlining (e.g., $\underline{\boldsymbol{r}}$); vectors over the finite field (code words) are always row vectors. Linear combinations over the field of complex numbers are marked by an overbar (e.g., $\bar{\boldsymbol{x}}_k$).

performance over linear equalization. Desiring an equalization according to the *minimum mean-squared error (MMSE)* criterion, for *lattice-reduction-aided decision-feedback equalization* and *successive integer-forcing equalization*, the augmented (stacked) channel matrix $\boldsymbol{\mathcal{H}} = \begin{bmatrix} \boldsymbol{H} \\ \sqrt{\zeta}\boldsymbol{I} \end{bmatrix}$, $\zeta \stackrel{\text{def}}{=} \sigma_n^2/\sigma_x^2$, is factorized according to[3] [14], [16]

$$(\boldsymbol{\mathcal{H}}^+)^{\mathsf{H}} = \boldsymbol{\mathcal{F}}^{\mathsf{H}}\, \boldsymbol{B}^{-\mathsf{H}}\, \boldsymbol{Z}^{-\mathsf{H}}\,, \tag{2}$$

where $\boldsymbol{Z} \in \mathbb{G}^{K\times K}$, $\mathbb{G} = \mathbb{Z}+\mathrm{j}\mathbb{Z}$, is a full-rank Gaussian-integer matrix, $\boldsymbol{B}$ is the upper triangular, unit main diagonal *feedback matrix*, and the left N_{R} columns of $\boldsymbol{\mathcal{F}}$ (with orthogonal rows) give the *feedforward matrix* $\boldsymbol{F}_{\mathrm{DFE}} = \big[\boldsymbol{\mathcal{F}}\big]_{(\text{left } N_{\mathrm{R}} \text{ columns})}$. Thereby, the factorization is performed such that the *column vectors* of $\boldsymbol{\mathcal{F}}^{\mathsf{H}}$ are as short as possible. As shown in [14] (cf. also [16]), one can restrict to *unimodular matrices* $\boldsymbol{Z}$ and the *Hermite–Korkine–Zolotarev (HKZ)* lattice reduction algorithm [8], [11] is optimum.[4]

In the LRA DFE structure (Fig. 1 (bottom)), the feedforward matrix $\boldsymbol{F}_{\mathrm{DFE}}$ guarantees that the noise is (spatially) white and that the cascade $\boldsymbol{F}_{\mathrm{DFE}}\boldsymbol{H}\boldsymbol{Z}^{-1}$ has an (almost) upper triangular form. This establishes a causality of the interference among the parallel data streams [2].

Incorporating $\boldsymbol{F}_{\mathrm{DFE}}$ into the channel, the remaining part of the receiver has to deal with

$$\boldsymbol{r} = \boldsymbol{B}\boldsymbol{Z}\boldsymbol{x} + \bar{\boldsymbol{n}} \stackrel{\text{def}}{=} \boldsymbol{B}\bar{\boldsymbol{x}} + \bar{\boldsymbol{n}}\,, \tag{3}$$

where $\bar{\boldsymbol{n}}$ is the effective disturbance after equalization including filtered channel noise and residual user interference.

In LRA linear equalization [3], [22], [5], not the users' signals are decoded but at the decoder input (noisy versions of) Gaussian integer linear combinations thereof are present. In contrast to LRA LE, in LRA DFE they are not decoded simultaneously in parallel but *successively*—the depicted feedback loop is processed branch by branch; due to the upper triangular form of $\boldsymbol{B}$ the processing order is $l = K, \ldots, 1$.

Noteworthy, (LRA) DFE can also be implemented in the *noise prediction structure*, shown in Fig. 2 (top), which gives the same performance [2]. Here, the feedforward matrix is given by $\boldsymbol{F}_{\mathrm{LE}} = \boldsymbol{B}^{-1}\boldsymbol{F}_{\mathrm{DFE}}$, which is identical to the feedforward matrix in linear equalization. Basically, the successive IF structure (Fig. 2 (bottom)) is similar to the LRA noise prediction structure but here the decoding results and noise samples are treated modulo $\boldsymbol{\Lambda}_{\mathrm{b}}$, the boundary lattice of the used signal constellation $\mathcal{A}$. Moreover, the integer interference is resolved over the finite field (as in the linear IF receiver).

These different orders of encoder inverse and inverse of $\boldsymbol{Z}$ is the main difference between the LRA and IF structures leading to different constraints on the codes. In LRA (linear and DFE) schemes, integer linear combinations in signal space have to be decodable; hence lattice codes are suited. In IF schemes, non-binary codes, tight to the prime signal constellation have to be used [22]. Since the LRA (DFE or noise prediction) structure offers a much more flexible code design and avoids the loss at low SNR due to modulo-folding of the noise, here we restrict ourselves to LRA DFE.

[3] $\boldsymbol{X}^{\mathsf{H}}$, $\boldsymbol{X}^{+}$, $\boldsymbol{X}^{-\mathsf{H}}$: Hermitian, pseudoinverse, inverse and Hermitian of $\boldsymbol{X}$.

[4] Thereby, the size reduction step is irrelevant; hence an *effective HKZ reduction* is sufficient [16].

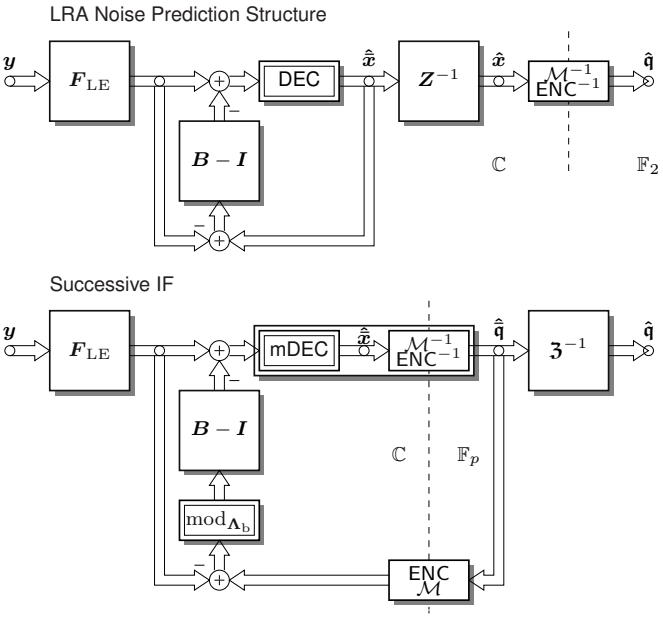


Fig. 2. Lattice-reduction-aided noise prediction structure (top) and successive integer-forcing receiver (bottom). $\mathfrak{z}$ is the finite-field equivalent of $\boldsymbol{Z}$.

III. Multilevel Codes and LRA DFE

Multilevel coding [9], [18] is an attractive strategy to coded modulation, since the code in signal space is generated via a set of conventional *binary component codes* $\mathfrak{C}_\mu$, $\mu = 0, \ldots, m-1$, and a mapping from binary address labels of $m = \log_2(M)$ bits to M signal points.

A. Mapping, Codes, and Lattices

Each constellation can be associated with a *mapping* $\mathcal{M}$ of binary information (labels) to signal points. We restrict ourselves to quadrature-amplitude modulation (QAM) constituent constellations and mapping according to the *set partitioning rule* [17]. Then, the mapping is given by its binary expansion w.r.t. the *base* $\phi = -1 + \mathrm{j}$; for an M-ary constellation it can then be written as [7], [5]

$$\mathcal{M}(\mathfrak{b}_{m-1}\ldots\mathfrak{b}_1\mathfrak{b}_0) = \mathrm{mod}_B\Big(\sum\nolimits_{\mu=0}^{m-1}\psi(\mathfrak{b}_\mu)\phi^\mu\Big) - O\,, \tag{4}$$

where $\psi(\cdot)$ is the common mapping from the finite-field ($\mathbb{F}_2$) elements "$\mathsf{0}$" and "$\mathsf{1}$" to the real numbers "0" and "1" ($\psi(\mathsf{0}) = 0$ and $\psi(\mathsf{1}) = 1$). $B = \phi^m$ defines the boundary region; $\mathrm{mod}_B(x) \stackrel{\text{def}}{=} x - B\,\lfloor x\,B^*/|B|^2\rceil$ is the complex modulo operation ($\lfloor\cdot\rceil$: rounding to the nearest Gaussian integer), and O is the offset for zero-mean constellations.

Via the mapping and having the binary component codes $\mathfrak{C}_\mu$ for the levels $\mu = 0, \ldots, m-1$ (w.l.o.g. for simplification with equal lengths N), the multilevel code is defined by

$$\mathcal{C}_{\mathrm{MLC}} = \mathrm{mod}_B\Big(\sum\nolimits_{\mu=0}^{m-1}\psi(\mathfrak{C}_\mu)\phi^\mu\Big) - O\,, \tag{5}$$

where $\psi(\cdot)$, $\mathrm{mod}(\cdot)$, and the offset O are applied componentwisely.

Eliminating the offset O and ignoring the modulo reduction (inherently assuming an infinite number of extra uncoded levels) multilevel codes can be lattices if the component codes are chosen suitably [12]. However, the respective constraints typically cannot be fulfilled in practical schemes (unless only the lowest level is encoded which results in lattice construction A [21])—in turn integer linear combinations of MLC codewords are not valid codewords of the code and cannot be decoded.

B. Carry Correction

To circumvent this problem and to enable the use of MLC in LRA linear schemes, in [5] (and independently for one-dimensional signaling in [1]) a generalized version of multistage decoding which incorporates a "*carry correction*" has been proposed. Thereby, the main idea is that the parallel decoders (cf. Fig. 1 (bottom)) can exchange decoding results.

Instead of decoding each linear combination (via multistage decoding) separately, the lowest level in each branch is decoded. To this end, we note that the *effective codewords* at level $\mu = 0$, i.e., the results of Gaussian integer linear combinations in signal space, are given by

$$\begin{bmatrix} \mathfrak{c}_{\mathrm{eff},1}^{(\mu)} \\ \vdots \\ \mathfrak{c}_{\mathrm{eff},K}^{(\mu)} \end{bmatrix} = \mathfrak{Z}_0 \begin{bmatrix} \mathfrak{c}_1^{(\mu)} \\ \vdots \\ \mathfrak{c}_K^{(\mu)} \end{bmatrix}, \tag{6}$$

where $\mathfrak{Z}_0 = [\mathfrak{z}_0^{(i,j)}]$ and $\mathfrak{z}_0^{(i,j)}$ is the least significant bit (LSB) of $z_{i,j}$ w.r.t. to the basis ϕ. As long as $\det(\boldsymbol{Z}) \in 1 + \phi\mathbb{G}$ [5], (6) can be solved and the original codewords of each user in the lowest level can be regenerated.

Having these estimates ($\hat{\mathfrak{c}}_k^{(0)}$), the "carries" to the higher levels can be calculated and subtracted. The contributions (over $\mathbb{C}$) $\underline{\boldsymbol{s}}_l^{(0)}$ of the superposition of these code words into the higher levels (carries) of user k calculate to ($\mu = 0$)

$$\begin{bmatrix} \underline{\boldsymbol{s}}_1^{(\mu)} \\ \vdots \\ \underline{\boldsymbol{s}}_K^{(\mu)} \end{bmatrix} = \boldsymbol{Z} \begin{bmatrix} \psi(\hat{\mathfrak{c}}_1^{(\mu)}) \\ \vdots \\ \psi(\hat{\mathfrak{c}}_K^{(\mu)}) \end{bmatrix}. \tag{7}$$

This procedure is repeated over the levels μ.

C. Carry Correction in DFE

Unfortunately, this procedure cannot be applied straightforwardly in LRA DFE. In linear equalization, *integer* linear combinations cause "interference" from the lower levels to the upper ones—causality over the levels is present. Such a causality w.r.t. the code levels does not exist in DFE since the channel is equalized only towards an upper triangular matrix. The not yet decoded upper levels cause interference via the *non-integer* off-diagonal entries of $\boldsymbol{B}$. The fractional part of $b_{l,k}$ determines how the upper levels of user k interfere with a particular level of linear combination $l \leq k$.

Hence, to eliminate interference of other users when decoding level μ of linear combination l, all lower levels $\nu = 0, \ldots, \mu - 1$ of all users have to be known and the upper levels $\nu = \mu, \ldots, m - 1$ of users $K, K - 1, \ldots, l + 1$ whose interference has to be subtracted via $\boldsymbol{B}$ also have to be known.

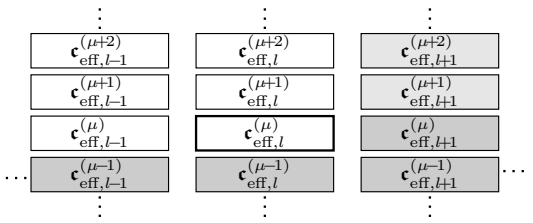


Fig. 3. Visualization of the individual levels of the multilevel construction. Situation when decoding the effective codeword $\mathfrak{c}_{\mathrm{eff},l}^{(\mu)}$. μ: coding level; l: linear combination. Already decoded words are dark gray shaded, tentative decisions are light gray shaded.

This, at first glance, prevents the application of multistage decoding with carry correction as developed for the linear equalization case. However, a small modification is sufficient to use the same philosophy in LRA DFE. The main idea is to employ *tentative decisions* on the higher levels. When having a decoding result $\hat{\mathfrak{c}}_{\mathrm{eff},l}^{(\mu)}$ for level μ of linear combination l (levels $\nu = 0, \ldots, \mu - 1$ are already available from previous decoding stages) *symbol-by-symbol decisions* $\underline{\check{\boldsymbol{x}}}_{\mathrm{up},l}$ on all upper levels jointly of this linear combination are additionally generated via quantization $\mathrm{Q}_{\mathbb{G}}\{\cdot\}$ to the signal point lattice $\mathbb{G}$—thereby, the code constraints in higher levels are simply ignored. As now for all levels (tentative) decoding results are available, a tentative estimate (row vector)

$$\underline{\check{\boldsymbol{x}}}_l = \sum\nolimits_{\nu=0}^{\mu} \underline{\boldsymbol{s}}_l^{(\mu)} \phi^{\nu} + \underline{\check{\boldsymbol{x}}}_{\mathrm{up},l} \phi^{\mu+1} \tag{8}$$

of linear combination l can be calculated. Thereby, the contributions $\underline{\boldsymbol{s}}_l^{(\mu)}$ of the lower levels to the higher ones are calculated as in (7). The tentative estimates of linear combinations $l + 1, \ldots, K$ are used in the feedback loop to eliminate the integer interference ("carry correction") and the non-integer residual interference. This procedure is repeated over the levels.

Fig. 3 visualizes the dependencies in the decoding process. The effective codewords at the individual levels of the multilevel construction are shown. Having decoded all effective codewords $\mathfrak{c}_{\mathrm{eff},l}^{(\nu)}$ at one level ν, the original codewords at this level can be calculated using (6) and the "carries" to the higher levels can be calculated using (7) and subtracted.

When decoding $\mathfrak{c}_{\mathrm{eff},l}^{(\mu)}$ (bold frame), the effective codewords (and hence initial codewords) of levels $0, \ldots, \mu - 1$ are already decoded (dark gray shaded). Due to the successive procedure (going from $l = K$ to 1, i.e., right to left in the figure), the effective codewords $l + 1, \ldots, K$ at levels μ have been decoded, too. In addition, hard (tentative) decisions on the upper levels $> \mu$ are generated (light gray shaded). Using the tentative decisions (8) of data streams $l + 1, \ldots, K$, the interference is subtracted via the feedback matrix $\boldsymbol{B}$. Hence, $\mathfrak{c}_{\mathrm{eff},l}^{(\mu)}$ can be decoded free of carries of lower levels and interference of other users. This is successively done for all

Alg. 1 Multistage Decoding with Carry Correction for DFE.

function $[\underline{\hat{\boldsymbol{c}}}_1, \ldots, \underline{\hat{\boldsymbol{c}}}_K] = \mathtt{MSD}(\underline{\boldsymbol{r}}_1, \ldots, \underline{\boldsymbol{r}}_K)$

```
for μ = 0, 1, ..., m − 1  {                          // loop over levels
    x̃_l = Σ_{ν=0}^{μ−1} s_l^(μ) φ^ν, l = 1, ..., K   // carries of lower levels
    for l = K, K − 1, ..., 1  {                      // loop over linear comb.
        r_{l,μ} = (r_l − Σ_{ℓ=l}^{K} B_{l,ℓ} x̃_ℓ)/φ^μ   // eliminate interference
        ĉ_{eff,l}^(μ) = DEC_{c^(μ)}{r_{l,μ}}         // decode level μ
        x̃_{up,l} = Q_G{(r_{l,μ} − ψ(ĉ_{eff,l}^(μ)))/φ}   // decisions upper levels
        x̃_l = x̃_l + ψ(ĉ_{eff,l}^(μ))φ^μ + x̃_{up,l}φ^{μ+1}   // update tentative dec.
    }
    solve (6) for ĉ_k^(μ)                            // calculate codewords
    calculate s_k^(μ) via (7)                        // calculate carries
}
ĉ_k = Σ_{μ=0}^{m−1} ψ(ĉ_k^(μ))φ^μ                    // codeword estimates
```

data streams at one coding level. The decoding process then continues with the next level.

In QAM signaling the next level operates at a 3 dB higher SNR; the tentative decisions are hence reliable enough. Thus, there are only a few erroneous tentative decisions at higher levels when compared to the lower SNR at the actual level. These errors are controllable by the codes at the actual level without a serious performance degradation. Moreover, for LRA DFE a unimodular integer matrix $\boldsymbol{Z}$ is optimal (cf. Sec. II). For such matrices it is guaranteed that (6) is solvable [5] and carry correction works.

In Alg. 1, a pseudo-code description of this generalized version of multistage decoding is given. Noteworthy, if $\boldsymbol{B} = \boldsymbol{I}$, this algorithms reduces to that in [5] and if additionally $\boldsymbol{Z} = \boldsymbol{I}$ conventional multistage decoding in parallel for the users results.

The complexity is dominated by the runs of the component decoders; each level of each user is decoded exactly once. Consequently, the same effort as K times conventional multistage decoding (parallel, individual decoding of the users) is required.

IV. Numerical Results

To study the performance of the above proposed decoding algorithm, numerical simulations have been conducted. As a simple example, we assume $K = 3$ users, each employing a 16QAM constellation. The same low-density parity-check (LDPC) codes, in particular *irregular repeat-accumulate codes* [10], as in [5] with rates $R_0/R_1/R_2/R_3 = .282/.753/.964/1$ (sum rate 3 bits per QAM symbol) and code length $N = 5000$ are employed as component codes.

To enlighten the effects of decoding integer linear combinations in the LRA DFE structure and to show the gains over LRA linear equalization, first the channel matrix is randomly chosen and kept fixed. The selected channel matrix reads

$$\boldsymbol{H} = \begin{bmatrix} 0.336 + 0.151\mathrm{j} & -0.566 - 0.014\mathrm{j} & -0.255 + 0.454\mathrm{j} \\ -1.101 + 0.581\mathrm{j} & 0.247 - 0.185\mathrm{j} & -0.373 - 0.465\mathrm{j} \\ -1.848 - 1.037\mathrm{j} & 0.019 + 0.758\mathrm{j} & 1.776 - 1.298\mathrm{j} \end{bmatrix}. \quad (9)$$

For LRA linear equalization we employ the Minkowski reduction, as for i.i.d. Gaussian channel matrices the restriction

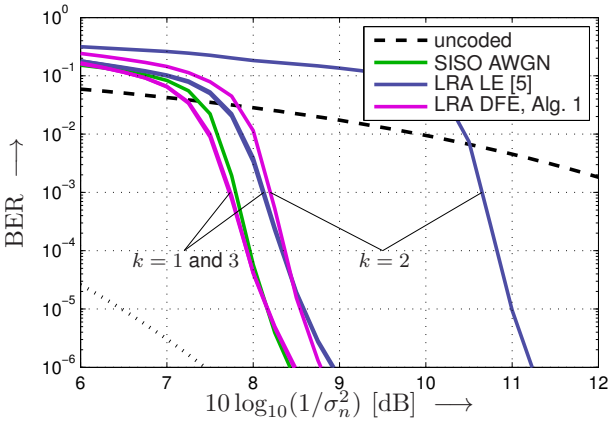


Fig. 4. Bit error rate over the inverse noise power (in dB). 16QAM. Component codes: LDPC codes of length $N = 5000$. Rate 3 bit/symbol. $N_{\mathrm{R}} = 3$ receive antennas, $K = 3$ users. Fixed channel matrix (9); averaging over 100000 codewords. Dotted: asymptotic behavior (curve for uncoded transmission shifted by the gross coding gain of 9 dB).

to unimodular matrices ($|\det(\boldsymbol{Z})| = 1$) causes no noticeable loss, cf. [4], [15]. The following integer matrix is obtained

$$\boldsymbol{Z}_{\mathrm{Mk}} = \begin{bmatrix} \mathrm{j} & 1 & -1-\mathrm{j} \\ -1+\mathrm{j} & 0 & -\mathrm{j} \\ -1 & 0 & -\mathrm{j} \end{bmatrix}. \quad (10)$$

For LRA DFE (see Sec. II) we employ the HKZ reduction on the factorization problem (2); here the integer matrix and the feedback matrix calculate to

$$\boldsymbol{Z}_{\mathrm{HKZ}} = \begin{bmatrix} -2 & 1 & -2\mathrm{j} \\ -1+\mathrm{j} & 0 & -\mathrm{j} \\ -1 & 0 & -\mathrm{j} \end{bmatrix}, \quad (11)$$

$$\boldsymbol{B}_{\mathrm{HKZ}} = \begin{bmatrix} 1 & 0.422 - 0.423\mathrm{j} & -0.609 - 0.395\mathrm{j} \\ 0 & 1 & -0.288 - 0.108\mathrm{j} \\ 0 & 0 & 1 \end{bmatrix}. \quad (12)$$

In all cases, the feedforward equalizers are calculated according to the MMSE criterion.

Fig. 4 shows the error rates of the information bits of the individual users over the inverse noise power (in dB). For comparison, the performance of uncoded transmission is shown (black, dashed) and that of the multilevel code (no linear combinations) over the single-input/single-output (SISO) AWGN channel (green). Noteworthy, due to the uncoded ($R_3 = 1$) highest level, the asymptotic (gross) coding gain is limited to 9 dB (dotted).

In all cases, user 2 has the worst performance, which is due to the noise enhancement in the feedforward filter. In LE this effect is much more pronounced (approximately 2.4 dB worse) than in case of DFE (gain by not equalizing the channel to (almost) identity matrix but only to upper triangular form). Users 1 and 3 perform almost the same (the curves lie almost on top of each other) for a given receiver type but better in case of DFE. This positive effect cannot be explained by reduced noise enhancement in the frontend as linear combinations number 2 and 3 almost have the same noise enhancement in the linear and the DFE case. The better performance is due to the fact that in the successive procedure correlated

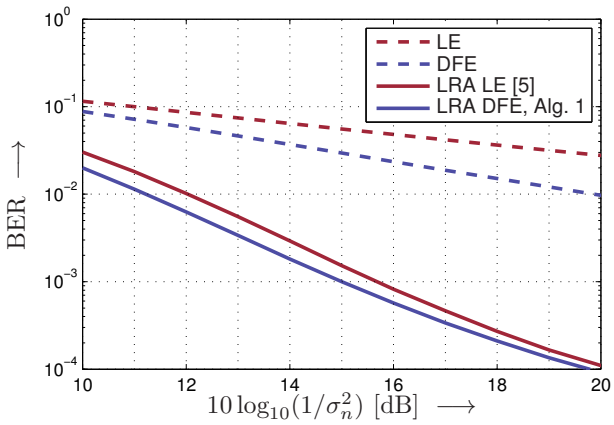


Fig. 5. Bit error rate over the inverse noise power (in dB). 16QAM. Component codes: LDPC codes of length $N = 5000$. Rate 3 bit/symbol. $N_{\mathrm{R}} = 3$ receive antennas, $K = 3$ users. Block-fading channel with i.i.d. circular symmetric complex Gaussian entries. Average over 100000 i.i.d. circular symmetric complex Gaussian channel matrices and 10 codewords per channel realization.

linear combinations have to be decoded [3]; these correlations are exploited in DFE but ignored in linear equalization. The performance of these two users is very close to that of the original code over the SISO AWGN channel.

Next, the channel matrix is randomly chosen with i.i.d. circular symmetric complex unit-variance Gaussian entries. A block-fading channel is assumed, where the channel matrix is constant over the codeword. Hence, the code cannot exploit temporal diversity. Given the channel matrix, the integer matrices are calculated using the Minkowski reduction (which gives the optimal unimodular matrix for LRA LE) and the HKZ reduction (which gives the optimal matrix for LRA DFE), respectively. $N_{\mathrm{R}} = 3$ receive antennas and $K = 3$ users are assumed; the codes and signal constellations from above are assumed.

Fig. 5 shows the average error rates of the information bits of the users over the inverse noise power (in dB). Besides LRA LE (with decoding algorithm from [5]) and LRA DFE (with decoding algorithm Alg. 1), results for conventional LE and DFE (both using the standard MSD decoding algorithm) are treated.

As can be seen, the LRA schemes (solid lines) show a much better performance than the conventional ones (dashed lines); the diversity order is improved from one to $N_{\mathrm{R}} = 3$, which is a well-known fact. Moreover, the DFE schemes (blue) are superior over the linear ones (red), both in the conventional (here the H-BLAST approach is present) and the LRA DFE case. LRA DFE outperforms LRA LE by approximately 1 dB with almost no extra cost in complexity.

V. Summary and Conclusions

In this paper, we have studied the application of multilevel codes in LRA decision-feedback equalization. Employing DFE, the noise prediction gain overlinear equalization can be utilized leading to better performance. A generalized version of multistage decoding incorporating carry correction and tentative decisions has been proposed. Only marginal additional complexity compared to independent decoding is required. Via the multilevel construction, state-of-the-art binary channel codes can be used and no lattice structure of the code is required. This simplifies implementation significantly or even makes coded LRA schemes applicable in practice.

References

[1] S.H. Chae, M. Jang, S.K. Ahn, J. Park, C. Jeong. Multilevel Coding Scheme for Integer-Forcing MIMO Receivers With Binary Codes. *IEEE Trans. Wireless Comm.*, vol. 16, no. 8, pp. 5428–5441, Aug. 2017.

[2] R.F.H. Fischer. *Precoding and Signal Shaping for Digital Transmission*, John Wiley & Sons, New York, 2002.

[3] R.F.H. Fischer, C. Windpassinger, C. Stierstorfer, C. Siegl, A. Schenk, Ü. Abay. Lattice-Reduction-Aided MMSE Equalization and the Successive Estimation of Correlated Data. *AEÜ—Int. Journal of Electronics and Communications*, vol. 65, no. 8, pp. 688–693, Aug. 2011.

[4] R.F.H. Fischer, M. Cyran, S. Stern. Factorization Approaches in Lattice-Reduction-Aided and Integer-Forcing Equalization. In *2016 Int. Zurich Seminar on Communications*, Zurich, Switzerland, March 2016.

[5] R.F.H. Fischer, J.B. Huber, S. Stern, P.M. Guter. Multilevel Codes in Lattice-Reduction-Aided Equalization. In *2018 Int. Zurich Seminar on Communications*, Zurich, Switzerland, March 2018.

[6] R.F.H. Fischer, S. Stern, J.B. Huber. Lattice-Reduction-Aided and Integer-Forcing Equalization: Structures, Criteria, Factorization, and Coding. *Foundations and Trends in Communications and Information Theory*, vol. 16, no. 1–2, pp. 1–155, Dec. 2019.

[7] G.D. Forney. Coset Codes. I. Introduction and Geometrical Classification. *IEEE Trans. Information Theory*, vol. 34, no. 5, pp. 1123–1151, Sep. 1988.

[8] C. Hermite. Extraits de lettres de M.Ch. Hermite à M. Jacobi sur différents objects de la théorie des nombres. *Journal für die reine und angewandte Mathematik*, vol. 40, pp. 261–277, 1850.

[9] H. Imai, S. Hirakawa. A New Multilevel Coding Method Using Error Correcting Codes. *IEEE Trans. Information Theory*, vol. 23, no. 3, pp. 371–377, May 1977.

[10] H. Jin, A. Khandekar, R. McEliece. Irregular Repeat-Accumulate Codes. In *2nd Int. Symp. on Turbo Codes and Rel. Topics*, Brest, France, 2000.

[11] A. Korkine, G. Zolotarev. Sur les formes quadratiques. *Mathematische Annalen*, vol. 6, pp. 366–389, 1873.

[12] W. Kositwattanarerk, F. Oggier. Connections Between Construction D and Related Constructions of Lattices. *Designs, Codes and Cryptography*, vol. 73, no. 2, pp. 441–455, Nov. 2014.

[13] B. Nazer, M. Gastpar. Compute-and-Forward: Harnessing Interference Through Structured Codes. *IEEE Trans. Information Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[14] O. Ordentlich, U. Erez, B. Nazer. Successive Integer-Forcing and its Sum-Rate Optimality. *51st Annual Allerton Conf. on Communication, Control, and Computing*, pp. 282–292, Oct. 2013.

[15] S. Stern, R.F.H. Fischer. Optimal Factorization in Lattice-Reduction-Aided and Integer-Forcing Linear Equalization. In *Int. ITG Conf. on Systems, Communications, and Coding*, Hamburg, Germany, Feb. 2017.

[16] S. Stern, R.F.H. Fischer. V-BLAST in Lattice Reduction and Integer Forcing. In *IEEE Int. Symp. Inf. Theo.*, Aachen, Germany, July 2017.

[17] G. Ungerböck. Channel Coding with Multilevel/Phase Signals. *IEEE Trans. Information Theory*, vol. 28, no. 1, pp. 55–67, Jan. 1982.

[18] U. Wachsmann, R.F.H. Fischer, J.B. Huber. Multilevel Codes: Theoretical Concepts and Practical Design Rules. *IEEE Trans. Information Theory*, vol. 45, no. 5, pp. 1361–1391, Jul. 1999.

[19] C. Windpassinger, R.F.H. Fischer. Low-Complexity Near-Maximum-Likelihood Detection and Precoding for MIMO Systems using Lattice Reduction. In *IEEE Information Theory Workshop*, pp. 345–348, Paris, France, March/April 2003.

[20] H. Yao, G. Wornell. Lattice-Reduction-Aided Detectors for MIMO Communication Systems. In *IEEE Global Telecommunications Conf.*, pp. 424–428, Taipei, Taiwan, Nov. 2002.

[21] R. Zamir. *Lattice Coding for Signals and Networks*. Cambridge University Press, Cambridge, U.K., 2014.

[22] J. Zhan, B. Nazer, U. Erez, M. Gastpar. Integer-Forcing Linear Receivers. *IEEE Trans. Information Theory*, vol. 60, no. 12, pp. 7661–7685, Dec. 2014.

Author Index