Diss. ETH No. 29852

PROBABILISTIC ROBUSTNESS GUARANTEES FOR MACHINE LEARNING SYSTEMS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES

(Dr. sc. ETH Zurich)

presented by

Maurice Giorgio Weber

MSc in Mathematics, ETH Zurich

born on 15.12.1991

Accepted on the recommondation of

Prof. Dr. Martin Vechev

Prof. Dr. Ce Zhang

Prof. Dr. Bo Li

Prof. Dr. J. Zico Kolter

2023

PROBABILISTIC ROBUSTNESS GUARANTEES FOR MACHINE LEARNING SYSTEMS

ABSTRACT

The widespread adoption of Machine Learning (ML) systems has ushered in an era of unparalleled opportunities, transforming industries, healthcare, and society at large. However, alongside these opportunities, ML faces formidable challenges that necessitate comprehensive understanding and need to be adequately addressed. The inherent complexity of ML systems renders them vulnerable to a wide spectrum of adversarial attacks and uncertainties at different stages of an ML pipeline. These vulnerabilities originate from the dynamic and unpredictable environments in which ML systems operate, where subtle perturbations in input data can lead to catastrophic consequences. Furthermore, the emergence of quantum computing and the eventual integration into ML systems introduces additional vulnerabilities along a novel dimension. In consequence, it is of paramount importance to develop methodologies that can guarantee the resilience of ML systems. This thesis is dedicated to the development of probabilistic robustness guarantees which address vulnerabilities arising at different stages of an ML pipeline, encompassing both model development, including training and data curation, and model deployment. Moreover, these guarantees extend to both classical and quantum computing frameworks, acknowledging the distinct challenges and opportunities that quantum computing introduces into the ML landscape. The contributions of this dissertation are threefold:

First, we focus on adversarial attacks that appear during the *model development* stage and develop RAB, which is a provably robust training process against backdoor attacks, a specific instance of a data poisoning attack. Our approach is an extension of the probabilistic robustness guarantees derived from the Neyman-Pearson Lemma and is based on smoothing a model over both training data and test instance. Next to this extension, we also propose several optimizations for specific model types that are needed for our approach to be effective on common, large scale datasets. Next to the theoretical development of the approach, we present extensive experimental results on both tabular data and computer vision datasets and show that our robust training pipeline not only improves the empirical robustness, but also provides a probabilistic certificate that guarantees that a backdoor attack has failed.

Second, we put our attention on the *model deployment* stage and consider input perturbations arising from semantic transformations, as well as shifts in the data distribution. Semantic transformations are typically governed by a low dimensional parameter space (e. g., rotation angle) and incur large image corruptions in terms of ℓ_p norms. This renders traditional probabilistic robustness guarantees, which provide bounds on the perturbation magnitude, ineffective. To address this issue, we propose TSS, a robustness certification framework for input corruptions arising from semantic perturbations. TSS leverages the Neyman-Pearson approach of randomized smoothing and applies smoothing over transformation parameters. In combination with novel techniques to bound interpolation errors and several transformation-specific certification approaches, TSS sets a new state-of-the-art for a large number of semantic transformations.

Orthogonal to this line of work, we study the robustness of ML systems against shifts in the data distribution and, in contrast to the previous results, take a population-based view. Rather than guaranteeing the robustness at a specific instance, here we present guarantees for the out-of-domain generalization by bounding the worst-case population loss over any distribution within an ϵ-Ball around the training distribution. In contrast to previous methods, our approach only requires blackbox access to the model function and is thus scalable to large models such as EfficientNet-B7 and BERT. In diverse experiments on both computer vision and natural language benchmarks, we show that our bounds accurately capture the worst-case change in performance arising from shifts in the data distribution.

Third, we develop robustness guarantees for *quantum* ML where we cover both the adversarial case and input corruptions arising from natural noise and decoherence. To that end, we first present a robustness guarantee for quantum classifiers where we extend the Neyman-Pearson approach to the quantum domain. Interestingly, as a consequence of the inherent probabilistic nature of readout from quantum circuits, this guarantee holds as a general property of quantum classifiers and does not require to actively inject noise for smoothing as is the case for classical ML models. Finally, at this stage in the development of Noisy intermediate-scale Quantum (NISQ) algorithms, natural noise and decoherence is an obstacle of arguably even greater importance than the adversarial scenario. As a last step, we consider approximations of an ideal quantum state where the approximation error arises either from noise, or from algorithmic shortcomings such as limited expressibility of Ansätze. We then extend techniques based on Quantum Hypothesis Testing (QHT) and on the non-negativity of Gram matrices in order to derive bounds on the worst-case error of quantum expectation values. In numerical simulations we study the Variational Quantum Eigensolver (VQE) and validate our bounds on several problems related to quantum chemistry.

ZUSAMMENFASSUNG

Die weitreichende Verbreitung von ML Systemen hat eine Ära beispielloser Möglichkeiten eingeläutet, die das Potenzial haben verschiedenste Branchen, sowie die Gesellschaft insgesamt zu transformieren. Gleichzeitig steht ML jedoch erheblichen Herausforderungen gegenüber, die ein umfassendes Verständnis und angemessene Massnahmen erfordern. Die inhärente Komplexität von ML-Systemen macht sie anfällig für eine breite Palette von Angriffen und Unsicherheiten in verschiedenen Phasen einer ML-Pipeline. Diese Schwachstellen resultieren aus den dynamischen und unvorhersehbaren Umgebungen, in denen ML-Systeme operieren, in denen subtile Störungen in den Eingabedaten zu schwerwiegenden Folgen führen können. Darüber hinaus führt die Entstehung der Quantencomputing-Technologie und deren potenzielle Integration in ML-Systeme zusätzliche Schwachstellen in einer neuen Dimension ein. Daher ist es von grosser Bedeutung, Methoden zu entwickeln, die die Widerstandsfähigkeit von ML-Systemen gewährleisten können. Diese Dissertation widmet sich der Entwicklung von probabilistischen Robustheitsgarantien, die Schwachstellen in verschiedenen Phasen einer ML-Pipeline behandeln, einschliesslich Modellentwicklung, Training und Datenkuratierung sowie Modellbereitstellung. Darüber hinaus erstrecken sich diese Garantien auf klassische und quantenbasierte Rechenmodelle und berücksichtigen die unterschiedlichen Herausforderungen und Möglichkeiten, die Quantencomputing in die ML-Landschaft einführt. Die Beiträge dieser Dissertation sind dreifach:

Erstens konzentriert sich die vorliegende Arbeit auf Schwachstellen, die während der Modellentwicklung auftreten, und entwickeln RAB, einen Trainings-Algorithmus, der Robustheitsgarantien gegen Backdoor Angriffe bietet. Solche Backdoor Angriffe sind eine spezifische Form eines Data Poisoning Angriffs. Der präsentierte Ansatz erweitert die probabilistischen Robustheitsgarantien, die aus dem Neyman-Pearson-Lemma abgeleitet sind, und basiert auf der Glättung eines Modells über Trainingsdaten und Testinstanzen. Neben dieser Erweiterung schlagen wir auch mehrere Optimierungen für spezifische Modelltypen vor, die erforderlich sind, damit unser Ansatz auf gebräuchlichen, gross angelegten Datensätzen wirksam ist. Neben der theoretischen Entwicklung des Ansatzes präsentieren wir umfangreiche experimentelle Ergebnisse sowohl für tabellarische Daten als auch für Computer Vision-Datensätze und zeigen, dass unsere robuste Trainings-Pipeline nicht nur die empirische Robustheit verbessert, sondern auch eine probabilistische Garantie bietet, dass ein Backdoor Angriff gescheitert ist.

Zweitens konzentrieren wir uns auf die Phase der Modellbereitstellung und betrachten Eingabestörungen, die aus semantischen Transformationen und Verschiebungen in der Datenverteilung resultieren. Semantische Transformationen werden in der Regel von einem niedrigdimensionalen Parameterbereich (z.B. Rotationswinkel) gesteuert und führen zu grossen Veränderungen, gemessen in ℓ_p-Normen. Dies macht herkömmliche probabilistische Robustheitsgarantien, die die maximal tolerierbare Störungsmagnitude bestimmen, unwirksam. Um dieses Problem anzugehen, schlagen wir TSS vor, ein Framework, welches Garantien bezüglich der Robustheit gegenüber semantischen Transformationen liefert. TSS nutzt den Neyman-Pearson-Ansatz des Randomi-

zed Smoothings und wendet Glättung über Transformationsparameter an. In Kombination mit neuartigen Techniken zur Begrenzung von Interpolationsfehlern und verschiedenen transformationsbasierten Zertifizierungsansätzen, setzt TSS einen neuen state-of-the-art für eine grosse Anzahl semantischer Transformationen.

Orthogonal zu diesem Thema untersuchen wir die Robustheit von ML-Systemen gegenüber Verschiebungen in der Wahrscheinlichkeitsverteilung von Daten, und nehmen, im Gegensatz zu den vorherigen Ergebnissen, eine Populationsbasierte Sichtweise ein. Statt die Robustheit für eine spezifische Instanz zu garantieren, präsentieren wir hier Garantien für die Generalisierung ausserhalb der Verteilung der Trainingsdaten, indem wir das maximale Risiko über jede Verteilung innerhalb eines ϵ-Bereichs um die Trainingsverteilung begrenzen. Im Gegensatz zu früheren Methoden erfordert unser Ansatz lediglich Blackbox-Zugriff auf die Modellfunktion und ist daher skalierbar auf grosse Modelle wie EfficientNet-B7 und BERT. In vielfältigen Experimenten, sowohl im Bereich Bilderkennung als auch in der Sprachverarbeitung, zeigen wir, dass unsere Garantien die worst-case Veränderung der Leistung bei Verschiebungen in der Datenverteilung präzise erfassen.

Drittens entwickeln wir Robustheitsgarantien für das Quanten-Maschinenlernen, bei denen wir sowohl das Angriffs-basierte Szenario, als auch Eingabestörungen durch natürliches Rauschen und Dekohärenz, abdecken. Zu diesem Zweck präsentieren wir zunächst eine Robustheitsgarantie für Quantenklassifikatoren, bei der wir den Neyman-Pearson-Ansatz auf den Quantenbereich ausdehnen. Aufgrund der inhärenten probabilistischen Natur des Auslesens von Quantenschaltkreisen, gilt diese Garantie als eine allgemeine Eigenschaft von Quantenklassifikationsalgorithmen und erfordert keine aktive Injektion von Rauschen zur Glättung, wie es bei klassischen ML-Modellen der Fall ist. Auf dem gegenwärtigen Stand der Entwicklung von NISQ-Algorithmen, ist das natürliche Rauschen und die Dekohärenz ein Hindernis von sogar noch grösserer Bedeutung als das Angriffs-basierte Szenario. Als letzten Schritt betrachten wir Approximationen eines idealen Quantenzustands, wobei der Approximationsfehler aus Rauschen oder algorithmischen Mängeln wie der begrenzten Ausdrucksfähigkeit von Ansätzen resultiert. Wir erweitern Techniken, die auf dem Quantenhypothesentesten und der Nicht-Negativität von Gram-Matrizen basieren, um Grenzen für den worst-case Fehler von quantenmechanischen Erwartungswerten abzuleiten. In numerischen Simulationen untersuchen wir VQE, ein variationsbasierter Algorithmus zur Bestimmung von Eigenwerten, und validieren unsere Robustheitsgarantien in Bezug auf verschiedene Probleme im Bereich der Quantenchemie.

ACKNOWLEDGMENTS

First and foremost, I would like to sincerely thank Ce for all the coaching, support and inspiration he has given me over the last years. Ce has provided me with the perfect mixture of guidance and freedom I needed to develop my skills as a researcher and to pursue my academic interests. In addition, I would also like to thank Ce for the always immensely helpful career advice he has given me and for supporting me in whatever goals I envisioned. I could not have asked for a better advisor.

I would like to thank Bo, who I have closely collaborated with on many projects, and whose guidance and scientific visions I have immensely valued. These were truly enjoyable collaborations for which I am very grateful. I would also like to thank Martin for taking over the role as a PhD advisor in the final months of my PhD and for being part of the examination committee. Finally, thank you also to Zico, and the examination committee, for taking the time to assess this thesis and for your valuable feedback.

I also want to thank my colleagues and collaborators from ETH, especially Jansen for providing me with his wisdom and guidance on everything related to quantum computing. Our quantum coffee sessions at MAME were always very enjoyable. I am also grateful for the other members of the DS3Lab who I had the privilege to work with. Thanks to Bojan, Cedric, Johannes, Luka, Merve, Susie, Xiaozhe, Xiaozhong, Yilmazcan, and Yongjun. During my PhD, I also had the opportunity to work with very talented students, thanks Anton, Felix, Himankar, and Valdemar for the work we have done together.

I was also fortunate to have had the opportunity to spend time and work with people outside of ETH. I would like to thank Linyi and Xiaojun from UIUC for the very enjoyable collaborations and their immensely valuable contributions to each project. I also thank Abhinav, Alán, Alba, Jakob, and Thi Ha from the university of Toronto for the nice collaboration we had. I also want to thank the people at Xanadu for hosting me in Toronto during a joyful summer. Finally, I would like to thank the amazing team at Together AI for the exciting collaborations on RedPajama.

I am also incredibly grateful for my dear friends and family, who have always supported me and helped me distracting my mind away from research whenever I needed it. Finally, I would like to thank Shaila, my Love, who has supported me during every moment of this journey. I am very lucky to have you by my side.

Zurich, December 2023

The results outlined in this dissertation are based on several works published as first or co-first author. Some of the descriptions, figures, and tables are taken directly from these papers. Although, this dissertation bears my name, much of this work is based on collaborations with different authors, most notably: Bo Li, Bojan Karlas, Ce Zhang, Linyi Li, Luka Rimanic, Nana Liu, Xiaojun Xu, and Zhikuan Zhao. This collection of research would not have been possible without the contributions of all these collaborators.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AC	Activation clustering
ART	Adversarial Robustness Toolbox
CNN	Convolutional Neural Network
CPTP	Completely positive and trace preserving
DNN	Deep Neural Network
GAN	Generative Adversarial Network
KNN	k-Nearest Neighbors
ML	Machine Learning
NC	Neural Cleanse
NISQ	Noisy intermediate-scale Quantum

POVM	Positive Operator-valued Measure
QML	Quantum Machine Learning
QHT	Quantum Hypothesis Testing
SCAn	Statistical Contamination Analyzer
SDP	Semidefinite Programming
SPA	Separable-Pair Approximation
SVM	Support Vector Machine
TSS	Transformation-Specific Smoothing
TSS-R	Transformation-Specific Smoothing for Resolvable Transformations
TSS-DR	Transformation-Specific Smoothing for Differentially Resolvable Transformations
UpCCGSD	Unitary Pair Coupled-Cluster Generalized Singles and Doubles
VarQTE	Variational Quantum Time Evolution
VQE	Variational Quantum Eigensolver

1 INTRODUCTION

1.1 MOTIVATION & RESEARCH SCOPE

Over the last decade, the field of ML has revolutionized our ability to process and interpret vast amounts of data, enabling applications ranging from image recognition and autonomous driving to the recent success of natural language models and conversational AI assistants. However the wide-spread adoption of these technologies, especially in security-critical scenarios, makes it paramount to understand how these systems operate and to have a sound understanding of their limitations, risks and vulnerabilities. In particular, ML systems are vulnerable to a vast array of threats originating from the environments in which they operate, including malicious attacks by adversaries, shifts in data distributions, and even naturally occurring noise, to name just a few. It is thus crucial to develop methods that enable guarantees on the correctness, fairness and safety of these ML systems as they are becoming ever more embedded in our everyday lives.

However, ML systems are inherently complex and consist of a multitude of individual stages, ranging from data collection and curation, to model design and development, before being deployed and monitored in real-life applications. This introduces further vulnerabilities and makes reasoning about guarantees on the robustness of these systems an exceedingly challenging task. Indeed, during development, ML systems are susceptible to data poisoning attacks, where adversaries are able to interfere with the data collection and model training process in order to bias the model towards certain patterns and exploit such biases once the model has been deployed. During deployment, vulnerabilities present themselves in different forms and shapes such as manually crafted adversarial examples or even innocent semantic transformations and shifts in the input distribution due to e. g., changing demographics, climate or of the broader environment. Moreover, the emergence of quantum computing introduces a further dimension to these challenges. While Quantum Machine Learning (QML) leverages the unique properties of quantum mechanics and promises to amplify the capabilities of ML systems, it also introduces novel vulnerabilities presenting themselves in the form of adversarial examples and, an obstacle of arguably even greater significance, natural noise and decoherence inherent to quantum systems. These vulnerabilities can differ profoundly from classical ML, making it imperative to address both frameworks comprehensively. The core question that this dissertation is attempting to answer is therefore:

How can we reason about robustness guarantees for different stages of an ML pipeline, considering the unique nature of vulnerabilities inherent to those stages and accounting for the presence of both classical and quantum computing frameworks?

In the remainder of this chapter, we break this question up into five research questions which we aim to answer in the main part of this dissertation. Keeping in mind the natural flow of ML pipelines, we start by highlighting vulnerabilities present during model development, and the kind of guarantees that we wish to obtain during this

stage. Subsequently we put our attention on the model deployment stage, starting with classical ML, before moving on to the quantum domain. Finally, we provide an overview of the organization and the reading threads, and present the publications that have led to the results outlined in this work.

1.1.1 *Robustness Guarantees during Model Development*

A core component in the initial cycles of any ML development pipeline concerns the collection and curation of training data used to train ML models designed for a specific use case. Often, dataset creators harvest their dataset from public sources on the web [172, 247], or by allowing outsiders the privilege to contribute data samples actively, thereby opening the door for potential adversaries to interfere with the ML model development process. Most notably, the feasibility of such attacks has been illustrated by the manipulation of commercial spam filters [161] or the Tay chatbot [231], and dataset poisoning attacks have been shown to be a realistic threat to large language models and web scale datasets [24, 232] among others. In response to these threats, the research community has proposed several empirical methods to defend ML models against such attacks [67, 76, 134, 234]. However, while these methods are empirically expected to improve the robustness, they are not designed to provide a *guarantee* and can, in principle, be bypassed by adaptive adversaries. In addition, while guarantees to defend against test-time attacks have been relatively well studied, it is challenging to directly extend these results to the data poisoning scenario, given that these threat models differ significantly. Indeed, only a relatively small number of methods have been proposed that are able to provide robustness guarantees for various forms of data poisoning attacks [107, 131, 184]. Among the various forms of data poisoning attacks, in a *backdoor attack*, an adversary poisons a dataset by adding a particular pattern to a subset of training instances such that the resulting model is biased towards these patterns, thereby planting a "backdoor" in the model. During deployment, the adversary can then exploit this vulnerability by knowing the backdoor pattern and controlling model predictions to, e. g., bypass malware filters or spam detection systems. It becomes apparent that, while ML models are vulnerable against attacks on the development stage of an ML pipeline, a sound and comprehensive understanding of corresponding robustness guarantees is lacking. We thus formulate our first research question in an attempt to fill this gap.

Question 1: How can we develop certifiably robust ML models against backdoor attacks?

Randomised smoothing [39, 128] has been proposed as an approach to provide probabilistic robustness guarantees against test-time adversarial attacks, scaling to ImageNet [185] size neural networks. In this thesis, we explore the extension of randomised smoothing to certify the robustness against backdoor attacks. Interestingly, other approaches have also used randomised smoothing to derive guarantees against label flipping attacks [184], against ℓ_0-norm feature- and label flipping attacks [276] and to bound the maximum number of poisoned instances an ML model can support without being successfully backdoored [131]. While the extension of randomised smoothing to the data poisoning scenario is a first step, we also explore how these guarantees can and, in fact, need to be optimized for specific model types such as neural networks or K-nearest neighbour classifiers.

1.1.2 *Robustness Guarantees during Model Deployment*

After an ML model has been fully developed and trained, it is being deployed in real-life applications where users make decisions based on its predictions, or even fully allow it to take control of actions, such as in autonomous driving scenarios. However, also during this stage of the ML pipeline, models remain vulnerable to their environment, be it in a malicious context, via the careful crafting of ℓ_p-norm bounded adversarial examples [7, 29, 71, 212, 221, 254], semantic transformations [21, 59, 69, 70, 91, 98, 170, 256] or via naturally occurring shifts in the data distributions [3, 9, 44, 78, 113, 229]. It is thus clear that ML models are subject to a plethora of vulnerabilities during deployment, which require careful analysis and understanding in order to safely and reliably use those systems in real-life applications. In response to these risks, a multitude of empirical defenses have been developed with the goal of improving the robustness against adversarial examples [146, 189, 197] and semantic transformations [59, 91], or by leveraging distributionally robust optimization techniques [10, 16, 52, 66, 121, 190] to improve the robustness against distribution shifts. As is the case for empirical defenses against vulnerabilities emerging during model development, while these test-time defense methods can increase the ML model robustness, they fail to provide guarantees and can either be evaded by adaptive adversaries, or by shifts in data distributions that are unknown at training time. This state of affairs has sparked the development of certifiable defense methods which, in addition to improving the empirical robustness, also provide a robustness guarantee. However, in the context of instance-level certification, the majority of existing techniques is designed to certify the robustness against ℓ_p-norm bounded perturbations [39, 74, 131, 177, 252], which poses a significant obstacle when transformations are governed by a low dimensional parameter space which incurs a large ℓ_p distance between clean and attacked test sample, as is the case for semantic transformations like rotations, Gaussian blur, or scaling. While certified defenses in this direction have been developed [65, 156, 202], they are either loose and computationally expensive [8, 156] or provide certification only against a relatively weaker set of attacks [65]. This leaves a gap open which we wish to explore in this thesis. The first question related to robustness during model deployment is therefore:

Question 2.1: How can we guarantee the robustness of ML models against input corruptions arising from semantic transformations that incur large ℓ_p-norm perturbations?

To provide an answer to this question, in this thesis, we explore the extension of the probabilistic guarantees provided by randomised smoothing to semantic transformations. Due to the large perturbation magnitude that such transformations incur, a direct application of the ℓ_p-norm guarantees is inadequate and requires further adjustment accounting for the unique properties of these transformations.

As highlighted previously, next to the instance-level perturbations, ML models are also vulnerable to shifts occurring in the data distributions [113]. In this case, rather than certifying a specific test instance, we are interested in bounds to population-level metrics, such as accuracy or test loss, which enable the certification of the out-of-domain generalization abilities of ML models. Existing methods that aim to provide such certification typically do not scale to medium and large scale models [16, 42, 66, 198, 203, 205] as they rely on properties of the models such as the Lipschitz constant and the model smoothness which, in practice, are difficult to obtain accurately. This lack of guarantees therefore motivates the next research question which we aim to answer in this thesis:

Question 2.2 How can we certify the out-of-domain generalization abilities of ML models while only allowing blackbox access to these models?

The crucial point is that we require only blackbox access to the model, allowing certification of large scale models. To that end, in this thesis, we explore the application of a technique pioneered by Weinhold [243], which provides lower bounds to quantum expectation values in the context of quantum chemistry.

1.1.3 *Robustness Guarantees for Quantum Machine Learning*

QML is an emerging subfield of quantum computing that aims to exploit the unique nature of quantum mechanics with the goal of enhancing capabilities of classical ML models in the form of drastic speed-ups of classical algorithms [41, 63, 179, 278], improved accuracy [1, 81, 140, 191] or robustness [50, 77, 125, 138, 141, 248]. With the emergence of noisy, intermediate-scale quantum computers [174], the field of QML has attracted increased interest in recent years [14, 55, 192]. With the eventual integration of quantum computing components in ML systems, natural questions about their reliability and safety arise, akin to the vulnerabilities present in current, purely classical ML systems. Indeed, like their classical counterparts, quantum classifiers have been shown just as vulnerable to adversarial examples [138, 141]. However, in the current NISQ era of quantum computing [174], an arguably even more pressing concern is the level of naturally occurring noise, inherent to current quantum computing systems, that these QML algorithms can tolerate. In light of these challenges, and orthogonal to our work on robustness guarantees for classical ML pipelines, in this thesis we also seek to explore robustness guarantees for QML algorithms. In a first step, we consider quantum classification models which, in analogy to their classical counterparts, are trained to classify data encoded as quantum states into a set of classes via optimization of parametrized quantum circuits. While first steps in this direction have been made [138], a complete and tight characterization of the robustness of QML models is lacking. Therefore, in this thesis, we seek to answer the following question:

Question 3.1: How can we enable tight robustness guarantees for quantum classification models, taking into account the unique nature of QML algorithms?

Typically, the prediction of a quantum classifier is formed by obtaining repeated measurements of a quantum observable which are then aggregated into an expectation value. In this thesis, we explore the probabilistic nature of such classifiers and seek to get an understanding of how classical probabilistic robustness guarantees, in particular the Neyman-Pearson approach [39], can be used to derive robustness guarantees inherent to quantum classifiers. Such a characterization would prove useful along at least two axes. First, the worst-case nature of the guarantee, naturally covers quantum adversarial examples as these can be considered a worst-case type of noise. In addition, in the context of natural quantum noise, the exact characterization of noise models is difficult in practice. The lack of such a model thus further motivates the need to derive worst-case robustness guarantees for quantum classifiers.

While the preceding research question, in its core, seeks to explore conditions under which the most likely measurement remains unaffected, we wish to take a further step in this thesis and explore how shifts in quantum states affect more general quantum algorithms whose outputs are based on expectation values of quantum observables.

Here we move away from the adversarial scenario and study perturbations that occur due to noise and decoherence, prevalent in NISQ era quantum devices, limited expressibility of Ansätze [160, 201], barren plateaus during optimization in variational hybrid quantum-classical algorithms [150, 167, 236], measurement noise and decoherence, and other experimental imperfections that occur frequently in practice [102, 241]. It is worth pointing out that this topic is related, but orthogonal, to quantum error mitigation [22] and quantum error correction [23, 199, 206] which aim to *reduce* or *eliminate*, rather than *characterize*, errors in quantum algorithms. While it is clear that such errors exist and can indeed be detrimental to quantum algorithms, an accurate characterization of errors from such a multitude of sources has received relatively less attention. Inspired by the early days of classical computing where errors stemming from faulty floating-point operations, temperature and voltage fluctuations and other imprecisions were prevalent and their magnitude estimated [230, 251], we explore the following, and final question:

Question 3.2: How can we characterize accurate error bounds on the output of quantum algorithms arising from imperfect representations of an ideal quantum state?

In this thesis, we take a worst-case point of view, in the sense that we do not assume any prior knowledge about how these imperfections may arise and only require bounds on how large the approximation error between the approximate and ideal state can possibly be. In other words, we explore bounds which only require blackbox access to the quantum systems of interest and are independent of whether these imperfect representations come from noise and decoherence or from algorithmic errors.

1.2 CONTRIBUTIONS

To address the research questions proposed in the previous section, this thesis makes the following set of contributions.

C1 To defend against backdoor attacks, we propose an extension of randomised smoothing [39, 128] which, together with our robust training pipeline, RAB, improves robustness empirically while providing a rigorous robustness guarantee. Next to contributions on the theoretical level, this also includes several practical optimizations for specific model classes.

C2.1 To provide robustness guarantees against adversarial attacks in the form of semantic transformations, we present TSS, a framework based on randomised smoothing, that enables the certification of transformation parameters and overcomes the limitations of classical ℓ_p-norm guarantees for certifying robustness against semantic adversarial attacks.

C2.2 To certify the out-of domain generalization of ML models, we derive bounds on the worst-case population loss over an uncertainty set of shifted distributions, expressed in terms of Hellinger distance balls around the source distribution. Since our bounds only require blackbox access to the model, we show that the certification scales to large-scale datasets and models.

C3.1 Exploiting the unique probabilistic nature of quantum classifiers, we make use of the Neyman-Pearson approach to QHT and derive tight robustness bounds for

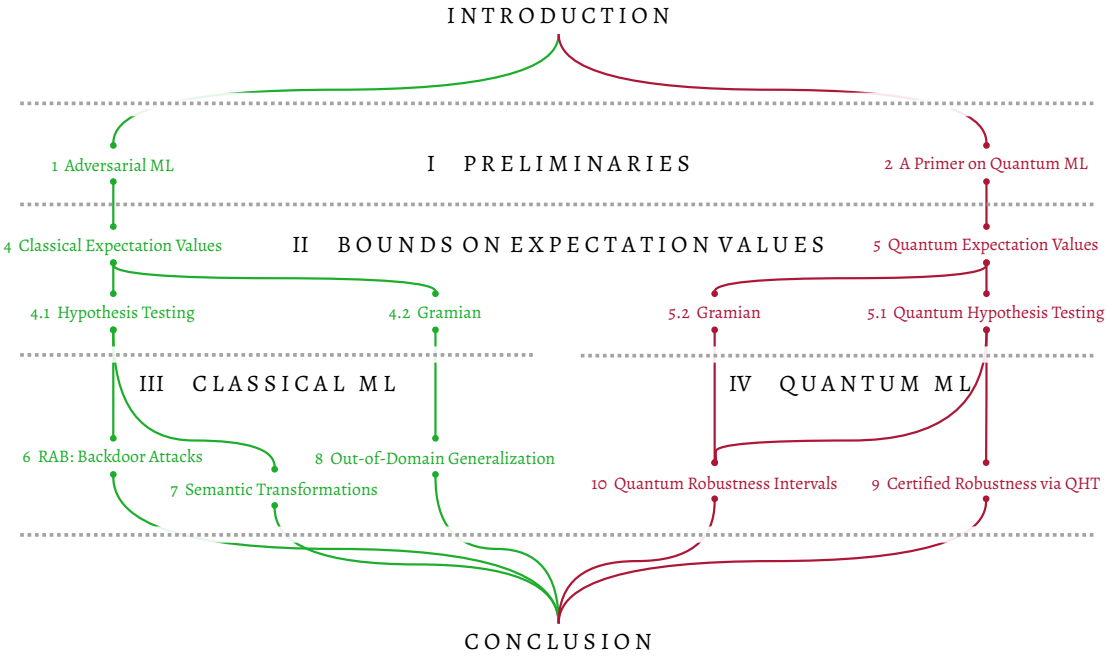


Figure 1: Structure of the thesis and reading threads. The reading thread covering classical ML is shown in green color, and the reading thread covering quantum ML is shown in red color.

quantum classifiers. Next to the guarantees, this bound reveals intriguing links between the characterization of robustness for quantum classifiers and hypothesis testing. Namely, while hypothesis testing is concerned with *discriminating* between quantum states, robustness is concerned with finding conditions under which classifiers *can not discriminate* quantum states. We make this intuition precise in terms of error probabilities associated with QHT.

C3.2 Finally, we present tight bounds on quantum expectation values that allow for an accurate characterization of the worst-case error of quantum algorithms stemming from imperfect representations of an ideal quantum state. These bounds only require knowledge of a bound on the similarity of approximate and ideal state and are otherwise agnostic to the nature of the approximation error. We validate our bounds in the context of the seminal VQE.

1.3 ORGANIZATION OF THE THESIS

We have organized this thesis into four main parts, and have made the separation largely based on whether a result is concerned with classical or quantum machine learning. In this way, a reader more interested in the classical part of this thesis, can follow the classical reading thread, while a reader more interested in the quantum part, can follow the quantum reading thread. The conceptual organization of the thesis is illustrated in Figure 1. In Part i, we introduce the background required for a sound understanding of the results introduced in this thesis. In Part ii, we present an overview of the theoretical results we have derived in order to answer the research questions outlined at the beginning of this thesis. We remark that these results are stated in the

most general way possible so as to illustrate the general applicability to multiple topics treated in this thesis. In addition, this generality also makes apparent the similarities and differences between results obtained for classical and quantum machine learning. In Part iii we present our work related to classical machine learning, covering robustness guarantees for backdoor attacks (Chapter 6), semantic transformations (Chapter 7), and out-of-domain generalization (Chapter 8). In Part iv, we present our results related to the quantum side of this thesis. In Chapter 9, we derive robustness bounds for quantum classifiers, and in Chapter 10 we present error bounds for quantum expectation values arising from imperfect representations of ideal quantum states. We conclude in Chapter 11 where we summarize the contributions and limitations of our work, and outline the potentials for future work.

1.4 AUTHOR'S PUBLICATIONS

This dissertation is based on the following publications, presented in the order of appearance in this dissertation († indicates equal contribution):

[1] **Maurice Weber**†, Xiaojun Xu†, Bojan Karlaš, Ce Zhang, and Bo Li. "Rab: Provable Robustness against Backdoor Attacks." In *2023 IEEE Symposium on Security and Privacy (SP)*. 2023.

[2] Linyi Li†, **Maurice Weber**†, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. "TSS: Transformation-Specific Smoothing for Robustness Certification." In: *2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*. 2021.

[3] **Maurice Weber**, Linyi Li, Boxin Wang, Zhikuan Zhao, Bo Li, and Ce Zhang. "Certifying Out-of-Domain Generalization for Blackbox Functions." In: *39th International Conference on Machine Learning (ICML)*. 2022.

[4] **Maurice Weber**, Nana Liu, Bo Li, Ce Zhang, and Zhikuan Zhao. "Optimal provable robustness of quantum classification via quantum hypothesis testing." In: *npj Quantum Information 7.1* (2021), p. 76.

[5] **Maurice Weber**, Abhinav Anand, Alba Cervera-Lierta, Jakob S Kottmann, Thi Ha Kyaw, Bo Li, Alán Aspuru-Guzik, Ce Zhang, and Zhikuan Zhao. "Toward reliability in the nisq era: Robust interval guarantee for quantum measurements on approximate states." In: *Physical Review Research 4.3* (2022), p. 033217.

Further publications which are outside of the scope of this thesis. Appearance is in chronological order († indicates equal contribution):

[6] **Maurice Weber**, Cedric Renggli, Helmut Grabner, and Ce Zhang. "Observer Dependent Lossy Image Compression." In: *42nd German Conference on Pattern Recognition (GCPR)*. 2020.

[7] Leonel Aguilar, ..., **Maurice Weber**, ..., Ce Zhang. "Ease. ML: A Lifecycle Management System for MLDev and MLOps." In: *Proceedings of the Annual Conference on Innovative Data Systems Research (CIDR)*. 2021.

[8] Nicolas Langer, **Maurice Weber**, ..., Ce Zhang. "The AI Neuropsychologist: Automatic scoring of memory deficits with deep learning." In: bioRxiv preprint, 2022.

[9] Haoxiang Wang†, **Maurice Weber**†, Josh Izaac, and Cedric Yen-Yu Lin. "Predicting Properties of Quantum Systems with Conditional Generative Models." *arXiv preprint:2211.16943*, 2022.

[10] Mintong Kang†, Linyi Li†, **Maurice Weber**, Yang Liu, Ce Zhang, and Bo Li. "Certifying Some Distributional Fairness with Subpopulation Decomposition." In: *Advances in Neural Information Processing Systems 35 (NeurIPS)*. 2022.

[11] **Maurice Weber**†, Carlo Siebenschuh†, Rory Butler†, Anton Alexandrov, Valdemar Thanner, Georgios Tsolakis, Haris Jabbar, Ian Foster, Bo Li, Rick Stevens, Ce Zhang. "WordScape: a Pipeline to extract multilingual, visually rich Documents with Layout Annotations from Web Crawl Data." In: *Advances in Neural Information Processing Systems 36 (NeurIPS)*. 2023.

Part I

PRELIMINARIES

2 ADVERSARIAL MACHINE LEARNING

We start this thesis by providing a brief review of the field of adversarial ML with an emphasis on provable defenses. While the concepts outlined here are supposed to be useful material to equip the reader with basic background on the field, this chapter should also help the reader to accurately place the present work in the landscape of adversarial ML.

2.1 ATTACKS

As deep neural networks became increasingly integrated into real-life applications, the security and reliability of these systems started to become a significant concern, especially in security-critical applications. It was first discovered in [212] that well-trained neural networks are susceptible to adversarial attacks, with further analyses conducted in [71]. Subsequently an abundance of attacks were discovered in the context of speech recognition and voice control systems [25, 273], traffic signs [123], physical-world attacks [61], and data poisoning attacks [34, 76, 263] to name just a few. More recently, attacks against neural language models have been proposed via adversarial prompts [26, 242, 281] and dataset poisoning [232].

The different types of adversarial attacks on ML systems can broadly be divided into three different categories:

(1) *Evasion Attacks* are among the most common attacks and try to evade a deployed ML system by manipulating test instances.

(2) *Poisoning Attacks*, also known as contamination attacks, occur during the development phase, where an adversary aims to poison the training data with the goal of eliciting a specific behaviour in the trained model, which can, typically during the deployment phase, be exploited.

(3) *Exploratory Attacks* do not assume access to the training data and aim to gain knowledge about the underlying model, training process, or training data.

In this thesis, we primarily deal with *evasion* attacks (Chapter 7 and Chapter 9) and *poisoning* attacks (Chapter 6).

2.2 EMPIRICAL DEFENSES

In response to the vulnerabilities outlined in the previous section, a multitude of defenses were developed which have the goal to *empirically* improve the robustness against adversarial attacks. Among the most prominent type of defenses is adversarial training [71, 124], which is based on the idea to mix adversarial examples into the training set. The hope is that that this procedure ensures that the model will predict the same class for benign and adversarial samples. In a different direction, based on the idea of model distillation [93], Papernot et al. [168] proposed to use the technique of distilling

a model as a defense. Further mechanisms to defend against evasion attacks based on Generative Adversarial Network (GAN)s [189], feature squeezing [258, 260] and reducing the transferability of adversarial examples [97] have also been proposed. Similarly, in the context of poisoning attacks, empirical defenses have been developed which in the majority rely on detection of poisoned instances [33, 67, 223, 234] or on augmenting the training process [144]

However, although these defense strategies have been shown effective against particular types of attacks, several challenges remain. Firstly, these defenses are not adaptive in the sense that, while they may block a specific kind of attack, they leave open vulnerabilities against other types of attack which might be a priori unknown. In addition, adversaries who know the defense strategies can bypass these by designing new attacks [27]. Furthermore, these defenses can incur significant computational overhead and degrade the predictive performance of ML models significantly. In particular the former challenge has motivated the development of provable defenses, which we review in the following section. This is also the central topic this thesis is concerned with.

2.3 PROVABLE DEFENSES

Provable defenses are defense methods against adversarial attacks which provide a guarantee for the robustness of an ML model under certain perturbation constraints, typically expressed in terms of ℓ_p-norm bounds on the perturbation magnitude where p is usually considered 1, 2, or ∞. Such certified robustness always serves as a lower bound to the actual, empirical robustness and much work has gone into improving such a lower bound by developing tight guarantees and training approaches favouring robustness. A recent survey on provably robust ML by Li et al. [135] proposes a taxonomy of provable defense methods and categorizes these into *complete* and *incomplete* defenses. A provable defense is considered to be complete if, whenever such a verification outputs "not verified" for an input x_0, then it is guaranteed that an adversarial x in a neighbourhod of x_0 exists. In contrast, a provable defense is considered incomplete, if the verification is allowed to abstain from verifying an input. A further distinction is then made between *deterministic* and *probabilistic* verification approaches. As the name suggests, a verification is considered to be deterministic, if the statement "verified" for an input x_0 is deterministically true. The verification is called probabilistic, if this statement only holds with high probability. It is worth remarking that, to the best of our knowledge, all currently known probabilistic provable defense are incomplete, while there are both complete and incomplete deterministic provable defenses known.

2.3.1 *Deterministic Guarantees*

Complete and deterministic approaches usually consider ℓ_∞-norm perturbations and support only feed-forward ReLU networks. In addition all these approaches have worst-case expontential time complexity, what makes these methods difficult to scale to large problem sizes (in terms of input dimension). Existing techniques in this category make use of solver-based verification [35, 175, 218], the extended simplex method [110, 111], or branch-and-bound techniques [64, 68, 237, 275]. It is worth pointing out that even though these verification approaches have exponential worst-case time complexity, in

practice they can still verify neural networks with up to 10^5 neurons, corresponding to moderate-size CIFAR-10 models.

As a remedy against scaling barriers to complete deterministic verification approaches, incomplete verification approaches via linear relaxations have been proposed [74, 153, 188, 202, 246, 252]. These approaches rely on the concept of ReLU polytopes and approximate ReLU activation functions using their convex hull of different shapes. Among these methods, interval bound propagation [74] has been shown to be the most scalable and can handle datasets up to Tiny ImageNet [259]. We refer the reader to the comprehensive survey [135] for further details and a more in-depth treatment of these methods.

2.3.2 *Probabilistic Guarantees*

Probabilistic guarantees sacrifice completeness and the deterministic nature of the previous approaches in exchange for scalability. Indeed, to the best of our knowledge, probabilistic approaches are the only robustness guarantees which scale to large datasets like ImageNet [135]. Probabilistic robustness guarantees are based on smoothed models and are otherwise known as randomized smoothing approaches. Smoothed models are constructed from a base model h by randomizing the model predictions over its inputs, where "input" can be understood in the most general sense. Formally, for a smoothing distribution μ, the smoothed model is given by the expectation

$$g(x) = \int_{\mathrm{supp}(\mu)} h(x+\delta)\, d\mu(\delta). \tag{1}$$

It is worth pointing out that the construction of such a smoothed model via an expectation value is what makes guarantees based on this technique probabilistic, since the integral can in general not be solved exactly and needs to be estimated via Monte Carlo estimation. Intuitively, in the case of classifiers, the noise effectively makes the decision boundaries smoother and suppresses regions with high curvature. Since adversarial examples aim to exploit precisely these high curvature regions, the vulnerability to such attacks is expected to be reduced. We remark that while, for the sake of clarity, we have considered additive noise in the definition of the smoothed model, this is in general not necessary. As we will see in Chapter 6 and Chapter 7, generalizations of this framework to other transformations are required in order to provide robustness guarantees for more diverse threat models.

Among the tightest randomized smoothing guarantees are those which are based on the Neyman-Pearson Lemma [39, 264], stemming from the optimality of the Neyman-Pearson tests from statistical hypothesis testing. On an intuitive level, the Neyman-Pearson approach determines the *minimum* perturbation $x_0 \to x_{adv}$ required such that an optimal hypothesis test for distinguishing the distribution of $x_0 + \delta$ from $x_{adv} + \delta$ has a low error probability. In the context of robustness, this threshold precisely determines the *maximum* perturbation a classifier can tolerate while still being guaranteed to make a correct prediction.

The majority of robustness guarantees derived via the randomized smoothing framework considers only zeroth-order information by directly querying predictions of the smoothed model [40, 56, 128, 133, 264], but methods which take first-order information into account also exist [132, 155]. Tightness improvements for these higher-order

methods can mainly be observed for ℓ_1 adversaries [135]. Finally, the choice of smoothing distribution also affects the guarantees obtained via randomized smoothing where, e. g., Gaussian noise leads to guarantees in ℓ_2 norm and Laplacian smoothing leads to guarantees in ℓ_1 norm. In this thesis, we will further investigate this in Chapter 7.

3

A PRIMER ON QUANTUM MACHINE LEARNING

In this section we provide a brief overview of quantum machine learning and present some of the basic concepts from quantum information theory used in the chapter on QML in this thesis. For a broader treatment of this topic we refer the reader to [14].

3.1 CONCEPTS

We start by outlining some of the basic concepts used in quantum information science, and which are relevant to the quantum part of this thesis. This treatment essentially corresponds to the three postulates of quantum mechanics, namely the definition of the *state space*, the *evolution* of a quantum system, and *measurements*.

STATE SPACE Associated to any isolated quantum mechanical system is a complex vector space with an inner product that induces a complete metric, i. e., a Hilbert space $\mathcal{H}$. Throughout this thesis we work in finite dimensional Hilbert spaces. Any quantum system is described by its state vector $|\psi\rangle \in \mathcal{H}$. The simplest such system is the *qubit* which has a two dimensional state space with basis vectors $|0\rangle, |1\rangle \in \mathcal{H}$, and can be written as

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \qquad \alpha, \beta \in \mathbb{C} \tag{2}$$

with the condition that the state is normalized,

$$|\langle\psi|\psi\rangle|^2 = |\alpha|^2 + |\beta|^2 = 1. \tag{3}$$

The *qubit* is the quantum counterpart to the classical *bit* as a carrier of information. While a bit can only be in the state 0 or 1, a qubit can be in states *other* than $|0\rangle$ or $|1\rangle$, i. e., in a superposition. Quantum states described by a state vector are called *pure* states and live in isolated quantum systems. However, in practice, due to interactions with the environment and other sources of noise, this assumption is often violated and quantum states are mixed. Mixed states can be seen as a probabilistic mixture of pure states,

$$\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|, \qquad \text{s.t.} \quad \sum_i p_i = 1 \tag{4}$$

where ρ is called the density operator describing the state. In more abstract terms, density operators are defined as positive semi-definite, Hermitian operators with trace equal to 1. We use the notation $\mathcal{S}(\mathcal{H})$ for the space of density operators acting on the Hilbert space $\mathcal{H}$. The fidelity of two quantum states ρ and σ measures how similar the states are and is defined as

$$F(\rho, \sigma) = \mathrm{Tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]^2 \tag{5}$$

which, in the case of pure states reduces to the squared overlap,

$$F(|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi|) = |\langle\psi|\phi\rangle|^2 . \tag{6}$$

An alternative to the fidelity is the trace distance which can be expressed as

$$T(\rho, \sigma) = \frac{1}{2} \|\rho - \sigma\|_1 \tag{7}$$

where $\|A\|_1 = \mathrm{Tr}\left[\sqrt{A^\dagger A}\right]$ is the Schatten-1 norm.

EVOLUTION The evolution of an isolated quantum system is described by a unitary matrix U acting on states of the system,

$$|\psi\rangle \mapsto U|\psi\rangle. \tag{8}$$

If the system is not isolated, then the evolution of a quantum system is described by quantum channels which are defined as Completely positive and trace preserving (CPTP) maps acting on density operators of a given state space. A map $\mathcal{E}\colon \mathcal{S}(\mathcal{H}) \to \mathcal{S}(\mathcal{H})$ is called positive if $\mathcal{E}(\rho)$ remains a positive operator and it is called completely positive if the the map $\mathcal{E} \otimes \mathbb{1}_n$ is a positive map for every $n \in \mathbb{N}$ and where $\mathbb{1}_n$ is the identity. The map $\mathcal{E}$ is called trace preserving if $\mathrm{Tr}[\mathcal{E}(\rho)] = \mathrm{Tr}[\rho]$ where Tr denotes the trace operator. For example, for any unitary matrix $\mathcal{U}$ the map

$$\rho \mapsto U\rho U^\dagger \tag{9}$$

is a CPTP map.

MEASUREMENT To extract information from a quantum system, a measurement is performed. Quantum measurements are described by a collection of measurement operators $\{M_m\}$ where the index m refers to the outcomes that can be measured using that set of operators. In full generality, the set of operators $\{M_m\}$ is a Positive Operator-valued Measure (POVM), i. e., a set of positive semi-definite Hermitian matrices with

$$\sum_m M_m = \mathbb{1}. \tag{10}$$

The probability with which measurement outcome m is observed is determined by the Born rule via the identiy

$$p(m) = \mathrm{Tr}\left[M_m \rho\right]. \tag{11}$$

When measuring a quantum system in state ρ, we change its state since we interfere with the system. The state of the system, immediately after measuring it, is then given by

$$\rho \mapsto \frac{A_m \rho A_m^\dagger}{\mathrm{Tr}\left[M_m \rho\right]}, \qquad M_m = A_m^\dagger A_m. \tag{12}$$

Properties of quantum systems are commonly expressed as the expectation value of an observable describing the property of interest. In abstract terms, an observable A is a Hermitian operator acting on the Hilbert space $\mathcal{H}$. The expectation value of A for a system in the state ρ is defined as

$$\langle A \rangle_\rho = \mathrm{Tr}\left[A\rho\right]. \tag{13}$$

This expectation value can be written as

$$\langle A \rangle_\rho = \sum_i \lambda_i \mathrm{Tr}\left[P_i \rho\right], \tag{14}$$

given the spectral decomposition $A = \sum_i \lambda_i P_i$ where λ_i are the (real) eigenvalues of A and P_i are projections onto the associated eigenspaces.

3.2 QUANTUM MACHINE LEARNING

The interplay between quantum computation and machine learning seeks to apply quantum algorithms enhance classical machine learning with the goal of enabling speed-ups of classical algorithms, or improve their accuracy. QML algorithms can be broadly categorized by the type of input data they are processing, and the type of device they are running on. In this thesis, we focus on QML algorithms that take quantum data, or classical data encoded in quantum states, as input. For other types of QML algorithms we recommend the review presented in [14].

We define a K-class quantum classifier of states of the quantum system $\mathcal{H}$ as a map $\mathcal{A}\colon \mathcal{S}(\mathcal{H}) \to \mathcal{C}$ which maps quantum states $\sigma \in \mathcal{S}(\mathcal{H})$ to class labels $k \in \mathcal{C} = \{1, \ldots, K\}$. Any such classifier is described by a CPTP map $\mathcal{E}$ and a POVM $\{M_m\}$. Formally, a quantum state σ is passed through the quantum channel $\mathcal{E}$ and subsequently the measurement $\{M_k\}$ is performed. Finally, the probability of measuring outcome m is identified with the class probability $\mathbf{y}_m(\sigma)$, i. e.

$$\sigma \mapsto \mathbf{y}_k(\sigma) := \mathrm{Tr}\left[\Pi_k \mathcal{E}(\sigma)\right]. \tag{15}$$

The final prediction is then given by the most likely class

$$\mathcal{A}(\sigma) \equiv \arg\max_k \mathbf{y}_k(\sigma). \tag{16}$$

We remark that we can treat the POVM elements M_m as projections $M_m = |m\rangle\langle m| \otimes \mathbb{1}_{d/K}$ which determines whether the output is classified into class m. This can be done without loss of generality by Naimark's dilation since $\mathcal{E}$ is kept arbitrary and potentially involves ancillary qubits and a general POVM element can be expressed as a projector on the larger Hilbert space. Throughout this work, we refer to $\mathcal{A}$ as the *classifier* and to $\mathbf{y}$ as the *score function*. In the context of QML, the input state σ can be an encoding of classical data by means of, for example, amplitude encoding or other types of encodings [125, 277], or inherently quantum input data, while $\mathcal{E}$ can be realized, for example, by a trained parametrized quantum circuit potentially involving ancillary registers [11].

Part II

BOUNDS ON EXPECTATION VALUES

4 CLASSICAL EXPECTATION VALUES

The goal of this and the subsequent chapter is to introduce the theoretical tools used to derive the core results presented in the five main chapters of this thesis. The common denominator that lays the theoretical groundwork among these chapters is the different ways in which expectation values can be bounded, under the assumption that we have minimal knowledge of the random variable over which the expectation value is taken. These observations apply to both the classical domain, where random variables commute, and to the quantum domain where the analogue to random variables are observables and probability distributions correspond to a density operator describing the quantum system of interest. Here we introduce the theory for classical probability spaces and present the results for quantum expectation values in Chapter 5.

To maintain generality, in this section we work in abstract probability spaces defined as triplets $(\Omega, \mathcal{F}, P)$ where Ω is the sample space, $\mathcal{F}$ is a σ-algebra over Ω and P is a probability measure on Ω. We consider real-valued random variables, i. e., measurable functions $X\colon \Omega \to \mathbb{R}$. The expected value of X is then defined as the Lebesgue integral

$$\mathbb{E}_{X\sim P}[X] = \int_{\Omega} X dP. \tag{17}$$

In the following sections, we present two techniques that allow to quantify the shift that occurs in the expected value $\mathbb{E}_{X\sim P}[X]$ when the probability measure P undergoes a change $P \to Q$.

4.1 BOUNDS VIA HYPOTHESIS TESTING

The first technique that we employ is based on statistical hypothesis testing and, in its core, is the technique used to derive the Neyman-Pearson based probabilistic robustness bounds [39, 264]. Hypothesis testing is a statistical problem that is concerned with the question of whether or not some hypothesis about one or more probability distributions is correct. A decision procedure for such a problem is called a statistical hypothesis test. Formally, the decision is based on the value of a realization x for a random variable X whose distribution is known to be either P, the null hypothesis, or Q, the alternative hypothesis. Given a sample $x \in \mathcal{X}$, a randomized test ϕ can be modeled as a function $\phi\colon \mathcal{X} \to [0, 1]$ which rejects the null hypothesis with probability $\phi(x)$ and accepts it with probability $1 - \phi(x)$. The two central quantities of interest are the probabilities of making a type I error, denoted by $\alpha(\phi; P)$ and the probability of making a type II error, denoted by $\beta(\phi; Q)$. The type I error corresponds to the situation where the test ϕ decides that the alternative is true, when in fact the null hypothesis is true. An error of type II occurs when the alternative is true but the test decides for the null. Formally, α and β are defined as

$$\alpha(\phi; P) := \mathbb{E}_{X\sim P}[\phi(X)] \qquad \text{(type-I error)}$$
$$\beta(\phi; Q) := \mathbb{E}_{X\sim Q}[1 - \phi(X)] \qquad \text{(type-II error)}$$

The problem is then to select the test ϕ which minimizes the probability of making a type II error, subject to the constraint that the probability of making a type-I error is below a given threshold α_0. When testing the null hypothesis $X \sim P$ against the alternative $X \sim Q$, the Neyman-Pearson Lemma [162] states that a likelihood ratio test ϕ_{NP} is optimal in the sense that it admits the smallest probability of making a type-II error, among all tests which have type-I error probability bounded by α_0. The likelihood ratio test ϕ_{NP} is defined as

$$\phi_{NP}(x) = \begin{cases} 1 & \Lambda(x) > t \\ q & \Lambda(x) = t \\ 0 & \Lambda(x) < t \end{cases}, \quad \text{with} \quad \Lambda(x) = \frac{f_Q(x)}{f_P(x)} \tag{18}$$

where $f_P = \frac{dP}{d\mu}$ and $f_Q = \frac{dQ}{d\mu}$ are the probability density functions with respect to a reference measure μ on $\mathcal{X}$. The values of t and q are chosen such that

$$\alpha(\phi_{NP}; P) = q\mathbb{P}_{X\sim P}[\Lambda(X) = t] + \mathbb{P}_{X\sim P}[\Lambda(X) > t] = \alpha_0 \tag{19}$$

Formally, optimality of the likelihood ratio test means that it is a solution to the optimization problem

$$\beta^*(\alpha_0; P, Q) = \inf\{\beta(\phi; Q) \,|\, \alpha(\phi; P) \leqslant \alpha_0\}. \tag{20}$$

In Section A.1, we explicitly show optimality and existence of a likelihood ratio test at a specific significance level α_0. We can now state the first bounds for expectation values.

Lemma 1. *Let $h\colon \mathcal{X} \to [0, 1]$ be a deterministic function and let P, Q be probability measures on $\mathcal{X}$. Let $\underline{m}, \overline{m} \in [0, 1]$ such that*

$$\underline{m} \leqslant \mathbb{E}_{X\sim P}[h(X)] \leqslant \overline{m}. \tag{21}$$

Then, we have

$$\beta^*(1 - \underline{m}; P, Q) \leqslant \mathbb{E}_{X\sim Q}[h(X)] \leqslant 1 - \beta^*(\overline{m}; P, Q). \tag{22}$$

Proof. We first show the lower bound. By assumption, we have

$$\mathbb{E}_{X\sim P}[1 - h(X)] \leqslant 1 - \underline{m} \tag{23}$$

and hence, by definition of β^* and viewing h as a hypothesis test,

$$\mathbb{E}_{X\sim Q}[h(X)] = \beta(1 - h; Q) \geqslant \beta^*(1 - \underline{m}; P, Q). \tag{24}$$

To show the upper bound, we proceed analogously. It follows directly from the assumption and the definition of β^* that

$$\mathbb{E}_{X\sim Q}[h(X)] = 1 - \beta(h; Q) \leqslant 1 - \beta^*(\overline{m}; P, Q). \tag{25}$$

□

While Lemma 1 follows as a direct consequence of the Neyman-Pearson Lemmma [162], many of the challenges arise when instantiating this result for specific distributions and for specific applications. In Chapter 6 we will apply this result to certify the robustness of ML models against backdoor attacks, where an instantiation using Gaussian noise leads to a first robustness bound which, in practice, needs further modifications to arrive at non-trivial robustness certificates. In Chapter 7, we instantiate Lemma 1 for several distributions P and Q, and use it to certify robustness against a multitude of semantic transformations like rotations, scaling, or Gaussian blur. Finally, in Section 5.1 we state the quantum analogue of Lemma 1.

4.2 BOUNDS VIA GRAM MATRICES

Here we introduce a second method to bound expectation values under shifts in the underlying distribution. The technique was pioneered by Weinhold [243] where it was used to derive bounds for expectation values of pure states in the context of quantum chemistry. Here, we derive the classical version of this result, while in Section 5.2 we present an extension of the original result for mixed quantum states.

Let Σ be the Borel σ-algebra and μ a measure on X, and consider the Hilbert space $L_2(\mathcal{X}, \Sigma, \mu)$ of real-valued, square-integrable functions $f\colon \mathcal{X} \to \mathbb{R}$, endowed with the inner product

$$\langle f, g\rangle = \int_{\mathcal{X}} fg \, d\mu. \tag{26}$$

We can identify any probability distribution P on $\mathcal{X}$, which is absolutely continuous[1] with respect to μ, with a unit vector in $L_2(\mathcal{X}, \Sigma, \mu)$, via the square root of its Radon-Nikodym derivative,

$$\psi_P := \sqrt{\frac{dP}{d\mu}}. \tag{27}$$

With this, we can now define the Hellinger distance, which measures the similarity between two probabilty measures P and Q and can be written in terms of the inner product (26).

Definition 1 (Hellinger-distance)**.** *Let P, Q be probability measures on $\mathcal{X}$ that are absolutely continuous with respect to a reference measure μ, $P, Q \ll \mu$. The Hellinger distance between P and Q is defined as*

$$H(P, Q)^2 := \frac{1}{2}\int_{\mathcal{X}} \left(\sqrt{\frac{dP}{d\mu}} - \sqrt{\frac{dQ}{d\mu}}\right)^2 d\mu = 1 - \langle \psi_P, \psi_Q\rangle \tag{28}$$

The Hellinger distance is independent of the choice of the reference measure μ.

Let $h \in L_\infty(\mathcal{X}, \Sigma, \mu)$ be any essentially bounded function. We can rewrite the expectation as

$$\mathbb{E}_{X\sim P}[h(X)] = \int_{\mathcal{X}} h dP = \int_{\mathcal{X}} h\sqrt{\frac{dP}{d\mu}}\sqrt{\frac{dP}{d\mu}}\, d\mu = \langle \psi_P, h\cdot\psi_P\rangle, \tag{29}$$

where the product $h\cdot\psi_P$ is defined as a pointwise multiplication. Similarly, the variance can be written as

$$\mathbb{V}_{X\sim P}[h(X)] = \mathbb{E}_{X\sim P}[h(X)^2] - \mathbb{E}_{X\sim P}[h(X)]^2 = \langle \psi_P, f^2\cdot\psi_P\rangle - \langle \psi_P, f\cdot\psi_P\rangle^2. \tag{30}$$

Let Q be a further probability measure on $\mathcal{X}$ and consider the Gram matrix G of the elements ψ_Q, ψ_P, and $h\cdot\psi_P$,

$$G := \begin{pmatrix} 1 & \gamma & \langle \psi_Q, h\cdot\psi_P\rangle \\ \gamma & 1 & \langle \psi_P, h\cdot\psi_P\rangle \\ \langle h\cdot\psi_P, \psi_Q\rangle & \langle h\cdot\psi_P, \psi_P\rangle & \langle h\cdot\psi_P, h\cdot\psi_P\rangle \end{pmatrix}. \tag{31}$$

1 We say that a measure ν on $\mathcal{X}$ is absolutely continuous with respect to another measure μ denoted by $\nu \ll \mu$, if for any $A \in \Sigma$, $\mu(A) = 0 \Rightarrow \nu(A) = 0$.

where we have set $\gamma = \langle \psi_Q, \psi_P \rangle$. Since the matrix G is positive semidefinite (see, e. g., Theorem 7.2.10 in [96]), its determinant is non-negative and is given by

$$\det(G) = a \cdot \langle \psi_Q, h \cdot \psi_P \rangle^2 + b \cdot \langle \psi_Q, h \cdot \psi_P \rangle + c \tag{32}$$

with coefficients

$$\begin{aligned} a = -1, \;\; b = 2 \cdot \gamma \cdot \langle \psi_P, h \cdot \psi_P \rangle \\ c = (1-\gamma^2)\langle h \cdot \psi_P, h \cdot \psi_P \rangle - \langle h \cdot \psi_P, \psi_P \rangle^2. \end{aligned} \tag{33}$$

In other words, the determinant $\det(G)$ is a second-degree polynomial $\pi(x)$ evaluated at $x = \langle \psi_Q, h \cdot \psi_P \rangle$. Its non-negativity then restricts the values which $\langle \psi_Q, h \cdot \psi_P \rangle$ can take to be bounded within the roots of π,

$$\frac{b}{2} - \sqrt{\frac{b^2}{4} + c} \leqslant \langle \psi_Q, h \cdot \psi_P \rangle \leqslant \frac{b}{2} + \sqrt{\frac{b^2}{4} + c}. \tag{34}$$

For positive $h \geqslant 0$, applying the Cauchy-Schwarz inequality yields

$$\begin{aligned} \frac{b}{2} - \sqrt{\frac{b^2}{4} + c} &\leqslant \langle \psi_Q, h \cdot \psi_P \rangle \\ &\leqslant \sqrt{\langle \psi_Q, h \cdot \psi_Q \rangle \cdot \langle \psi_P, h \cdot \psi_P \rangle} \\ &= \sqrt{\mathbb{E}_{X\sim P}[h(X)] \cdot \mathbb{E}_{X\sim Q}[h(X)]} \end{aligned} \tag{35}$$

Finally, under the condition that $c \leqslant 0$, we can square both sides of the inequality and rearrange terms to get the lower bound

$$\mathbb{E}_Q[h(X)] \geqslant \mathbb{E}_P[h(X)] - 2\sqrt{\gamma^2\,(1-\gamma^2)\,\mathbb{V}_P[h(X)]} - (1-\gamma^2)\left[\mathbb{E}_P[h(X)] - \frac{\mathbb{V}_P[h(X)]}{\mathbb{E}_P[h(X)]}\right] \tag{36}$$

It is worth noting that this bound is monotonically decreasing with decreasing γ and expectation $\mathbb{E}_P[h(X)]$, and it is decreasing as the variance $\mathbb{V}_P[h(X)]$ gets larger. This allows us to estimate this bound based on estimates of these quantities. An upper bound can be obtained by applying the above reasoning to the function $\tilde{h} := M_h - h$ where $M_h = \sup_{x\in\mathcal{X}} |h(x)|$. In summary, we have shown the following result:

Theorem 1. *Let $h\colon \mathcal{X} \to [0, M]$ be a deterministic, square-integrable function and let P be a probability measure on $\mathcal{X}$. Let $\underline{m}, \overline{m} \in [0, M]$, and $\overline{v} \in [0, M^2]$ such that*

$$\underline{m} \leqslant \mathbb{E}_{X\sim P}[h(X)] \leqslant \overline{m}, \quad \textit{and} \quad \mathbb{V}_{X\sim P}[h(X)] \leqslant \overline{v}. \tag{37}$$

Then, for any probability measure Q on $\mathcal{X}$ with $H(P, Q) \leqslant \rho$, and $\delta^2 \leqslant 1 - \sqrt{\frac{\overline{v}}{\overline{v}+\underline{m}^2}}$, we have the lower bound

$$\mathbb{E}_{X\sim Q}[h(X)] \geqslant \underline{m} - 2\sqrt{C_\rho(1-C_\rho)\overline{v}} - C_\rho\frac{\underline{m}^2 - \overline{v}}{\underline{m}} \tag{38}$$

where we have defined $C_\rho = \rho^2(2-\rho^2)$. Similarly, for $\rho^2 \leqslant 1 - \sqrt{\frac{\overline{v}}{\overline{v}+(M-\overline{m})^2}}$, we have the upper bound

$$\mathbb{E}_{X\sim Q}[h(X)] \leqslant \overline{m} + 2\sqrt{C_\rho(1-C_\rho)\overline{v}} + C_\rho\frac{(M-\overline{m})^2 - \overline{v}}{M-\overline{m}}. \tag{39}$$

While this bound is stated in a generic manner here, we will use this technique in Chapter 8 to derive guarantees for out-of-domain generalization of ML models. In Section 5.2, we will see the quantum analogue of this Theorem and circle back to the original derivation proposed by Weinhold in [243], before presenting a generalization.

It is worth pointing out the similarities and differences between this result, which we have derived using the non-negativity of Gram matrices and the mapping of probability measures to a suitable Hilbert space, and the result presented in Lemma 1 which is based on the Neyman-Pearson Lemma. First, we notice that both results do not rely on properties of the function h, apart from the expectation value and, in the case of Theorem 1, the variance. These are both quantities which can be measured from observations and only require black-box access to the function. Second, both results rely on a notion of statistical similarity and are faithful in the sense that as Q approaches P, the bounds approach the expectation values of h under the distribution P. Finally, the bounds based on the Neyman-Pearson Lemma only rely on the first moment of the distribution of $h(X)$, while the Gramian bounds additionally rely on the second moment of the distribution via its variance.

5 QUANTUM EXPECTATION VALUES

Similar to the classical case, here we derive bounds on expectation values of quantum observables, the quantum counterpart to random variables. Specifically, let $\mathcal{H}$ be a Hilbert space with finite dimension $d = \dim(\mathcal{H}) < \infty$, and let $\sigma \in \mathcal{S}(\mathcal{H})$ be a quantum state. For an observable $A \in \mathcal{L}(\mathcal{H})$, its expectation under σ is then defined as

$$\langle A \rangle_\sigma = \operatorname{Tr}[A\sigma] \tag{40}$$

In the following two sections, we present the quantum analogues of the bounds presented in Section 4.1 and Section 4.2, allowing us to bound the shift incurred to the expectation $\langle A \rangle_\sigma$ when σ undergoes a change $\sigma \to \rho$.

5.1 BOUNDS VIA QUANTUM HYPOTHESIS TESTING

QHT is typically formulated in terms of state discrimination where several quantum states have to be discriminated through a measurement [88]. In binary QHT, the aim is to decide whether a given unknown quantum system is in one of two states corresponding to the null and alternative hypothesis. Any such test is represented by an operator $0 \leqslant M \leqslant \mathbb{1}_d$, which corresponds to rejecting the null hypothesis in favor of the alternative hypothesis. Analogous to classical hypothesis testing, the two central quantities of interest are the probabilities of making an error of type I or and error of type II. The former corresponds to rejecting the null hypothesis when it is true, while the latter occurs if the null is accepted when the alternative hypothesis is true. Specifically, for density operators $\sigma \in \mathcal{S}(\mathcal{H})$ and $\rho \in \mathcal{S}(\mathcal{H})$ describing the null and alternative hypothesis, the type-I error probability is defined as $\alpha(M; \sigma)$, and the type-II error probability as $\beta(M; \rho)$, so that

$$\alpha(M; \sigma) := \operatorname{Tr}[\sigma M] \qquad \text{(type-I error)}$$
$$\beta(M; \rho) := \operatorname{Tr}[\rho(\mathbb{1} - M)] \qquad \text{(type-II error)}$$

Here we consider the Neyman-Pearson approach to QHT [90], where the two types of errors are associated with a different cost.[1] Given a maximal allowed probability for the type I error, the goal is to minimize the probability of the type II error. Specifically, one aims to solve the Semidefinite Programming (SDP) problem

$$\begin{aligned} \beta^*(\alpha_0; \sigma, \rho) := \text{minimize} \quad & \beta(M; \rho) \\ \text{s.t.} \quad & \alpha(M; \sigma) \leqslant \alpha_0, \\ & 0 \leqslant M \leqslant \mathbb{1}_d. \end{aligned} \tag{41}$$

Optimal tests can be expressed in terms of projections onto the eigenspaces of the operator $\rho - t\sigma$ where t is a non-negative number. More specifically, for $t \geqslant 0$ let $P_{t,+} :=$

1 In contrast, in the Bayesian setting, the hypotheses σ and ρ occur with some prior probabilities π_0 and π_1 and the goal is to find a test which minimizes the total error probability. A Bayes optimal test M is one that minimizes the posterior probability $\pi_0 \cdot \alpha(M) + \pi_1 \cdot \beta(M)$.

$\{\rho - t\sigma > 0\}$, $P_{t,-} := \{\rho - t\sigma < 0\}$ and $P_{t,0} := \mathbb{1} - P_{t,+} - P_{t,-}$ be the projections onto the eigenspaces of $\rho - t\sigma$ associated with positive, negative and zero eigenvalues. The quantum analogue to the Neyman-Pearson Lemma [162] shows optimality of Helstrom operators [88] which essentially correspond to likelihood ratio tests

$$M_t := P_{t,+} + X_t, \quad 0 \leqslant X_t \leqslant P_{t,0}. \tag{42}$$

The choice of the scalar $t \geqslant 0$ and the operator X_t is such that the preassigned type-I error probability α_0 is attained. In Section B.1.1 we derive an explicit construction of these operators and prove their optimality. We can now state bounds on expectation values using QHT.

Lemma 2. *Let $A \in \mathcal{L}(\mathcal{H})$ be a quantum observable with $0 \leqslant A \leqslant \mathbb{1}_d$, and let $\rho, \sigma \in \mathcal{S}(\mathcal{H})$ be quantum states. Let $\underline{m}, \overline{m} \in [0, 1]$ such that*

$$\underline{m} \leqslant \langle A \rangle_\sigma \leqslant \overline{m}. \tag{43}$$

Then, we have

$$\beta^*(1 - \underline{m}; \sigma, \rho) \leqslant \langle A \rangle_\rho \leqslant 1 - \beta^*(\overline{m}; \sigma, \rho). \tag{44}$$

Proof. The proof of this result follows immediately from the definition of β^*. Indeed, note that, by definition

$$\begin{aligned} \beta^*(1 - \underline{m}; \sigma, \rho) &= \inf\{\beta(M; \rho) \,|\, \alpha(M; \sigma) \leqslant 1 - \underline{m}\} \\ &\leqslant \beta(\mathbb{1}_d - A; \rho) = \langle A \rangle_\rho. \end{aligned} \tag{45}$$

Similarly, the upper bound can be seen using

$$\begin{aligned} \beta^*(\overline{m}; \sigma, \rho) &= \inf\{\beta(M; \rho) \,|\, \alpha(M; \sigma) \leqslant \overline{m}\} \\ &\leqslant \beta(A; \rho) = 1 - \langle A \rangle_\rho. \end{aligned} \tag{46}$$

and thus

$$\langle A \rangle_\rho \leqslant 1 - \beta^*(\overline{m}; \sigma, \rho). \tag{47}$$

This concludes the proof. □

In analogy to the result based on classical hypothesis testing presented in Lemma 1, this Lemma is a direct consequence of the optimality of the Helstrom operators. The main challenge when using these bounds is to derive an operationally convenient form that makes the dependence of the bounds on a distance between two quantum states explicit. In the following Theorem, we establish the explicit form which presents itself as a lower bound on the optimal type-II error probability β^*. It is interesting to note that the distance that emerges here is the fidelity between quantum states which is essentially the quantum analogue of the Hellinger distance seen in Section 4.2.

Lemma 3. *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H}_d)$ be arbitrary quantum states, $\alpha_0 \in [0, 1]$ and $\epsilon \in [0, 1 - \alpha_0]$. Suppose that $\mathcal{F}(\rho, \sigma) \geqslant 1 - \epsilon$. Then*

$$\beta^*(\alpha_0; \rho, \sigma) \;\geqslant\; \alpha_0(2\epsilon - 1) + (1 - \epsilon) - 2\sqrt{\alpha_0 \epsilon (1 - \alpha_0)(1 - \epsilon)} \tag{48}$$

with equality if the states σ and ρ are pure and $\mathcal{F}(\rho, \sigma) = 1 - \epsilon$.

The proof of this result is based on an explicit construction of the Helstrom operators (42). We state the full proof in Section B.1. In Chapter 9 we use these results to establish a robustness guarantee for quantum classifiers, while in Chapter 10 we use the result to bound expectation values of quantum observables in the context of naturally occurring quantum noise.

5.2 BOUNDS VIA GRAM MATRICES

Here we present the quantum analogue to the technique presented in Section 4.2 which is based on the non-negativity of Gram matrices. In contrast to the classical case, here we don't need to construct an explicit mapping of probability measures to a Hilbert space since quantum expectation values are naturally expressed in terms of inner products. To provide intuition, we start with a review of the technique presented in [243] for the case of pure states and provide the full proof for mixed states in Section B.2.

Consider a Hermitian operator $A \in \mathcal{L}(\mathcal{H})$ and pure states $|\psi\rangle$ and $|\phi\rangle$. The Gram matrix for the vectors $|\psi\rangle$, $|\phi\rangle$, $A|\phi\rangle$ is given by

$$G = \begin{pmatrix} 1 & \gamma & \langle\psi|A|\phi\rangle \\ \gamma & 1 & \langle\phi|A|\phi\rangle \\ \langle\phi|A|\psi\rangle & \langle\phi|A|\phi\rangle & \langle\phi|A^2|\phi\rangle \end{pmatrix} \tag{49}$$

where we have set $\gamma = \langle\psi|\phi\rangle$ which, without loss of generality, is assumed to be real-valued (otherwise multiply each state by a global phase). The Gram matrix G is positive semidefinite, its determinant is non-negative and is given by

$$\det(G) = a \cdot \Re(\langle\psi|A|\phi\rangle)^2 + b \cdot \Re(\langle\psi|A|\phi\rangle) + c \tag{50}$$

with coefficients

$$a = -1, \;\; b = 2\gamma\langle A\rangle_\phi, \;\; c = (1-\gamma^2)\langle A^2\rangle_\phi - \langle A\rangle_\phi^2 - \Im(\langle\psi|A|\phi\rangle)^2 \tag{51}$$

and where $\Re(z)$ and $\Im(z)$ denote the real and imaginary parts of a complex number $z \in \mathbb{C}$. In analogy to the classical derivation, we view the determinant $\det(G)$ as a second-degree polynomial $\pi(x)$ evaluated at $x = \Re(\langle\psi|A|\phi\rangle)$. Due to its non-negativity, the roots of π then bound the values which $\Re(\langle\psi|A|\phi\rangle)$ can take to

$$\gamma\langle A\rangle_\phi - \Delta A_\phi\sqrt{1-\gamma^2} \leqslant \Re(\langle\psi|A|\phi\rangle) \leqslant \gamma\langle A\rangle_\phi + \Delta A_\phi\sqrt{1-\gamma^2} \tag{52}$$

where we have taken into account that $\Im(\langle\psi|A|\phi\rangle)^2 \geqslant 0$ and defined the variance as

$$(\Delta A_\phi)^2 = \langle A^2\rangle - \langle A\rangle^2. \tag{53}$$

For positive semidefinite operators $A \geqslant 0$, and $\gamma \geqslant 1 - \frac{\langle A\rangle_\phi^2}{\langle A^2\rangle_\phi}$ the Cauchy-Schwarz inequality yields the lower bound

$$\langle A\rangle_\psi \geqslant (2\gamma^2-1)\langle A\rangle_\phi - 2\Delta A_\phi\sqrt{\gamma^2(1-\gamma^2)} + (1-\gamma^2)\frac{\langle A^2\rangle_\phi}{\langle A\rangle_\phi^2}. \tag{54}$$

While this result only holds for pure states, we can extend this to mixed states by making use of purifications and working in the enlarged Hilbert space. Here we state the full result and provide a detailed derivation in Section B.2.

Lemma 4. *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H}_d)$ be density operators with fidelity $\mathcal{F}(\rho,\sigma) \geqslant 1-\epsilon$ for some $\epsilon \geqslant 0$ and let $A \geqslant 0$ be an observable. Let $\underline{m}, \bar{v} \geqslant 0$ such that*

$$\langle A\rangle_\sigma \geqslant \underline{m} \quad \textit{and} \quad \Delta A_\sigma \leqslant \bar{v}. \tag{55}$$

For ϵ with $\epsilon \leqslant \frac{\overline{m}^2}{\overline{m}^2+\bar{v}^2}$, a lower bound of $\langle A\rangle_\rho$ can be expressed as

$$\langle A\rangle_\rho \geqslant (1-\epsilon)\overline{m} - 2\bar{v}\sqrt{\epsilon(1-\epsilon)} + \epsilon\frac{\bar{v}^2}{\overline{m}}. \tag{56}$$

In Chapter 10 we apply this result in the context of the VQE and present bounds on expectation values for specific observables relevant in quantum chemistry, as well as lower and upper bounds on Eigenvalues of observables.

Part III

CLASSICAL MACHINE LEARNING

6 PROVABLE ROBUSTNESS AGAINST BACKDOOR ATTACKS

We start the classical part of this thesis at the early stages of any ML pipeline. At this stage, next to important design decisions such as the choice of model and training algorithm, ML practitioners and researchers alike build training datasets. These can be contributed by outsiders and labelled by crowd workers, from publicly available sources, such as the world wide web, or privately held data sources such as customer data. If an adversary has access to this process, either directly or indirectly, they can poison the dataset by manipulating training instances in ways that can be exploited once the model has been deployed. It is the topic of this chapter to provide a means to guarantee that such an attack fails, i. e., that the model is provably robust. In the subsequent chapters, we will then focus on attacks to which the model is vulnerable once it has been trained, during deployment.

6.1 INTRODUCTION

6.1.1 *Overview*

Building ML algorithms that are robust to adversarial attacks has been an emerging topic over the last decade. There are mainly two different types of adversarial attacks, namely (1) evasion attacks, and (2) Data poisoning attacks. In the former, attackers manipulate the test examples against a trained ML model, while in the latter, attackers are allowed to perturb the training set. Here we focus on backdoor attacks, a particular type of data poisoning attack while in the subsequent chapters we focus on evasion attacks. To carry out a backdoor attack, an attacker adds small backdoor patterns to a subset of training instances such that the trained model is biased toward test instances with the same patterns [34, 75]. In this work, we present the first certification process, referred to as RAB, which offers provable robustness for ML models against backdoor attacks.

Both evasion and data poisoning attacks have attracted intensive interests from academia as well as industry [71, 255, 263, 279]. In response, several empirical solutions have been proposed as defenses against evasion attacks [27, 142, 260, 266]. For instance, adversarial training has been proposed to retrain the ML models with generated adversarial examples [146], and quantization has been applied to either inputs or neural network weights to defend against potential adversarial instances [260]. However, recent studies have shown that these defenses are not resilient against intelligent adversaries responding dynamically to the deployed defenses [7, 27].

As a result, one exciting line of research aims to develop *certifiably robust* algorithms against *evasion attacks*, including both deterministic and probabilistic certification approaches [135]. In particular, among these certified robustness approaches, only randomized smoothing and its variations are able to provide certified robustness against evasion attacks on large-scale datasets such as ImageNet [39, 128, 265]. On a high level, the randomized smoothing-based approaches are able to certify the robustness of a

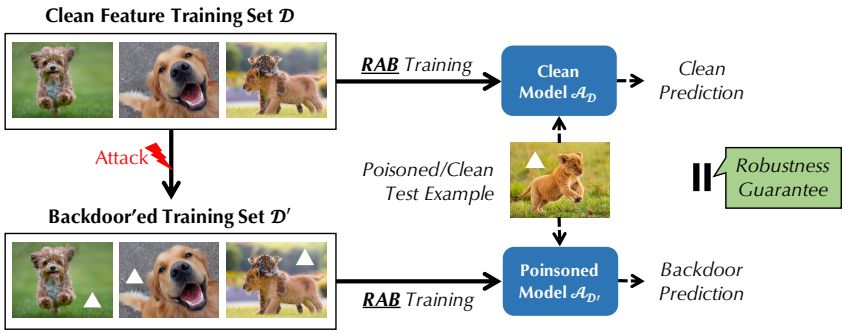


Figure 2: Given a poisoned dataset $\mathcal{D}'$ – generated by adding backdoor patterns Δ to some instances in the dataset $\mathcal{D}$ with clean features – RAB guarantees that, for all test examples x, $\mathcal{A}_{\mathcal{D}'}(x) = \mathcal{A}_{\mathcal{D}}(x)$, with high probability when the magnitude of the backdoor pattern Δ is within the certification radius.

smoothed classifier, by making a consistent prediction for an adversarial input as long as the perturbation is within a certain radius. The smoothed classifier is obtained by taking the expectation over the possible outputs given a set of randomized inputs which are generated by adding noise drawn from a smoothing distribution.

Despite these recent developments on certified robustness against *evasion attacks*, only empirical studies have been conducted to defend against *backdoor attacks* [67, 76, 134, 234], and the question of how to improve and certify the robustness of given ML models against backdoor attacks remains largely unanswered. To the best of our knowledge, there is no certifiably robust strategy to deal with backdoor attacks yet. Naturally, we ask: *Can we develop certifiably robust ML models against backdoor attacks?*

It is clear that extending existing certification methods against evasion attacks to attacks that occur during training is challenging given these two significantly different threat models. For instance, even certifying a label flipping training-time attack is non-trivial as illustrated in [184], which proposes to certify against a label flipping attack by setting a limit to how many labels in the training set may be flipped such that it does not affect the final prediction. Similar to this work, [184] also leverages randomized smoothing. As backdoor attacks involve both label flipping and instance pattern manipulations, providing certifications can be even more challenging.

6.1.2 *Contributions*

In this work, we present a certification process, referred to as RAB, which offers probabilistic robustness guarantees for ML models against backdoor attacks. As shown in Figure 2, our certification goal is to guarantee *that a test instance, which may contain backdoor patterns, will be classified consistently, independent of whether the models were trained on data with or without backdoors, as long as the embedded backdoor patterns are within an* L_p*-ball of radius* R. We formally define the corresponding threat model and our certification goal in Section 6.1.5.

Our approach to achieving this is mainly inspired by randomized smoothing, a technique to certify robustness against evasion attacks [39], but goes significantly beyond it due to the different settings. Our **first contribution** is to develop a theoretical framework to generalize randomized smoothing to a much larger family of functions and smoothing distributions. This allows us to support cases in which a classifier is a func-

tion that takes as input a test instance *and* a training set. With our framework, we can *(1) provide robustness certificates against both evasion and dataset poisoning attacks; (2) certify any classifier which takes as input a tuple of test instance and training dataset* and *(3) prove that the derived robustness bound is tight*. Given this general framework, we can enable a basic version of the proposed RAB framework. At a high level, as shown in Figure 3, given a training set $\mathcal{D}$, RAB generates N additional "smoothed" training sets $\mathcal{D} + \epsilon_i$ by adding noise ϵ_i ($i \in \{1, \ldots, N\}$) drawn from a smoothing distribution. Then, for each of these N training sets, a corresponding classifier is trained, resulting in an ensemble of N different classifiers. These models are then aggregated to generate a "smoothed classifier" for which we prove that its output will be consistent regardless of whether there are backdoors added during training, as long as the backdoor patterns satisfy certain conditions.

However, this basic version is not enough to provide satisfactory certified robustness against backdoor attacks. When we instantiate our theoretical framework with a practical training pipeline to provide certified robustness against backdoor attacks, we need to further develop nontrivial techniques to improve (1) the certification radius and (2) the certification efficiency. Our **second contribution** consists of two non-trivial technical optimizations. To improve the *certification radius*, we certify Deep Neural Network (DNN) classifiers with a data augmentation scheme enabled by hash functions. This provides additional guidance for improving the certified robustness against backdoor attacks and we hope that it can inspire other research in the future. To improve the *certification efficiency*, we observed that for certain families of classifiers, namely K-nearest neighbor classifiers, we can develop an efficient algorithm to compute the smoothing result *exactly, eliminating the need to resort to Monte Carlo algorithms as for generic classifiers*.

Finally, our **third contribution** is an extensive benchmark, evaluating our framework RAB on multiple ML models and provide the first collection of certified robustness bounds on a diverse range of datasets, namely MNIST, CIFAR-10, ImageNette, and on spambase tabular data. We hope that these experiments and benchmarks can provide future directions for improving the robustness of ML models against backdoor attacks.

In summary, in this work we make the following set of technical contributions:

- We propose a unified framework to certify the model robustness against both evasion and backdoor attacks and prove that our robustness bound is tight.
- We provide the first certifiable robustness bound for general ML models against backdoor attacks considering *different* smoothing noise distributions.
- We propose an exact and efficient smoothing algorithm for k-Nearest Neighbors (KNN) models eliminating the need to sample random noise during training.
- We conduct extensive reproducible large-scale experiments and provide a benchmark for certified robustness against three representative backdoor attacks for multiple types of models on diverse datasets. We also provide a series of ablation studies to further analyze the factors that affect model robustness against backdoor attacks.

6.1.3 *Related Work*

BACKDOOR ATTACKS There have been several works developing optimal poisoning attacks against ML models such as SVM and logistic regression [15, 134]. Furthermore, [159] proposes a similar optimization-based poisoning attack against neural networks that can only be applied to shallow MLP models. In addition to these optimization-based poisoning attacks, the backdoor attacks are shown to be very effective against deep neural networks [34, 75]. The backdoor patterns can be either static or generated dynamically [263]. Static backdoor patterns can be as small as one pixel, or as large as an entire image [34].

EMPIRICAL DEFENSES AGAINST BACKDOOR ATTACKS Given the potentially severe consequences caused by backdoor attacks, multiple defense approaches have been proposed. NeuralCleanse [234] proposes to detect the backdoored models based on the observation that there exists a "short path" to make an instance to be predicted as a malicious one. [33] improves upon the approach by using model inversion to obtain training data, and then applying GANs to generate the "short path" and apply anomaly detection algorithm as in Neural Cleanse. Activation Clustering [32] leverages the activation vectors from the backdoored model as features to detect backdoor instances. Spectral Signature [223] identifies the "spectral signature" in the activation vector for backdoored instances. STRIP [67] proposes to identify the backdoor instances by checking whether the model will still provide a confident answer when it sees the backdoor pattern. SentiNet [36] leverages computer vision techniques to search for the parts in the image that contribute the most to the model output, which are very likely to be the backdoor pattern. In [144], differential privacy has been leveraged as a defense against poisoning attacks. Note that RAB can not guarantee that the trained models are differentially private, although both aim to decrease the model sensitivity intuitively. A further empirical defense against backdoor attacks is proposed in [82] using covariance estimation with the aim of amplifying the spectral signature of backdoored instances.

CERTIFIED DEFENSES AGAINST POISONING ATTACKS Another interesting application of randomized smoothing is presented in [184] to certify the robustness against label-flipping attacks and randomize the entire training procedure of the classifier by randomly flipping labels in the training set. This work is orthogonal to ours in that we investigate the robustness with respect to perturbations on the training inputs rather than labels. BagFlip [276] proposes a robustness guarantee against both feature perturbations and label flips by combining bagging and randomised smoothing. However, the technique does not scale to large problem sizes due to the requirement of preparing large lookup tables ($\mathcal{O}(N^2)$) for the computation of the certified radius during inference. In a further line of work on provable defenses against poisoning attacks, [131] proposes an ensemble method, deep partition aggregation (DPA). Similar to our work, DPA is related to randomized smoothing, however, in contrast to our work, the goal is to certify the number of poisoned instances for which the prediction remains unaffected. Similarly, [107] use an ensemble technique to certify robustness against poisoning attacks. This is also orthogonal to ours as it certifies the number of poisoned instances, rather than the trigger size. The same certification goal is considered in [108], but is restricted to nearest neighbor algorithms and derives an intrinsic certificate by viewing them as ensemble methods. Drews et al. [49] provide data poisoning robustness

guarantees for decision-trees by using techniques based on abstract interpretation. In addition to these works aiming to certify the robustness of a single model, [267] provides a new way to certify the robustness of an end-to-end sensing-reasoning pipeline. Finally, [257] propose a technique to certify robustness against backdoor attacks within the federated learning framework by controlling the global model smoothness. Furthermore, a technical report also proposes to directly apply the randomized smoothing technique to certify robustness against backdoor attacks without any evaluation or analysis [233]. In addition, as we will show, directly applying randomized smoothing will not provide high certified robustness bounds. Contrary to that, here, we first provide a unified framework based on randomized smoothing, and then propose the RAB robust training process to provide certified robustness against backdoor attacks based on the framework. We provide the tightness analysis for the robustness bound, analyze different smoothing distributions, and propose the hash function-based model deterministic test-time augmentation approach to achieve good certified robustness. In addition, we analyze different ML models with corresponding properties such as model smoothness to provide guidance to further improve the certified robustness.

6.1.4 *Background on Backdoor attacks*

A backdoor attack aims to inject “backdoor" patterns into the training set and associate such patterns with a specific adversarial target (i. e., label). As a result, during testing time, any test instance with such a pattern will be misclassified as the preselected adversarial target [34, 76]. ML models with injected backdoors are called *backdoored models* and they are typically able to achieve performance similar to clean models on benign data, making it challenging to detect whether the model has been backdoored.

There are several ways to categorize backdoor attacks. First, based on the *adversarial target design*, the attacks can be characterized either as *single target attacks* or *all-to-all attacks*. In a *single target attack*, the backdoor pattern will cause the poisoned classifier to always return a designed target label. An *all-to-all* attack leverages the backdoor pattern to permute the classifier results. The second categorization is based on *different types of backdoor patterns*. There are *region based* and *blending* backdoor attacks. In the *region based* attack, a specific region of the training instance is manipulated in a subtle way that will not cause human notification [76, 279]. In particular, it has been shown that such backdoor patterns can be as small as only one or four pixels [223]. On the other hand, Chen et al. [34] show that by blending the whole instance with a certain pattern such as a fixed random noise pattern, it is possible to generate effective backdoors to poison the ML models. In this work, we focus on certifying the robustness against general backdoor attacks, where the attacker is able to add any specific or uncontrollable random backdoor patterns for arbitrary adversarial targets.

6.1.5 *Method Overview*

Here we first present the threat model, and then introduce the method overview, where we define our robustness guarantee.

NOTATION We write random variables as uppercase letters X and use the notation $\mathbb{P}_X$ to denote the probability measure induced by X and write f_X to denote the prob-

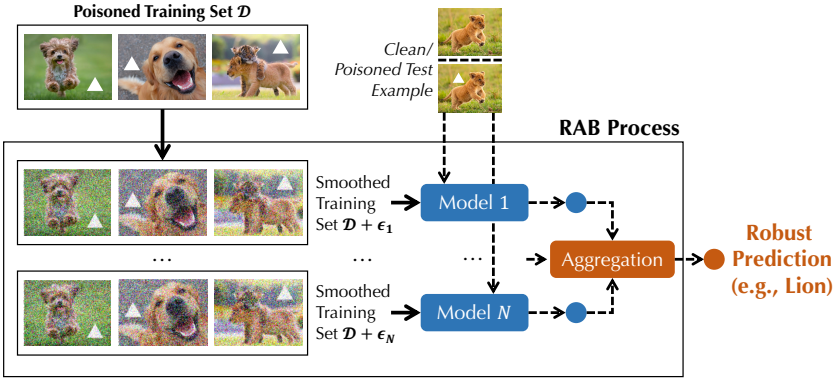


Figure 3: RAB robust training process. Given a poisoned training set $\mathcal{D}+\Delta$ and a training process $\mathcal{A}$ vulnerable to backdoor attacks, RAB generates N smoothed training sets $\{\mathcal{D}_i\}_{i\in[N]}$ and trains N different classifiers $\mathcal{A}_i$.

ability density function. Realizations of random variables are written in lowercase letters. For discrete random variables, we use lowercase letters to denote their probability mass function, e.g. $p(y)$ for distribution over labels. Feature vectors are taken to be d-dimensional real vectors $x \in \mathbb{R}^d$ and the set of labels y for a C-multiclass classification problem is given by $\mathcal{C} = \{1, \ldots, C\}$. A training set $\mathcal{D}$ consists of n (feature, label)-pairs $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. For a dataset $\mathcal{D}$ and a collection of n feature vectors $d = \{d_1, \ldots, d_n\}$, we write $\mathcal{D} + d$ to denote the set $\{(x_1 + d_1, y_1), \ldots, (x_n + d_n, y_n)\}$. We view a classifier as a deterministic function that takes as input a tuple with a test instance x and training set $\mathcal{D}$ and returns a class label $y \in \mathcal{C}$. Formally, given a dataset $\mathcal{D}$ and a test instance x, a classifier h learns a conditional probability distribution $p(y|x, \mathcal{D})$ over class labels and outputs the label which is deemed most likely under the learned distribution p:

$$h(x, \mathcal{D}) = \arg\max_y p(y|x, \mathcal{D}). \tag{57}$$

We omit the dependence on model parameters throughout this chapter and tacitly assume that the model is optimized based on the training dataset $\mathcal{D}$ via some optimization schemes such as stochastic gradient descent.

THREAT MODEL An adversary carries out a backdoor attack against a classifier h and a clean dataset $\mathcal{D} = \{(x_i, y_i)\}$. The attacker has in mind a target backdoor pattern Ω_x and a target class $\tilde{y}$ and the adversarial goal is to alter the dataset such that, given a clean test example x, adding the backdoor pattern to x (i.e., $x + \Omega_x$) will *alter* the classifier output $\tilde{y}$ with high probability. In general, the attack can replace r training instances (x_i, y_i) by backdoored instances $(x_i + \Omega_x, \tilde{y}_i)$. We remark that the attacker could embed distinct patterns to each instance and our result naturally extends to this case. Thus, summarizing the backdoor patterns as the collection $\Delta(\Omega_x) := \{\delta_1, \ldots, \delta_r, 0, \ldots, 0\}$, we formalize a backdoor attack as the transformation $(\mathcal{D}, \Omega_x, \tilde{y}) \to \mathcal{D}_{BD}(\Omega_x, \tilde{y})$ with

$$\mathcal{D}_{BD}(\Omega_x, \tilde{y}) = \{(x_i + \delta_i, \tilde{y}_i)\}_{i=1}^{r} \cup \{(x_i, y_i)\}_{i=r+1}^{n} \tag{58}$$

We often write $\mathcal{D}_{BD}(\Omega_x)$ instead of $\mathcal{D}_{BD}(\Omega_x, \tilde{y})$ when our focus is on the backdoor pattern Ω_x instead of the target class $\tilde{y}$. The backdoor attack succeeds on test example x whenever

$$h(x + \Omega_x, \mathcal{D}_{BD}(\Omega_x)) = \tilde{y} \tag{59}$$

CERTIFICATION GOAL One natural goal to defend against the above backdoor attack is to ensure that the prediction of $h(x + \Omega_x, \mathcal{D}_{BD}(\Omega_x))$ is *independent* of the backdoor patterns $\Delta(\Omega_x)$ which are present in the dataset, i.e.,

$$h(x + \Omega_x, \mathcal{D}_{BD}(\Omega_x)) = h(x + \Omega_x, \mathcal{D}_{BD}(\emptyset)) \tag{60}$$

where $\mathcal{D}_{BD}(\emptyset)$ is the dataset without any embedded backdoor patterns ($\delta_i = 0$). When this is true, the attacker obtained *no additional information* by knowing the pattern Ω_x embedded in the training set. That is to say, given a test instance which may contain a backdoor pattern, its prediction stays the same, independent of whether the models were trained with or without backdoors. We assume that the defender has full control of the training process. See Section 6.6 for more discussions on the assumptions and limitations of RAB.

ROBUSTNESS CERTIFICATION We aim to obtain a robustness bound R such that, whenever the sum of the magnitude of backdoors is below R, the prediction of the backdoored classifier is the same as when the classifier is trained on benign data. Formally, if $\mathcal{D}_{BD}(\Omega_x)$ denotes the backdoored training set, and $\mathcal{D}$ the training set containing clean features, we say that a classifier is *provably robust* whenever

$$\sqrt{\sum_{i=1}^{r} \|\delta_i\|_2^2} < R \tag{61}$$

implies that $h(x + \Omega_x, \mathcal{D}_{BD}(\Omega_x)) = h(x + \Omega_x, \mathcal{D}_{BD}(\emptyset))$. Our approach to obtaining the aforementioned robustness guarantee is based on randomized smoothing, which leads to the robust RAB training pipeline, as is illustrated in Figure 3. Given a clean dataset $\mathcal{D}$ and a backdoored dataset $\mathcal{D}_{BD}(\Omega_x)$, the goal of the defender is to make sure that the prediction on test instances embedded with the pattern Ω_x is the same as for models trained with $\mathcal{D}_{BD}(\emptyset)$.

Different from randomized smoothing-based certification against evasion attacks, here it is not enough to only smooth the test instances. Instead, in RAB, we will first add noise vectors, sampled from a smoothing distribution, to the given training instances, to obtain a collection of "smoothed" training sets. We subsequently train a model on each training set and aggregate their final outputs together as the final "smoothed" prediction. After this process, we show that it is possible to leverage the Neyman Pearson lemma to derive a robustness condition for this smoothed RAB training process. Additionally, the connection with the Neyman Pearson lemma also allows us to prove that the robustness bound is tight. Note that the RAB framework requires the training instances to be "smoothed" by a set of independent noises drawn from a certain distribution.

ADDITIONAL CHALLENGES We remark that, within this RAB training and certification process, there are several additional challenges. First, after adding noise to the training data, the clean accuracy of the trained classifier typically drops due to the distribution shift in the training data. To mitigate this problem, we add a deterministic value, based on the hash of the trained model, to test instances (Section 6.4), which minimizes the distribution shift and leads to improved accuracy scores. Second, considering different smoothing distributions for the training data, we provide rigorous analysis and a

robustness bound for both Gaussian and uniform smoothing distributions (Section 6.3). Third, we note that the proposed training process requires sampling a large number of randomly perturbed training sets. As this is computationally expensive, we propose an efficient polynomial-time algorithm for KNN classifiers (Section 6.4).

OUTLINE The remainder of this chapter is organized as follows. Section 6.2 presents the proposed general theoretical framework for certifying robustness against evasion and poisoning attacks, the tightness of the derived robustness bound, and sheds light on a connection between statistical hypothesis testing and certifiable robustness. Section 6.3 explains the proposed approach, RAB, for certifying robustness against backdoor attacks under the general framework with Gaussian and uniform noise distributions. Section 6.4 analyzes the robustness properties of DNN and KNN classifiers and presents algorithms to certify robustness for such models. Experimental results are presented in Section 6.5. Finally, Section 6.6 discusses the limitations of our work and concludes this chapter.

6.2 UNIFIED FRAMEWORK FOR CERTIFIED ROBUSTNESS

In this section, we propose a unified theoretical framework for certified robustness against evasion and poisoning attacks for classification models. Our framework is based on the intuition that randomizing the prediction or training process will "smoothen" the final prediction and therefore reduce the vulnerability to adversarial attacks. This principle has been successfully applied to certifying robustness against evasion attacks for classification models [39]. We first formally define the notion of a smoothed classifier where we extend upon previous work by randomizing *both* the test instance and the training set. We then leverage the Neyman Pearson lemma to derive a generic robustness condition in Theorem 2 and show that this robustness condition is tight.

SMOOTHED CLASSIFIERS On a high level, a smoothed classifier g is derived from a base classifier h by introducing additive noise to the input consisting of test and training instances. In a nutshell, the intuition behind randomized smoothing classifiers is that noise reduces the occurrence of regions with high curvature in the decision boundaries, resulting in reduced vulnerability to adversarial attacks. Recall that a classifier h, here serving as a base classifier, is defined as $h(x, \mathcal{D}) = \arg\max_y p(y|x, \mathcal{D})$ where p is learned from a dataset $\mathcal{D}$ and defines a conditional probability distribution over labels y. The final prediction is given by the most likely class under this learned distribution. A smoothed classifier is defined by

$$q(y|x, \mathcal{D}) = \mathbb{P}_{X,D}\left(h(x+X, \mathcal{D}+D) = y\right) \tag{62}$$

where we have introduced random variables $X \sim \mathbb{P}_X$ and $D \sim \mathbb{P}_D$ which act as smoothing distributions and are assumed to be independent. We emphasize that D is a collection of n independent and identically distributed random variables $D^{(i)}$, each of which is added to a training instance in $\mathcal{D}$. The final, smoothed classifier then assigns the most likely class to an instance x under this new, "smoothed" model q, so that

$$g(x, \mathcal{D}) = \arg\max_y q(y|x, \mathcal{D}). \tag{63}$$

Within this formulation of a smoothed classifier, we can also model randomized smoothing for defending against evasion attacks by setting the training set noise to be zero, i.e.

$D \equiv 0$. We emphasize at this point that the smoothed classifier g implicitly depends on the choice of noise distributions $\mathbb{P}_X$ and $\mathbb{P}_D$. In Section 6.3 we instantiate this classifier with Gaussian noise and with uniform noise and show how this leads to different robustness bounds.

A GENERAL CONDITION FOR PROVABLE ROBUSTNESS We now derive a tight robustness condition by drawing a connection between statistical hypothesis testing and the robustness of classification models subject to adversarial attacks. We allow adversaries to conduct an attack on either *(i) the test instance* x, *(ii) the training set* $\mathcal{D}$ or *(iii) a combined attack on test and training set*. The resulting robustness condition is of a general nature and is expressed in terms of the optimal type II errors for likelihood ratio tests. We remark that this theorem is a more general version of the result presented in [39], by extending it to general smoothing distributions and smoothing on the training set. In Section 6.3 we will show how this result can be used to obtain robustness bound in terms of L_p-norm bounded backdoor attacks. We show that smoothing on the training set makes it possible for certifying the robustness against backdoors, and the general smoothing distribution allows us to explore the robustness bound certified by different smoothing distributions.

Theorem 2. *Let* q *be the smoothed classifier as in* (62) *with smoothing distribution* $Z := (X, D)$ *with* X *taking values in* $\mathbb{R}^d$ *and* D *being a collection of* n *independent* $\mathbb{R}^d$*-valued random variables,* $D = (D^{(1)}, \ldots, D^{(n)})$. *Let* $\Omega_x \in \mathbb{R}^d$ *and let* $\Delta := (\delta_1, \ldots, \delta_n)$ *for backdoor patterns* $\delta_i \in \mathbb{R}^d$. *Let* $y_A \in \mathcal{C}$ *and let* $p_A, p_B \in [0, 1]$ *such that* $y_A = g(x, \mathcal{D})$ *and*

$$q(y_A | x, \mathcal{D}) \geqslant p_A > p_B \geqslant \max_{y \neq y_A} q(y | x, \mathcal{D}). \tag{64}$$

If the optimal type II errors, for testing the null $Z \sim \mathbb{P}_0$ *against the alternative* $Z + (\Omega_x, \Delta) \sim \mathbb{P}_1$, *satisfy*

$$\beta^*(1 - p_A; \mathbb{P}_0, \mathbb{P}_1) + \beta^*(p_B; \mathbb{P}_0, \mathbb{P}_1) > 1, \tag{65}$$

then it is guaranteed that $y_A = \arg\max_y q(y | x + \Omega_x, \mathcal{D} + \Delta)$.

Proof. The proof of this result is a direct consequence of Lemma 1 applied to the distributions $\mathbb{P}_0$ and $\mathbb{P}_1$. To see this, note that

$$\begin{aligned} q(y_A | x, \mathcal{D}) &= \mathbb{P}_{X,D}\left[h(x + X, \mathcal{D} + D) = y_A\right] \\ &= \mathbb{E}_{X,D}\left[\mathbb{1}\{h(x + X, \mathcal{D} + D) = y_A\}\right] \\ &\geqslant p_A. \end{aligned} \tag{66}$$

Thus, applying Lemma 1 to the function $(X, \mathcal{D}) \mapsto \mathbb{1}\{h(x + X, \mathcal{D} + D) = y_A\}$ and setting $\underline{m} = p_A$ yields

$$\begin{aligned} q(y_A | x + \Omega_x, \mathcal{D} + \Delta) &= \mathbb{E}_{X,D}\left[\mathbb{1}\{h(x + \Omega_x + X, \mathcal{D} + \Delta + D) = y_A\}\right] \\ &\geqslant \beta^*(1 - p_A; \mathbb{P}_0, \mathbb{P}_q) \end{aligned} \tag{67}$$

Similarly, applying Lemma 1 to the function $(X, \mathcal{D}) \mapsto \mathbb{1}\{h(x + X, \mathcal{D} + D) = y\}$, for any $y \neq y_A$, and setting $\underline{m} = p_B$, yields

$$\begin{aligned} q(y | x + \Omega_x, \mathcal{D} + \Delta) &= \mathbb{E}_{X,D}\left[\mathbb{1}\{h(x + \Omega_x + X, \mathcal{D} + \Delta + D) = y_B\}\right] \\ &\leqslant 1 - \beta^*(p_B; \mathbb{P}_0, \mathbb{P}_1). \end{aligned} \tag{68}$$

It follows that whenever

$$\beta^*(p_B; \mathbb{P}_0, \mathbb{P}_1) + \beta^*(1 - p_A; \mathbb{P}_0, \mathbb{P}_q) > 1, \tag{69}$$

it is guaranteed that $y_A = \arg\max_y q(y|x + \Omega_x, \mathcal{D} + \Delta)$. This concludes the proof. □

We now make some observations about Theorem 2 to get a better intuition on the robustness condition (65):

- Different smoothing distributions lead to robustness bounds in terms of different norms. For example, Gaussian noise yields a robustness bound in L_2 norm while Uniform noise leads to other L_p norms.
- The robustness condition (65) does not make any assumption on the underlying classifier other than on the class probabilities predicted by its smoothed version.
- The random variable $Z + (\Omega_x, \Delta)$ models a general adversarial attack including evasion and backdoor attacks.
- If no attack is present, i.e., if $(\Omega_x, \Delta) = (0, 0)$, then we get the trivial condition $p_A > p_B$.
- As p_A increases, the optimal type II error increases for the given backdoor (Ω_x, Δ). Thus, in the simplified setup where $p_A + p_B = 1$ and the robustness condition reads $\beta^*(1 - p_A; \mathbb{P}_0, \mathbb{P}_1) > 1/2$, the distribution shift caused by (Ω_x, Δ) can increase. Thus, as the smoothed classifier becomes more confident, the robust region becomes larger.

While the generality of Theorem 2 allows us to model a multitude of threat models, it bears the challenge of how one should instantiate this theorem such that it is applicable to defend against a specific adversarial attack. In addition to the flexibility with regard to the underlying threat model, we are also provided with flexibility regarding the smoothing distributions, resulting in different robustness guarantees. This again begs the question, which smoothing distributions result in useful robustness bounds. In Section 6.3, we will show how this theorem can be applied to obtain the robustness guarantee against backdoor attacks described in Section 6.1.5.

Next, we show that our robustness condition is tight in the following sense. If (64) is all that is known about the smoothed classifier g, then there is no perturbation (Ω_x, Δ) that violates (65). On the other hand, if (65) is violated, then we can always construct a smoothed classifier g^* such that it satisfies the class probabilities (64) but is not robust against this perturbation. We provide the proof of this result in Section 6.7.1.

Theorem 3. *Suppose that* $1 \geqslant p_A + p_B \geqslant 1 - (C - 2) \cdot p_B$ *where* C *is the number of classes. If the adversarial perturbations* (Ω_x, Δ) *violate* (65)*, then there exists a base classifier* h^* *such that the smoothed classifer* g^* *is consistent with the class probabilities* (64) *and for which* $g^*(x + \Omega_x, \mathcal{D} + \Delta) \neq y_A$.

6.3 PROVABLE ROBUSTNESS AGAINST BACKDOOR ATTACKS

It is not straightforward to use the result from Theorem 2 to get a robustness certificate against backdoor attacks in terms of L_p-norm bounded backdoor patterns. In this section, we aim to answer the question: *how can we instantiate this result to obtain robustness*

guarantees against backdoor attacks? In particular, we show that by leveraging Theorem 2, we obtain the robustness guarantee defined in Section 6.1.5. To that end, we derive robustness bounds for smoothing with isotropic Gaussian noise and we also illustrate how to derive certification bounds using other smoothing distributions. Since isotropic Gaussian noise leads to a better radius, we will use this distribution in our experiments as a demonstration.

6.3.1 *Method Outline*

Suppose that we are given a base classifier that has been trained on a *backdoored* dataset that contains r training samples which are infected with backdoor patterns $\Delta(\Omega_x)$. Our goal is to derive a condition on the backdoor patterns $\Delta(\Omega_x)$ such that the prediction for $x + \Omega_x$ with a classifier trained on the backdoored dataset $\mathcal{D}_{BD}(\Delta(\Omega_x))$ is the same as the prediction (on the same input) that a smoothed classifier would have made, had it been trained on a dataset without the backdoor triggers, $\mathcal{D}_{BD}(\emptyset)$. In other words, we obtain the guarantee that *an attacker can not achieve their goal of systematically leading the test instance with the backdoor pattern to the adversarial target*, meaning they will always obtain the same prediction as long as the added pattern δ satisfies the robustness conditions (i. e., bounded magnitude).

6.3.1.1 *Gaussian Smoothing*

We obtain this certificate by instantiating Theorem 2 in the following way. Suppose an attacker injects backdoor patterns $\Delta(\Omega_x) = \{\delta_1, \ldots, \delta_r\} \subset \mathbb{R}^d$ to $r \leqslant n$ training instances of the training set $\mathcal{D}$, yielding the backdoored training set $\mathcal{D}_{BD}(\Delta(\Omega_x))$. We then train the base classifier on this poisoned dataset, augmented with additional noise on the feature vectors $\mathcal{D}_{BD}(\Delta(\Omega_x)) + D$, where D is the smoothing noise added to the training features. We obtain a prediction of the smoothed classifier g by taking the expectation with respect to the distribution of the smoothing noise D. Suppose that the smoothed classifier obtained in this way predicts a malicious instance $x + \Omega_x$ to be of a certain class with probability at least p_A and the runner-up class with probability at most p_B. Our result tells us that, as long as the introduced patterns satisfy condition (65), we get the guarantee that the malicious test input would have been classified equally as when the classifier had been trained on the dataset with clean features $\mathcal{D}_{BD}(\emptyset)$. In the case where the noise variables are isotropic Gaussians with standard deviation σ, the condition (65) yields a robustness bound in terms of the sum of L_2-norms of the backdoors.

Corollary 1 (Gaussian Smoothing). *Let $\Delta = (\delta_1, \ldots, \delta_n)$ and Ω_x be $\mathbb{R}^d$-valued backdoor patterns and let $\mathcal{D}$ be a training set. Suppose that for each i, the smoothing noise on the training features is $D^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$. Let $y_A \in \mathcal{C}$ such that $y_A = g(x + \Omega_x, \mathcal{D} + \Delta)$ with class probabilities satisfying*

$$q(y_A | x + \Omega_x, \mathcal{D} + \Delta) \geqslant p_A > p_B \geqslant \max_{y \neq y_A} q(y | x + \Omega_x, \mathcal{D} + \Delta). \tag{70}$$

Then, if the backdoor patterns are bounded by

$$\sqrt{\sum_{i=1}^{n} \|\delta_i\|_2^2} < \frac{\sigma}{2} \left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B) \right), \tag{71}$$

it is guaranteed $y_A = g(x + \Omega_x, \mathcal{D}) = g(x + \Omega_x, \mathcal{D} + \Delta)$.

This result shows that, whenever the norms of the backdoor patterns are below a certain value, we obtain the guarantee that the classifier makes the same prediction on the test data with backdoors as it does when trained without embedded patterns in the training set. We can further simplify the robustness bound in (71) if we can assume that an attacker poisons at most $r \leqslant n$ training instances with one single pattern δ. In this case, the bound (71) is given by

$$\|\delta\|_2 < \frac{\sigma}{2\sqrt{r}} \left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right). \tag{72}$$

We see that, as we know more about the capabilities of an attacker and the nature of the backdoor patterns, we are able to certify a larger robustness radius, proportional to $1/\sqrt{r}$.

6.3.2 *Other Smoothing Distributions*

Given the generality of our framework, it is possible to derive certification bounds using other smoothing distributions. However, different smoothing distributions lead to vastly different performance and a comparative study among different smoothing distributions is interesting future work. Here, we will illustrate one example of smoothing using a uniform distribution.

Corollary 2 (Uniform Smoothing). *Let* $\Delta = (\delta_1, \ldots, \delta_n)$ *and* Ω_x *be* $\mathbb{R}^d$ *valued backdoor patterns and let* $\mathcal{D}$ *be a training set. Suppose that for each* i*, the smoothing noise on the training features is* $D^{(i)} \overset{iid}{\sim} \mathcal{U}([a, b])$*. Let* $y_A \in \mathcal{C}$ *such that* $y_A = g(x + \Omega_x, \mathcal{D} + \Delta)$ *with class probabilities satisfying*

$$q(y_A | x + \Omega_x, \mathcal{D} + \Delta) \geqslant p_A > p_B \geqslant \max_{y \neq y_A} q(y | x + \Omega_x, \mathcal{D} + \Delta). \tag{73}$$

Then, if the backdoor patterns satisfy

$$1 - \left(\frac{p_A - p_B}{2}\right) < \prod_{i=1}^{n} \left(\prod_{j=1}^{d} \left(1 - \frac{|\delta_{i,j}|}{b - a}\right)_+ \right) \tag{74}$$

where $(x)_+ = \max\{x, 0\}$*, it is guaranteed that* $y_A = g(x + \Omega_x, \mathcal{D}) = g(x + \Omega_x, \mathcal{D} + \Delta)$.

As in the Gaussian case, the robustness bound in (74) can again be simplified in a similar fashion, if we assume that an attacker poisons at most $r \leqslant n$ training instances with one single pattern δ. In this case, the bound (74) is given by

$$1 - \left(\frac{p_A - p_B}{2}\right) < \left(\prod_{j=1}^{d} \left(1 - \frac{|\delta_j|}{b - a}\right)_+ \right)^r. \tag{75}$$

We see again that, as the number of infected training samples r gets smaller, this corresponds to a larger bound since the RHS of (75) gets larger. In other words, if we know that the attacker injects fewer backdoors, then we can certify a backdoor pattern with a larger magnitude.

Algorithm 1 DNN-RAB for training certifiably robust DNNs.

Require: Poisoned training dataset $\mathcal{D} = \{(x_i + \delta_i, \tilde{y}_i)_{i=1}^n\}$, noise scale σ, model number N

1: **for** $k = 1, \ldots, N$ **do**
2: Sample $\epsilon_{k,1}, \ldots, \epsilon_{k,n} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$.
3: $\mathcal{D}_k = \{(x_i + \delta_i + \epsilon_{k,i}, \tilde{y}_i)_{i=1}^n\}$.
4: $h_k = \texttt{train\_model}(\mathcal{D}_k)$.
5: Sample u_k from $\mathcal{N}(0, \sigma^2 \mathbb{1}_d)$ deterministically with random seed based on $\texttt{hash}(h_k)$.
6: **end for**
7: **return** Model collection $\{(h_1, u_1), \ldots, (h_N, u_N)\}$

DISCUSSION. We emphasize that we focus on protecting the system against attackers who aim to *trigger* a targeted error with a specific *backdoor pattern*. The system can still be vulnerable to other types of *poisoning attacks*. One such example is the label flipping attack, in which one flips the labels of a subset of examples while keeping the features untouched. Interestingly, one concurrent work explored the possibility of using randomized smoothing to defend against label flipping attacks [184]. Developing a single framework to be robust against both backdoor and label flipping attacks is an exciting future direction, and we expect it to require nontrivial extensions of both approaches to achieve non-trivial certified accuracy. Furthermore, while we focus the experiments on Gaussian smoothing and L_2-norm guarantees, it is in principle possible to certify other L_p-norms with different smoothing distributions. Indeed, for evasion attacks, in Chapter 7, we use exponential smoothing noise with certificates in L_1-norm.

6.4 INSTANTIATING THE FRAMEWORK WITH SPECIFIC ML MODELS

In the preceding sections, we presented our approach to certifying robustness against backdoor attacks. Here, we will analyze and provide detailed algorithms for the RAB training pipeline for two types of ML models: deep neural networks and K-nearest neighbor classifiers. The success of backdoor poisoning attacks against DNNs has caused a lot of attention recently. Thus, we first aim to evaluate and certify the robustness of DNNs against backdoor attacks. Secondly, given the fact that KNN models have been widely applied in different applications, either based on raw data or on trained embeddings, it is of great interest to know about the robustness for this type of ML models. Specifically, we are inspired by a result from [109] and develop an *exact* efficient smoothing algorithm for KNN models, such that we do not need to draw a large number of random samples from the smoothing distribution for these models. This makes our approach considerably more practical for KNNs since it avoids the expensive training of a large number of models, as is required with generic classification algorithms including DNNs.

6.4.1 *Deep Neural Networks*

In this section, we consider smoothed models which use DNNs as base classifiers. For a given test input x_{test}, the goal is to calculate the prediction of g on $(x_{test}, \mathcal{D} + \Delta)$

Algorithm 2 Certified inference with RAB-trained models.

Require: Test sample x, noise scale σ, models $\{(h_k, u_k)\}_{k=1}^{N}$, backdoor magnitude $\|\delta\|_2$, number of poisoned training samples r

counts = $|\{k\colon\ h_k(x+u_k,\ \mathcal{D}+\epsilon_k) = y\}|$ for $y = 1, \ldots, C$
y_A, y_B = top two indices in counts
n_A, n_B = counts[y_A], counts[y_B]
p_A, p_B = calculate_bound(n_A, n_B, N, α).
if $p_A > p_B$ **then**
  $R = \frac{\sigma}{2\sqrt{r}}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right)$
  if $R \geqslant \|\delta\|_2$ **then**
    return prediction y_A, robust radius R.
  end if
end if
return ABSTAIN

according to Corollary 1 and the corresponding certified bound given in the right hand side of Eq. (71). In the following, we first describe the training process and then the inference algorithm.

6.4.1.1 *RAB Training for DNNs*

First, we draw N samples $d_1, \ldots, d_N$ from the distribution of

$$D \sim \prod_{i=1}^{n} \mathcal{N}(0,\ \sigma^2 \mathbb{1}_d). \tag{76}$$

Given the N samples of training noise (each consisting of $|\mathcal{D}| = n$ noise vectors), we train N DNN models on the datasets $\mathcal{D} + d_k$ for $k = 1, \ldots, N$ and obtain classifiers $h_1, \ldots, h_N$. Along with each model h_k, we draw a random noise u_k from $\mathcal{N}(0,\ \sigma^2 \mathbb{1}_d)$ with a random seed based on the hash of the trained model file. This noise vector is stored along with the model parameters and added to each test input during inference. The reason for this is that, empirically, we observed that inputting test samples without this additional augmentation leads to poor prediction performance since the ensemble of models $\{h_1, \ldots, h_N\}$ has to classify an input that has not been perturbed by Gaussian noise, while it has only "seen" noisy samples, leading to a mismatch between training and test distributions. Algorithm 1 shows the pseudocode describing RAB-training for DNN models.

6.4.1.2 *Inference*

To get the prediction of the smoothed classifier on a test sample x_{test} we first compute the empirical majority vote as an unbiased estimate

$$\hat{q}(y|x,\ \mathcal{D}) = \frac{|\{k\colon\ h_k(x_{test}+u_k,\ \mathcal{D}+d_k) = y\}|}{N} \tag{77}$$

of the class probabilities and where u_k is the (model-) deterministic noise vector sampled during training in Algorithm 1. Second, for a given error tolerance α, we compute p_A and p_B using one-sided $(1-\alpha)$ lower confidence intervals for the binomial distribution with parameters n_A and n_B and N samples. Finally, we invoke Corollary 1 and first

compute the robust radius according to Eq. (72), based on p_A, p_B the smoothing noise parameter σ and the number of poisoned training samples r. If the resulting radius R is larger than the magnitude of the backdoor samples δ, the prediction is certified, i.e. the backdoor attack has failed on this particular sample. Algorithm 2 shows the pseudocode for the DNN inference with RAB.

6.4.1.3 *Model-deterministic Test-time Augmentation*

One caveat in directly applying Equation (77) is the mismatch of the training and test distribution — during training, all examples are perturbed with sampled noise, whereas the test example is without noise. In practice, we see that this mismatch significantly decreases the test accuracy. One natural idea is to also add noise to the test examples, however, this requires careful design (e.g., simply drawing k independent noise vectors and applying them to (77) will lead to a less powerful bound). We thus modify the inference function given a learned model h_k in the following way. Instead of directly classifying an unperturbed input x_{test}, we use the hash value of the trained h_k model parameters as the random seed and sample $u_k \sim \mathcal{N}_{hash(h_k)}(0, \sigma^2 \mathbb{1}_d)$. In practice, we use SHA256 hashing[249] of the trained model file. In this way, the noise we add is a deterministic function of the trained model, which is equivalent to altering the inference function in a deterministic way, $\tilde{h}_k(x_{test}) = h_k(x_{test} + u_k)$. We show in the experiments that this leads to significantly better prediction performance in practice. Note that the reason for using a hash function instead of random sampling every time is to ensure that the noise generation process is deterministic, so the choice of different hash functions is flexible.

6.4.2 *K-Nearest Neighbors*

If the base classifier h is a K-nearest neighbor classifier, we can evaluate the corresponding smoothed classifier *exactly* and efficiently, in polynomial time, if the smoothing noise is drawn from a Gaussian distribution. In other words, for this type of model, we can eliminate the need to approximate the expectation value via Monte Carlo sampling and evaluate the classifier exactly. Finally, it is worth remarking that bypassing the need to do Monte Carlo sampling ultimately results in a considerable speed-up as it avoids the expensive training of independent models as is required for generic models including DNNs.

A KNN classifier works in the following way: Given a training set $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\}$ and a test example x, we first calculate the similarity between x and each x_i, $s_i := \kappa(x_i, x)$ where κ is a similarity function. Given all these similarity scores $\{s_i\}_i$, we choose the K most similar training examples with the largest similarity score $\{x_{\sigma_i}\}_{i=1}^K$ along with corresponding labels $\{y_{\sigma_i}\}_{i=1}^K$. The final prediction is made according to a majority vote among the top-K labels.

Similar to DNNs, we obtain a smoothed KNN classifier by adding Gaussian noise to training points and evaluate the expectation with respect to this noise distribution

$$q(y|x, \mathcal{D}) = \mathbb{P}\left(\mathrm{KNN}(x, \mathcal{D} + D) = y\right) \tag{78}$$

where $D = (D^{(1)}, \ldots, D^{(n)}) \sim \prod_{i=1}^n \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$. The next theorem shows that (78) can be computed exactly and efficiently when we measure the similarity with euclidean distance quantized into finite number similarity of levels.

Theorem 4. *Given n training instances, a C-multiclass KNN classifier based on quantized euclidean distance with L similarity levels, smoothed with isotropic Gaussian noise can be evaluated exactly with complexity $\mathcal{O}(K^{2+C} \cdot n^2 \cdot L \cdot C)$.*

Proof (sketch). The first step to computing (78) is to notice that we can summarize all possible arrangements $\{x_{\sigma_i} + D^{(\sigma_i)}\}_{i=1}^{K}$ of top-K instances leading to a prediction by using tally vectors $\gamma \in [K]^C$. A tally vector has as its k-th element the number of instances in the top-K with label k, $\gamma_k = \#\{y_{\sigma_i} = k\}$. In the second step, we partition the event that a tally vector γ occurs into events where an instance i with similarity β is in the top-K but would not be in the top-$(K-1)$. These first two steps result in a summation over $\mathcal{O}(K^C \cdot n \cdot L \cdot C)$ terms. In the last step, we compute the probabilities of the events $\{\text{tally}\,\gamma \wedge \kappa(x_i + D^{(i)}, x) = \beta\}$ with dynamic programming in $\mathcal{O}(n \cdot K^2)$ steps, resulting in a final time complexity of $\mathcal{O}(K^{2+C} \cdot n^2 \cdot L \cdot C)$. □

If $K = 1$, an efficient algorithm can even achieve time complexity linear in the number of training samples n. We refer the reader to Section 6.7.4 for details and the algorithm.

6.5 EXPERIMENTS

In this section, we present an extensive experimental evaluation of our approach and provide a benchmark for certified robustness for DNN and KNN classifiers on different datasets. In addition, we consider three different types of backdoor attack patterns, namely one-pixel, four-pixel, and blending-based attacks. The attack patterns are illustrated in Figure 4 which shows that these patterns can be hard to spot by a human, in particular for the one-pixel pattern on high-resolution images. At a high level, our experiments reveal the following set of observations: [1]

- RAB is able to achieve comparable robustness on benign instances compared with vanilla trained models, and achieves non-trivial *certified accuracy* under a range of realistic backdoor attack settings.
- There is a gap between the certified accuracy provided by RAB and empirical robust accuracy achieved by the state-of-the-art empirical defenses against backdoor attacks without formal guarantees, which serves as the upper bound of the certified accuracy; however, such a gap is reasonably small and we are optimistic that future research can further close this gap.
- RAB's efficient KNN algorithm provides a very effective solution for tabular data.
- Simply applying randomized smoothing to RAB is not effective and careful optimizations (e.g., deterministic test-time augmentation) are necessary.

6.5.1 *Experiment Setup*

We follow the popular transfer learning setting for poisoning attacks [76, 186, 193, 196, 280] in our experiments, specifically [194]. We first use models initialized with pre-trained weights obtained from a clean dataset, and then finetune the model with a subset of training data containing backdoored instances. Preliminary experiments and

1 Our code is available at https://github.com/AI-secure/Robustness-Against-Backdoor-Attacks

existing work [238] showed that it is difficult to successfully inject backdoors if only a subset of parameters is finetuned. As a result, we always finetune the entire set of model parameters.

DATASETS AND MODEL We consider four different datasets, namely the MNIST dataset [127] consisting of 60,000 images of handwritten digits from 0-9, the CIFAR-10 dataset[120] which includes 50,000 images of 10 different classes of natural objects such as horse, airplane, automobile, etc. Furthermore, we perform evaluations on the high-resolution ImageNette dataset [99] which is a 10-class subset of the original large-scale ImageNet dataset [47]. Finally, we evaluate the KNN model on a tabular dataset, namely the UCI Spambase dataset [51], which consists of bag-of-words feature vectors on E-mails and determines whether the message is spam or not. The dataset contains 4,601 data cases, each of which is a 57-dimensional input. We use 0.1% of the MNIST and CIFAR-10 training data to finetune our models; on ImageNette and Spambase, we use 1% for finetuning. For evaluations on DNNs, we choose the CNN architecture from [75] on MNIST and the ResNet used in [39] on CIFAR-10, whereas for ImageNette, we use the standard ResNet-18 [85] architecture.

TRAINING PROTOCOL We set the number of sampled noise vectors (i.e. augmented datasets) to $N = 1,000$ on MNIST and CIFAR, and $N = 200$ on ImageNette, leading to an ensemble of 1,000 and 200 models, respectively. The added smoothing noise is sampled from the Gaussian distribution with location parameter $\mu = 0$ and scale $\sigma = 0.5$ for MNIST and Spambase. For CIFAR-10 and ImageNette we use $\mu = 0$ and set the scale to $\sigma = 0.2$. The impact of different σ is shown in Section 6.5.2.4. The confidence intervals for the binomial distribution are calculated with an error rate of $\alpha = 0.001$. For the KNN models, we use $K = 3$ neighbors and set the number of similarity levels to $L = 200$, meaning that the similarity scores according to euclidean distance are quantized into 200 distinct levels.

BASELINES OF EMPIRICAL BACKDOOR REMOVAL BASED DEFENSES Since this is the first work providing rigorous certified robustness against backdoor attacks, there is no baseline that allows a fair comparison of the certified accuracy. We remark that a technical report [233] directly applies the randomized smoothing technique to certify robustness against backdoors without evaluation or analysis. However, as we will show in Section 6.5.2.5, directly applying randomized smoothing without deterministic test-time augmentation does not provide high certified robustness. We will, on the other hand, compare our empirical robust accuracy with the state-of-the-art empirical defenses. We briefly review these defenses in the following.

Activation clustering (AC) [32] extracts the activation of the last hidden layer of a trained model and uses clustering analysis to remove training instances with anomalies. We use the default parameter setting provided in the Adversarial Robustness Toolbox (ART) [165]. Spectral Signature (Spectral) [223] uses matrix decomposition on the feature representations to detect and remove training instances with anomalies. We again use the default parameter setting provided in ART. Sphere [207] performs dimensionality reduction and removes instances with anomalies in the lower dimensions. The top-15% anomaly instances are removed. Neural Cleanse (NC) [234] first reverse-engineers a "pseudo-trigger" for each class. Then, to detect and remove anomaly instances, the distances between each instance with and without the pseudo-trigger are

compared, and the most similar ones are recognized as anomaly instances. We use pixel-level distance as the distance metric, 100 epochs for trigger generation, and an initial $\lambda = 0.01$ for MNIST and $\lambda = 0.0001$ for CIFAR and ImageNette. Statistical Contamination Analyzer (SCAn) [215] first performs an expectation minimization algorithm to decompose two subgroups over a small clean dataset. Then, for each class in the train set, the parameters of a mixture model for all the data are estimated, before we calculate the likelihood for anomaly detection. To identify the backdoored instances, we recognize the smaller set in the most anomalous class as the backdoored instances. For Mixup [18], following their data augmentation technique, we use a 4-way mixup training algorithm to train the model over the train set. The convex coefficients are drawn from a Dirichlet distribution with $\alpha = 1.0$.

The initial goal of all these approaches, with the exception of Mixup, is to **detect** backdoored instances, i.e., to determine whether there exists a trigger. To apply them as a defense (i.e., to train a clean model despite the existence of backdoored data), we make adaptations either following the original paper (AC, Spectral, Sphere and NC) or by our design (SCAn) so that we remove training data with anomalies detected by these approaches and retrain a clean model. Some detection cannot be adapted to the defense task, such as [262], and are not included in the comparison.

EVALUATION METRICS We evaluate the model accuracy trained on the backdoored dataset with vanilla training and RAB training strategies. In particular, we evaluate both the model performance on benign instances (benign accuracy) and backdoored instances for which the attack was successful against the vanilla model (empirical robust accuracy). With RAB, we are also able to calculate the certified accuracy, which means that the RAB model not only certifies that the prediction is the same as if it were trained on the clean dataset, but also that the prediction is equal to the ground truth. The certified accuracy is defined below.

$$\text{Certified Acc.} = \frac{1}{n} |\{x_i : R_i > \|\delta\|_2 \wedge \hat{y}_i = y_i\}| \tag{79}$$

where R_i is the robus radius according to (71), $\hat{y}_i$ is the predicted label, and y_i is the ground truth for input x_i.

We emphasize that we only evaluate the backdoored test instances for which the attack is successful against the vanilla trained models, which is why the vanilla models always have 0% empirical robust accuracy on these backdoored instances in Table 1. This is to evaluate against the effective backdoor attacks and better illustrate the comparison between RAB-trained models with vanilla and baseline backdoor defense models (empirical robust accuracy). Such empirical robust accuracy of different methods serves as an upper bound for the certified accuracy.

BACKDOOR PATTERNS We evaluate RAB against three representative backdoor attacks, namely a one-pixel pattern in the middle of the image, a four-pixel pattern, and blending a random, but fixed, noise pattern to the entire image [34]. We visualize all backdoor patterns on different datasets in Figure 4. We control the perturbation magnitude of the attack via the L_2-norm of the backdoor patterns, setting $\|\delta\|_2 = 0.1$ for all attacks where δ is the backdoor pattern. On MNIST, we inject 10% backdoored instances and 5% for CIFAR and ImageNette respectively. If not described differently, the attack goal is to fool the model into predicting `0` on MNIST, `airplane` on CIFAR and

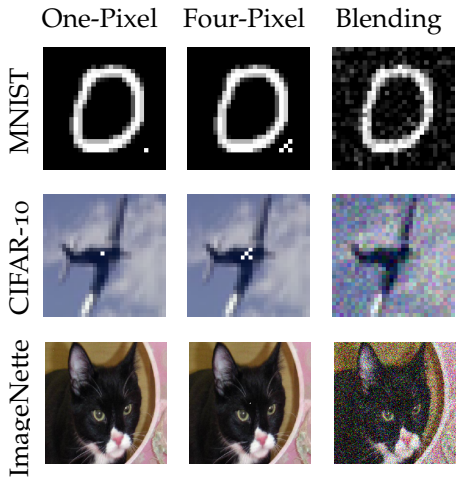


Figure 4: Illustrations of the applied backdoor patterns.

Table 1: Evaluation on DNNs with different datasets. We use $\sigma = 0.5$ for MNIST and $\sigma = 0.2$ for CIFAR-10 and ImageNette. Vanilla denotes DNNs without RAB training and RAB-cert is the certified accuracy of RAB. The highest empirical robust accuracies are **bolded**. The robust accuracy scores are evaluated only on successfully backdoored instances.

	Backdoor Pattern	Acc. on Benign Instances		Empirical Robust Acc.								**Certified Robust Acc.**
		Vanilla	RAB	Vanilla	RAB	AC [32]	Spectral [223]	Sphere [207]	NC [234]	SCAn [215]	Mixup [18]	RAB-cert
MNIST	One-pixel	92.7%	92.6%	0%	41.2%	64.3%	3.4%	3.1%	**76.2%**	45.6%	34.5%	23.5%
	Four-pixel	92.7%	92.6%	0%	40.7%	56.9%	2.8%	2.1%	**79.9%**	45.4%	33.2%	24.1%
	Blending	92.9%	92.6%	0%	39.6%	**63.6%**	3.0%	1.8%	63.0%	44.7%	28.3%	23.1%
CIFAR-10	One-pixel	59.9%	56.7%	0%	**42.9%**	31.4%	31.2%	16.5%	15.7%	12.9%	26.5%	24.5%
	Four-pixel	59.4%	56.8%	0%	**42.8%**	28.9%	31.4%	15.0%	16.8%	16.5%	31.8%	24.1%
	Blending	60.5%	56.8%	0%	**42.8%**	27.4%	28.0%	16.5%	16.6%	15.8%	30.0%	24.1%
ImageNette	One-pixel	93.0%	91.6%	0%	38.6%	44.7%	47.8%	29.6%	**69.9%**	35.2%	55.1%	15.9%
	Four-pixel	93.7%	91.5%	0%	38.4%	54.2%	52.8%	42.1%	**67.9%**	49.7%	51.6%	12.6%
	Blending	94.8%	91.8%	0%	29.9%	46.3%	18.4%	31.0%	**66.7%**	33.3%	56.3%	9.2%

`tench` on ImageNette. In Section C.1, we also consider an all-to-all attack goal [75] so that the fooled model will change its prediction conditioned on the original label.

It is possible to use different backdoor patterns via optimization and other approaches. However, since our goal is to provide *certified* robustness against backdoor attacks, a task that is by definition agnostic to the specific backdoor pattern but only depends on the magnitude of the pattern and the number of backdoored training instances, we mainly focus on these representative backdoor patterns. In addition, we only evaluate the backdoor attack to poison the dataset, while other attacks that interfere with the training process are not evaluated [180], as RAB is a robust training pipeline against training data manipulation based poisoning attacks.

6.5.2 *Deep Neural Networks*

In this section we evaluate RAB against backdoor attacks on different models and datasets. We present both the certified robust accuracy of RAB, as well we the empirical robust accuracy comparison between RAB and baseline defenses. Furthermore, we also present several ablation studies to further explore the properties of RAB.

Table 2: Evaluation on KNNs with $K = 3$ on the UCI Spambase tabular dataset. We use $\sigma = 0.5$ for Spam. Vanilla denotes DNNs without RAB training and RAB-cert is the certified accuracy of RAB. The highest empirical robust accuracies are **bolded**. The robust accuracy scores are evaluated only on successfully backdoored instances.

	Backdoor Pattern	Accuracy on Benign Instances		Empirical Robust Acc.						**Certified Robust Acc.**
		Vanilla	RAB	Vanilla	RAB	AC [32]	Spectral [223]	Sphere [207]	SCAn [215]	RAB-cert
UCI Spambase	One-pixel	98.7%	98.4%	0%	**54.6%**	9.0%	9.6%	2.4%	10.5%	36.4%
	Four-pixel	98.7%	98.4%	0%	**50.0%**	9.6%	9.6%	3.0%	11.2%	33.3%
	Blending	98.7%	98.4%	0%	**58.3%**	8.1%	8.1%	1.7%	9.9%	41.7%

6.5.2.1 *Certified Robustness with RAB*

We first evaluate the certified robustness of RAB on DNNs against different backdoor patterns on different datasets. We also present the performance of RAB on benign instances and backdoored instances empirically. Table 1 lists the benchmark results on MNIST, CIFAR-10, and ImageNette, respectively. From the results, we can see that RAB achieves significantly non-trivial certified robust accuracy against backdoor attacks at a negligible cost of benign accuracy; while there are no certified results for any other method. The slight drop in benign accuracy results from training on noisy instances. However, this loss in benign accuracy is less than 3% in most cases and is clearly outweighed by the achieved certified robust accuracy. In particular, RAB achieves over 23% *certified accuracy* on the backdoored instances for MNIST and CIFAR-10, and around 12% for ImageNette. In other words, we can successfully certify for these instances that our model predicts the same result as if it were trained on the clean training set. We run the experiment multiple times and show in Section C.3 that the standard deviation is less than 1% in most cases. We also show the abstain rate of certification in Section C.3 and observe that it is generally low. If the abstain rate is high, we can perform in a similar way as in [39] to obtain a variation of our theorem to certify the radius by some margin.

6.5.2.2 *Empirical Robustness: without RAB vs. with RAB*

In addition to the certificates that RAB can provide, the RAB training process also provides good robust accuracy *empirically*, without theoretical guarantees. In Table 1, the "RAB" column reports the empirical robust accuracy — *how often can a malicious input that successfully attacks a vanilla model trick RAB?* We can see that, RAB achieves high empirical robust accuracy, and such empirical robust accuracy achieved by either RAB or other methods serves as an upper bound for the certified robust accuracy provided by RAB under the "RAB-certified" column. It is shown that RAB achieves around 40% empirical robust accuracy on the backdoored instances for MNIST and CIFAR-10, and over 30% for ImageNette. In Section C.1, we also evaluate an empirical adversarial attack on the RAB model and observe a similar behavior as for vanilla models.

6.5.2.3 *Comparison with State-of-the-art Empirical Backdoor Defenses*

Another line of research is to develop empirical methods to automatically detect and remove backdoored training instances. *How does RAB compare with these methods?* We empirically compare the robustness of RAB with other state-of-the-art baseline methods introduced in Section 6.5.1, as shown in Table 1. we observe that although RAB is

not specifically designed for empirical defense, it achieves comparable empirical robust accuracy compared with these baseline methods. RAB outperforms about half of the baselines methods on MNIST and ImageNette and all the baselines on CIFAR-10. Interestingly, our approach performs better on CIFAR-10 than on other tasks while other baselines usually perform badly on CIFAR-10. We attribute this observation to the fact that the benign accuracy on CIFAR-10 is comparably low, so that the baselines based on analyzing feature representations or on model reverse engineering are largely affected and the performance is thus worse. By comparison, RAB only needs to add noise to smooth the training process without analyzing model properties, and is hence less affected by the model itself (similarly, the performance of Mixup is less affected too).

In Section C.1, we additionally evaluate the defenses against a more challenging *all-to-all attack* where many baseline approaches fail, and RAB still achieves good performance. We also show that our approach can be applied to an SVM model for three tabular datasets in Section C.2, while existing approaches cannot work well since there is no distinct "activation layer" in a simple SVM model. Furthermore, for very large attack perturbations, the certification will fail as shown in Section C.1; however, RAB still achieves non-trivial empirical robustness.

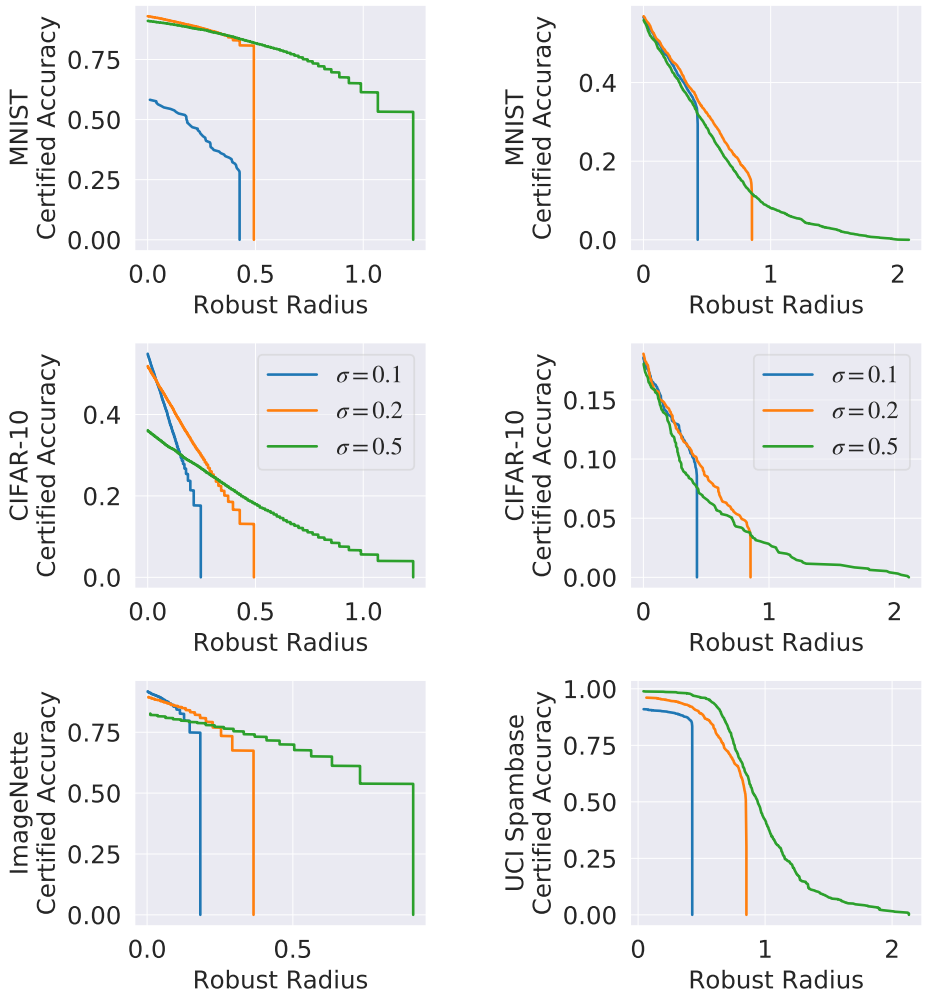
6.5.2.4 *Certified Accuracy Under different Radii*

We further discuss how different certified radii affect the certified accuracy. In Figure 5, we present the certified accuracy as a function of the robust radius given different values for the smoothing parameter σ against blending attack. The conclusions on other backdoor patterns are similar.

In the figures, we plot the certified accuracy of all test cases (instead of only on successfully attacked cases) so that the overall trend can be seen. We can see that the certified accuracy decreases with increased radii and, at a certain point, it suddenly goes to zero, which aligns with existing observations on certified robustness against evasion attacks [39]. Furthermore, stronger noise harms the certified accuracy at a small radius, while improving it at a larger robust radius. It is thus essential to choose an appropriate smoothing noise magnitude according to the task. The certified accuracy of KNN is comparatively low due to its simple structure, but it achieves non-trivial certified accuracy at a larger radius as we do not need Monte Carlo sampling which would result in a finite sampling error that decreases the certified robustness.

6.5.2.5 *Ablation Study: Impact of Deterministic Test-time Augmentation*

We compare the certification accuracy of RAB with and without deterministic test-time augmentation in Figure 6. We plot the certified accuracy of all test cases instead of only on successfully attacked cases to show the comparison on the entire dataset. We observe that the certified accuracy significantly improves with the proposed hash function based deterministic test-time augmentation, especially at small certification radii and with a particularly large gap on ImageNette dataset — without the augmentation, the certified accuracy is only around 20%, while it increases to around 80% with the augmentation. This shows that it is important to include the test-time augmentation during inference, and directly adopting randomized smoothing may not provide satisfactory certified accuracy. The detailed empirical and certified robust accuracies are shown in Section C.3.



(a) KNN. (b) DNN.

Figure 5: Certified accuracy of DNN and KNN at different radii with different smoothing parameters σ against blending attack.

6.5.3 *K-Nearest Neighbors*

Here we present the benchmarks based on our proposed efficient algorithm for KNN models. We perform experiments on the UCI spambase tabular datasets and show the results for $K = 3$ in Table 2. The NC baseline relies on gradient-based reverse engineering, while Mixup relies on mixing label information during training, so these two methods are not included here. The other baselines use intermediate feature vectors in DNN models, which do not exist in KNN models. Therefore, we use the output prediction vector as the feature vector. From the results, we see that for KNN models, RAB achieves good performance for both empirical and certified robustness and outperforms all the baselines, indicating its advantages for specific domains.

This comparison might seem unfair at first glance, since the considered baselines are based on deep feature representations, which are absent in the KNN case. However, firstly, we emphasize that none of the approaches, including RAB, use deep features for this comparison and have hence access to the same amount of information. Secondly,

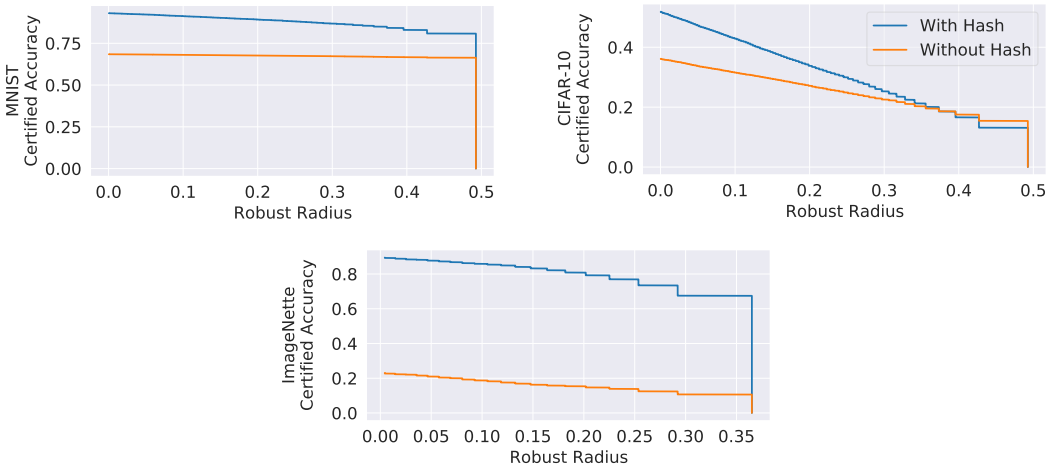


Figure 6: Comparison of the certified accuracy at different radii with and without the proposed deterministic test-time augmentation. The accuracy is evaluated against blending attack with smoothing parameter $\sigma = 0.2$.

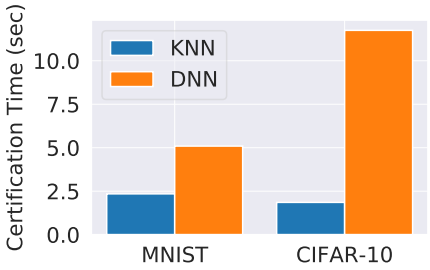


Figure 7: RAB runtime analysis.

this comparison reveals an important property of our approach: while the baselines struggle to handle ML models beyond DNN, RAB is applicable to a wider range of models and still yields non-trivial empirical and certified robust accuracy. To enable a comparison for KNN models which is more favorable to the baselines, we consider kernel KNN with a CNN as the kernel function. From the table in Section C.2, we see that for this scenario, some baselines indeed outperform RAB.

Figure 7 illustrates the runtime of the exact algorithm for KNN vs. the sampling-based method of DNN. We observe that for certifying one input on KNN with $K = 3$ neighbors, using the proposed *exact* certification algorithm takes only 2.5 seconds, which is around 2-3 times faster than the vanilla RAB on MNIST and 6-7 times faster on CIFAR-10. In addition, the runtime is agnostic to the input size but related to the size of the training set. It would be interesting future work to design similar efficient certification algorithms for DNNs. Nevertheless, the KNN algorithm remains slower than the algorithm without certification (which is 1000 times faster than the RAB DNN pipeline), and the improvement of running time is an important future direction.

6.6 CONCLUSION

In this chapter, we have provided theoretical results and practical techniques to certify robustness against backdoor attacks. These results were first enabled by an instantia-

tion of Lemma 1, and required careful further analysis and modifications to provide non-trivial guarantees in practice. Based on these modifications, we have proposed the RAB pipeline, enabling both a heuristic and certified defense mechanism against backdoor attacks. We have validated this pipeline for DNN and KNN models and provided comprehensive benchmarks of certified model robustness against backdoor attacks on diverse datasets.

While these contributions provide first steps towards provable robustness during the model development stage, and answered our first research question in the affirmative, the approach is not without its limitations. Indeed, perhaps the most significant limitation of RAB is that it introduces non-negligible compute requirements. To certify the robustness, we need to train and evaluate multiple models (here, 1000 for MNIST/CIFAR-10 and 200 for ImageNette), which is expensive despite the fact that it is parallelizable and can be accelerated with multiple GPUs. Nevertheless, with our polynomial-time KNN algorithm, we have shown a first step towards mitigating the computational cost and leave further endeavors in this direction as future work. Another limitation is the defender's knowledge of the attack. Indeed, to *certify* the robustness, the defender needs to know 1) an upper bound on the backdoor trigger magnitude (in terms of an L_p norm), 2) an upper bound on the number of poisoned training instances, and, 3) control over the training process. However, to use RAB only as a defense (i.e. without any certificate), the defender only needs to control the training process while 1) and 2) are not needed. The assumption 3) restricts RAB to be a robust training algorithm given an untrusted dataset. In other words, RAB cannot be used to defend against backdoor attacks that interfere with the training process (e.g., [180]).

6.7 PROOFS

Here we provide the proofs for the results stated in the previous sections. We write $\alpha(\phi) = \alpha(\phi; \mathbb{P}_0)$ and $\beta(\phi) = \beta(\phi; \mathbb{P}_0, \mathbb{P}_1)$ for type-I and -II error probabilities.

6.7.1 *Proof of Theorem 3*

Proof. We show tightness by constructing a base classifier h^*, such that the smoothed classifier is consistent with the class probabilities (64) for a given (fixed) input $(x_0, \mathcal{D}_0)$ but whose smoothed version is not robust for adversarial perturbations (Ω_x, Δ) that violate (65). Let ϕ_A and ϕ_B be two likelihood ratio tests for testing the null $Z \sim \mathbb{P}_0$ against the alternative $Z + (\Omega_x, \Delta) \sim \mathbb{P}_1$ and let ϕ_A be such that $\alpha(\phi_A) = 1 - p_A$ and ϕ_B such that $\alpha(\phi_B) = p_B$. Since (Ω_x, Δ) violates (65), we have that $\beta(\phi_A) + \beta(\phi_B) \leqslant 1$. Let p^* be given by

$$p^*(y|x, \mathcal{D}) = \begin{cases} 1 - \phi_A(x - x_0, \mathcal{D} - \mathcal{D}_0) & y = y_A \\ \phi_B(x - x_0, \mathcal{D} - \mathcal{D}_0) & y = y_B \\ \frac{1 - p^*(y_A|x, \mathcal{D}) - p(y_B|x, \mathcal{D})}{C-2} & \text{o.w.} \end{cases} \tag{80}$$

where the notation $\mathcal{D} - \mathcal{D}_0$ denotes subtraction on the features but not on the labels. Note that for binary classification, $C = 2$ we have that $\phi_A = \phi_B$ and hence p^* is well defined since in this case, by assumption $p_A + p_B = 1$. If $C > 2$, note that it follows immediately from the definition of p^* that $\sum_k p^*(y|x, \mathcal{D}) = 1$. Not that we have (point-

wise) $\phi_A \geqslant \phi_B$ provided that $p_A + p_B \leqslant 1$. It follows that for $y \neq y_A, y_B$ we have $p^*(y|x, \mathcal{D}) \propto \phi_A - \phi_B \geqslant 0$. Thus, p^* is a well defined (conditional) probability distribution over labels and $h^*(x, \mathcal{D}) := \arg\max_y p^*(y|x, \mathcal{D})$ is a base classifier. Furthermore, to see that the corresponding smoothed classifier q^* is consistent with the class probabilities (64), consider

$$q^*(y_A|x_0, \mathcal{D}_0) = \mathbb{E}(1 - \phi_A(X, D)) = p_A \tag{81}$$

and

$$q^*(y_B|x_0, \mathcal{D}_0) = \mathbb{E}(\phi_B(X, D)) = \alpha(\phi_B) = p_B. \tag{82}$$

In addition, for any $y \neq y_A, y_B$, we have $q^*(y|x_0, \mathcal{D}_0) = (1 - p_A - p_B)/(C - 2) \leqslant p_B$ since by assumption $p_A + p_B \geqslant 1 - (C - 2) \cdot p_B$. Thus, q^* is consistent with the class probabilities (64) In addition, note that $q^*(y_A|x_0 + \Omega_x, \mathcal{D}_0 + \Delta) = 1 - \beta(\phi_A)$ and $\beta(\phi_B) = q^*(y_B|x_0 + \Omega_x, \mathcal{D}_0 + \Delta)$. Since by assumption $1 - \beta(\phi_A) < \beta(\phi_B)$ we see that indeed $y_A \neq g^*(x_0 + \Omega_x, \mathcal{D}_0 + \Delta)$. □

6.7.2 *Proof of Corollary 1*

Proof. We prove this statement by direct application of Theorem 2. Let $Z = (X, D)$ be the smoothing distribution for q and let $\tilde{Z} := (\Omega_x, \Delta) + Z$ and $\tilde{Z}' := (0, -\Delta) + \tilde{Z}$. Correspondingly, let $\tilde{q}(y|x, \mathcal{D}) = q(y|x + \Omega_x, \mathcal{D} + \Delta)$. By assumption, we have that $\tilde{q}(y_A|x, \mathcal{D}) \geqslant p_A$ and $\max_{y \neq y_A} \tilde{q}(y|x, \mathcal{D}) \leqslant p_B$. We will now apply Theorem 2 to the smoothed classifier $\tilde{q}$. There exist likelihood ratio tests ϕ_A and ϕ_B for testing $\tilde{Z}$ against $\tilde{Z}'$ such that, if

$$\beta(\phi_A) + \beta(\phi_B) > 1 \tag{83}$$

then it follows that $y_A = \arg\max_y \tilde{q}(y|x, \mathcal{D} - \Delta)$. The statement then follows, since $\tilde{q}(y|x, \mathcal{D} - \Delta) = \arg\max_y \tilde{q}(y|x + \Omega_x, \mathcal{D} + \Delta)$. We will now construct the corresponding likelihood ratio tests and show that (83) has the form (71). Note that the likelihood ratio between $\tilde{Z}$ and $\tilde{Z}'$ at $z = (x, d)$ is given by

$$\Lambda(z) = \exp\left(\sum_{i=1}^{n} \langle d_i, -\delta_i \rangle_\Sigma + \frac{1}{2}\langle \delta_i, \delta_i \rangle_\Sigma\right) \tag{84}$$

where $\Sigma = \sigma^2 \mathbb{1}_d$ and $\langle a, b \rangle_\Sigma := \sum_{i=1}^{n} a_i b_i / \sigma^2$. Thus, since singletons have probability 0 under the Gaussian distribution, any likelihood ratio test for testing $\tilde{Z}$ against $\tilde{Z}'$ has the form

$$\phi_t(z) = \begin{cases} 1, & \Lambda(z) \geqslant t. \\ 0, & \Lambda(z) < t. \end{cases} \tag{85}$$

For $p \in [0, 1]$, let

$$t_p := \exp(\Phi^{-1}(p)\sqrt{\sum_{i=1}^{n} \langle \delta_i, \delta_i \rangle_\Sigma - \frac{1}{2}\sum_{i=1}^{n} \langle \delta_i, \delta_i \rangle_\Sigma}) \tag{86}$$

and note that $\alpha(\phi_{t_p}) = 1 - p$ since

$$\alpha(\phi_{t_p}) = 1 - \Phi\left(\frac{\log(t_p) + \frac{1}{2}\sum_{i=1}^{n}\langle\delta_i, \delta_i\rangle_\Sigma}{\sqrt{\sum_{i=1}^{n}\langle\delta_i, \delta_i\rangle_\Sigma}}\right) \tag{87}$$

where Φ is the CDF of the standard normal distribution. Thus, the test $\phi_A \equiv \phi_{t_A}$ with $t_A \equiv t_{p_A}$ satisfies $\alpha(\phi_A) = 1 - p_A$ and the test $\phi_B \equiv \phi_{t_B}$ with $t_B \equiv t_{1-p_B}$ satisfies $\alpha(\phi_B) = p_B$. Computing the type II error probability of ϕ_A yields

$$\beta(\phi_A) = \Phi(\Phi^{-1}(p_A) - \sqrt{\sum_{i=1}^{n}\langle\delta_i, \delta_i\rangle_\Sigma}). \tag{88}$$

and, similarly, the type II error probability of ϕ_B is given by

$$\beta(\phi_B) = \Phi(\Phi^{-1}(1 - p_B) - \sqrt{\sum_{i=1}^{n}\langle\delta_i, \delta_i\rangle_\Sigma}). \tag{89}$$

Finally, we see that $\beta(\phi_A) + \beta(\phi_B) > 1$ is satisfied if and only if

$$\sqrt{\sum_{i=1}^{n}\|\delta_i\|_2^2} < \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)). \tag{90}$$

□

6.7.3 *Proof of Corollary 2*

Proof. We proceed analogously to the proof of Corollary 1 but with a uniform distribution on the feature vectors $D \sim \mathcal{U}([a, b])$ and construct the likelihood ratio tests in the uniform case. Denote by $S := \prod_{i=1}^{n} S_i$, $S_i := \prod_{j=1}^{d}[a + \delta_{ij}, b + \delta_{ij}]$ the support of $\tilde{D} := \Delta + D$ and by $S' := \prod_{i=1}^{n}[a, b]^d$ the support of $\tilde{D}' := D$. Note that the likelihood ratio between $\tilde{Z}$ against $\tilde{Z}'$ at $z = (x, w, v)$ for any $w \in S \cup S'$ is given by

$$\Lambda(z) = \frac{f_{\tilde{Z}'}(z)}{f_{\tilde{Z}}(z)} = \frac{f_{\tilde{W}'}(w)}{f_{\tilde{W}}(w)} = \begin{cases} 0 & w \in S \setminus S', \\ 1 & w \in S \cap S', \\ \infty & w \in S' \setminus S. \end{cases} \tag{91}$$

and that any likelihood ratio test for testing $\tilde{Z}$ against $\tilde{Z}'$ has the form (299). We now construct such likelihood ratio tests ϕ_A, ϕ_B with $\alpha(\phi_A) = 1 - p_A$ and $\alpha(\phi_B) = p_B$. Specifically, we compute q_A, t_A such that these type I error probabilities are satisfied. Notice that

$$p_0 := \mathbb{P}(\tilde{W} \in S \setminus S') = 1 - \prod_{i=1}^{n}\left(\prod_{j=1}^{d}\left(1 - \frac{|\delta_{ij}|}{b - a}\right)_+\right) \tag{92}$$

where $(x)_+ = \max\{x, 0\}$. For $t \geqslant 0$ we have $\mathbb{P}(\Lambda(\tilde{Z}) \leqslant t) = p_0$ if $t < 1$ and 1 otherwise. Thus $t_p := \inf\{t \geqslant 0 \colon \mathbb{P}(\Lambda(Z) \leqslant t) \geqslant p\}$ is given by $t_p = 0$ if $p \leqslant p_0$ and $t_p = 1$ if

$p > p_0$. We notice that, if $p_A \leqslant p_0$, then $t_A \equiv t_{p_A} = 0$. This implies that the type II error probability of the corresponding test ϕ_A is 0 since in this case

$$\beta(\phi_A) = 1 - \mathbb{P}\left(\Lambda(\tilde{Z}') > 0\right) - q_A \cdot \mathbb{P}\left(\Lambda(\tilde{Z}') = 0\right) \tag{93}$$
$$= 1 - \mathbb{P}\left(\tilde{D}' \in S'\right) - q_A \cdot \mathbb{P}\left(\tilde{D}' \in S \setminus S'\right) = 0. \tag{94}$$

Similarly, if $1 - p_B \leqslant p_0$ then $t_B \equiv t_{p_B} = 0$ and we obtain that the corresponding test ϕ_B satisfies $\beta(\phi_B) = 0$. In both cases $\beta(\phi_A) + \beta(\phi_B) > 1$ can never be satisfied and we find that $p_A > p_0$ and $1 - p_B > p_0$ is a necessary condition. In this case, we have that $t_A = t_B = 1$. Let q_A and q_B be defined as

$$q_A := \frac{\mathbb{P}(\Lambda(\tilde{Z}) \leqslant 1) - p_A}{\mathbb{P}(\Lambda(\tilde{Z}) = 1)} = \frac{1 - p_A}{1 - p_0}, \tag{95}$$
$$q_B := \frac{\mathbb{P}(\Lambda(\tilde{Z}) \leqslant 1) - (1 - p_B)}{\mathbb{P}(\Lambda(\tilde{Z}) = 1)} = \frac{1 - (1 - p_B)}{1 - p_0}. \tag{96}$$

Clearly, the corresponding likelihood ratio tests ϕ_A and ϕ_B have significance $1 - p_A$ and p_B respectively. Furthermore, notice that

$$\mathbb{P}\left(\tilde{D}' \in S' \setminus S\right) = \mathbb{P}\left(\tilde{D} \in S \setminus S'\right) = p_0 \tag{97}$$
$$\mathbb{P}\left(\tilde{D}' \in S' \cap S\right) = \mathbb{P}\left(\tilde{D} \in S' \cap S\right) = 1 - p_0 \tag{98}$$

and hence $\beta(\phi_A)$ is given by

$$\beta(\phi_A) = 1 - \mathbb{P}\left(\Lambda(\tilde{Z}') > 1\right) - q_A \cdot \mathbb{P}\left(\Lambda(\tilde{Z}') = 1\right) \tag{99}$$
$$= 1 - p_0 - q_A \cdot (1 - p_0) \tag{100}$$
$$= p_A - p_0. \tag{101}$$

and similarly

$$\beta(\phi_B) = 1 - \mathbb{P}\left(\Lambda(\tilde{Z}') > 1\right) - q_B \cdot \mathbb{P}\left(\Lambda(\tilde{Z}') = 1\right) \tag{102}$$
$$= 1 - p_0 - q_B \cdot (1 - p_0) \tag{103}$$
$$= 1 - p_B - p_0. \tag{104}$$

Finally, the statement follows, since $\beta(\phi_A) + \beta(\phi_B) > 1$ if and only if

$$1 - \left(\frac{p_A - p_B}{2}\right) < \prod_{i=1}^{n} \left(\prod_{j=1}^{d} \left(1 - \frac{|\delta_{ij}|}{b - a}\right)_+ \right). \tag{105}$$

□

6.7.4 *Proof of Theorem 4*

We first formalize KNN classifiers which use quantizeed Euclidean distance as a notion of similarity. Specifically, let $B_1 = [0, b_1), \ldots, B_L := [b_{L-1}, \infty)$ be similarity buckets with increasing $b_1 < b_2 < \ldots, b_{L-1}$ and associated similarity levels $\beta_1 > \beta_2 > \ldots > \beta_L$. Then for $x, x' \in \mathbb{R}^d$ we define their similarity as $\kappa(x, x') := \sum_{l=1}^{L} \beta_l \mathbb{1}_{B_l}(\|x - x'\|_2^2)$

where $\mathbb{1}_{B_l}$ is the indicator function of B_l. Given a dataset $D = (x_i, y_i)_{i=1}^n$ and a test instance x, we define the relation

$$x_i \succeq x_j \iff \begin{cases} \kappa(x_i, x) > \kappa(x_j, x) & \text{if } i > j \\ \kappa(x_i, x) \geqslant \kappa(x_j, x) & \text{if } i \leqslant j \end{cases} \tag{106}$$

which says that the instance x_i is more similar to x, if either it has strictly larger similarity or, if it has the same similarity as x_j, then x_i is more similar if $i < j$. With this binary relation, we define the set of K nearest neighbours of x as $I_K(x, \mathcal{D}) := \{i: \left|\{j \neq i: x_j \succeq x_i\}\right| \leqslant K-1\} \subseteq [n]$ and summarize the per class votes in I_K as a label tally vector $\gamma_k(x, \mathcal{D}) := \#\{i \in I_K(x, \mathcal{D}): y_i = k\}$. The KNN prediction is given by $\mathrm{KNN}(x, \mathcal{D}) = \arg\max_k \gamma_k(x, \mathcal{D})$.

Proof. Our goal is to show that we can compute the smoothed classifier q with $Z = (0, D)$, $D \sim \prod_{i=1}^n \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$ and defined by the probability

$$q(y|x, \mathcal{D}) = \mathbb{P}_D\left(\mathrm{KNN}(x, \mathcal{D} + D) = y\right) \tag{107}$$

in time $\mathcal{O}(K^{2+C} \cdot n^2 \cdot L \cdot C)$. For ease of notation, let $X_i := x_i + D^{(i)}$ and $S_i := \kappa(X_i, x)$ and note that

$$p_i^l := \mathbb{P}\left(S_i = \beta_l\right) = F_{d,\lambda_i}\left(\frac{b_l}{\sigma^2}\right) - F_{d,\lambda_i}\left(\frac{b_{l-1}}{\sigma^2}\right) \tag{108}$$

where F_{d,λ_i} is the CDF of the non-central χ^2-distribution with d degrees of freedom and non-locality parameter $\lambda_i = \|x_i + \delta_i - x\|_2^2 / \sigma^2$. Note that we can write (107) equivalently as

$$\mathbb{P}_D\left(\arg\max_{k'} \gamma_{k'}(x, \mathcal{D} + D)) = y\right) \tag{109}$$

and thus

$$q(y|x, \mathcal{D}) = \sum_{\gamma \in \Gamma_y} \mathbb{P}_D\left(\gamma(x, \mathcal{D} + D) = \gamma\right) \tag{110}$$

where $\Gamma_y := \{\gamma \in [K]^C: \arg\max_{k'} \gamma_{k'} = y\}$. Next, we notice that the event that a tally vector γ occurs, can be partitioned into the events that lead to the given γ and for which instance i has similarity β_l and is in the top-K but not in the top-$(K-1)$. We define these to be the boundary events

$$\mathcal{B}_i^l(\gamma) := \{\forall c: \#\{j \in I_c: X_j \succeq X_i\} = \gamma_c, S_i = \beta_l\} \tag{111}$$

where $I_c = \{i: y_i = c\}$. The probability that a given tally vector γ occurs is thus given by

$$\mathbb{P}_D\left(\gamma(x, \mathcal{D} + D) = \gamma\right) = \sum_{i=1}^n \sum_{l=1}^L \mathbb{P}\left(\mathcal{B}_i^l(\gamma)\right). \tag{112}$$

For fixed i we notice that the for different classes, the events $\{\#\{j \in I_c: X_j \succeq X_i\} = \gamma_c\}$ are pairwise independent, conditioned on $\{S_i = \beta_l\}$. Writing $P_c^l(i, \gamma)$ for the conditional probability

$$\mathbb{P}\left(\#\{i \in I_c: y_i = c\} = \gamma_c | S_i = \beta_l\right) \tag{113}$$

yields

$$\mathbb{P}\left(\mathcal{B}_i^l(\gamma)\right) = p_i^l \cdot \prod_{c=1}^{C} P_c^l(i, \gamma) \tag{114}$$

and hence

$$q(y|x, \mathcal{D}) = \sum_{\gamma \in \Gamma_y} \sum_{i=1}^{n} \sum_{l=1}^{L} p_i^l \cdot \prod_{c=1}^{C} P_c^l(i, \gamma) \tag{115}$$

which requires $\mathcal{O}(K^C \cdot n \cdot L \cdot C)$ evaluations of P_c^l. The next step is to compute the probabilities P_c^l. For that purpose, we need to open up the binary relation $\succeq$. Suppose first that $y_i \neq c$. Then the event that exactly γ_c instances of class c satisfy $X_j \succeq X_i$ is the same as the event that for some $r \leqslant \gamma_c$ exactly r instances with index larger than i have similarity strictly larger than X_i and exactly $\gamma_c - r$ instances with an index smaller than i have similarity larger or equal than X_i. Now suppose that $y_i = c$. Then, the event that exactly γ_c instances of the same class c satisfy $X_j \succeq X_i$ is the same as the event that for some $r \leqslant \gamma_c$ exactly r instances with an index larger than i have similarity strictly larger than X_i and exactly $\gamma_c - r - 1$ instances with an index smaller than i have similarity larger or equal than X_i. This reasoning allows us to write P_c^l in terms of functions

$$R_c^l(i, r) := \mathbb{P}\left(|\{j \in I_c \colon S_j > \beta_l, j > i\}| = r\right) \tag{116}$$

$$Q_c^l(i, r) := \mathbb{P}\left(|\{j \in I_c \colon S_j \geqslant \beta_l, j < i\}| = r\right) \tag{117}$$

as

$$P_c^l(i, \gamma) = \begin{cases} \sum_{r=0}^{\gamma_c} R_c^l(i, r) \cdot Q_c^l(i, \gamma_c - r) & y_i \neq c \\ \sum_{r=0}^{\gamma_c - 1} R_c^l(i, r) \cdot Q_c^l(i, \gamma_c - r - 1) & y_i = c. \end{cases}$$

The functions R_l^c and Q_l^c exhibit a recursive structure that we wish to exploit to get an efficient algorithm. Specifically, we write

$$\alpha_i^l := \mathbb{P}(S_i \leqslant \beta_l) = \sum_{s=l}^{L} p_i^l, \tag{118}$$

and $\bar{\alpha}_i^l := 1 - \alpha_i^l$ and use the following recursion

$$R_l^c(i-1, r) = \begin{cases} R_c^l(i, r) & y_i \neq c \\ \bar{\alpha}_i^l \cdot R_c^l(i, r-1) + \alpha_i^l \cdot R_c^l(i, r) & y_i = c \end{cases}$$

starting at $i = n$ and $r = 0$ and with initial values $R_c^l(i, -1) = 0$, $R_l^c(n, 0) = 1$ and $R_l^c(n, r) = 0$ for $r \geqslant 1$. Similarly,

$$Q_c^l(i+1, r) = \begin{cases} Q_c^l(i, r) & y_i \neq c \\ \bar{\alpha}_i^{l+1} \cdot Q_c^l(i, r-1) + \alpha_i^{l+1} \cdot Q_c^l(i, r) & y_i = c \end{cases} \tag{119}$$

starting the recursion at $i = 1$ and $r = 0$ and with initial values $Q_c^l(i, -1) = 0$, $Q_c^l(1, 0) = 1$ and $Q_c^l(1, r) = 0$ for $r \geqslant 1$. Evaluating P_c^l requires $\mathcal{O}(K)$ calls to R_c^l and Q_c^l each. The computation of R_c^l and Q_c^l can be achieved in $\mathcal{O}(n \cdot K)$ if the values α_i^l are computed beforehand and stored separately (requiring $\mathcal{O}(n \cdot L)$ computations). The entire computation has complexity $\mathcal{O}(K^{C+2} \cdot n^2 \cdot L \cdot C)$. □

7 TRANSFORMATION-SPECIFIC SMOOTHING FOR ROBUSTNESS CERTIFICATION

In the previous chapter we have focused on vulnerabilities of ML systems at the training stage and presented theoretical as well as empirical results which provide guidance on the certification of robustness against backdoor attacks, a particular type of data poisoning attack. In this and the subsequent chapters, we focus our attention on vulnerabilities to which ML systems are prone to during deployment. Specifically, here we study the certification of robustness against instance-level input perturbations arising from semantic transformations such as rotations, changes in contrast or Gaussian blur. In the subsequent chapter we will take a population based view and study the certification of robustness against distribution shifts.

7.1 INTRODUCTION

7.1.1 *Overview*

Recent advances in ML have enabled a plethora of applications in diverse tasks such as image recognition [84], game playing [157, 200] and natural language processing [20, 48, 220]. Despite all of these advances, ML systems are also found exceedingly vulnerable to adversarial attacks – image recognition systems can be adversarially misled [71, 212, 254], and malware detection systems can be evaded easily [219, 261]. In response, recent research has attempted to provide answers to the implied risks and developed empirical defense techniques [143, 222], as well as certified defenses [39, 136, 218, 252, 259].

While empirical defenses are prone to being broken by adaptive attackers [7, 60, 71, 256], certified defenses take a more conservative view and provide robustness conditions under which an ML model is guaranteed to be robust. Such certification techniques usually follow the pattern that the ML model is provably robust against arbitrary adversarial attacks, as long as the perturbation magnitude is below a certain threshold, measured in an ℓ_p norm. However, certifying robustness only against ℓ_p perturbations is not sufficient for attacks based on semantic transformations. For instance, image rotation, scaling, and other semantic transformations are able to mislead ML models effectively [59, 69, 70, 256]. These transformations are both common and practical [21, 91, 170]. For example, it has been shown [98] that brightness and contrast attacks can achieve 91.6% attack success on CIFAR-10, and 71%-100% attack success rate on ImageNet [91]. In practice, brightness- and contrast-based attacks have been demonstrated to be successful in autonomous driving [170, 217]. These types of transformations incur large ℓ_p-norm differences and are thus beyond the reach of existing certifiable defenses [17, 83, 122, 188] which can only certify relatively small magnitudes. To address these shortcomings, here we attempt to provide robustness guarantees against semantic transformations which incur large changes when measured in ℓ_p norm.

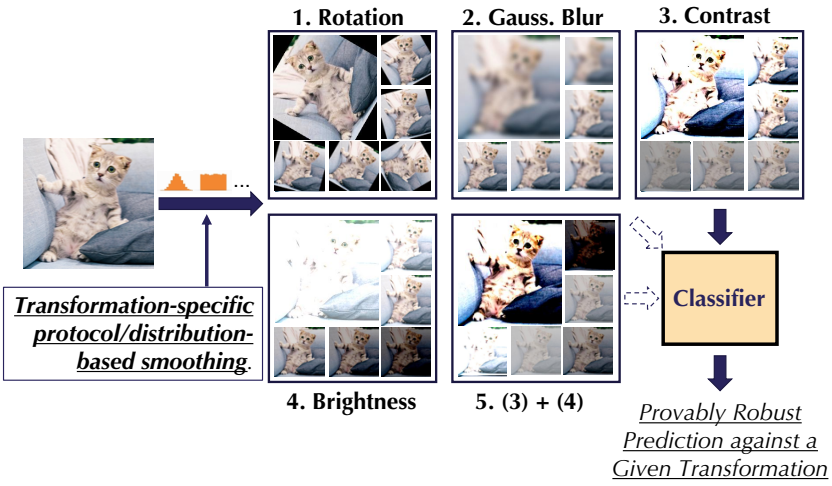


Figure 8: TSS, a general robustness certification framework for diverse semantic transformations. We develop a range of different transformation-specific smoothing protocols and various techniques to provide substantially better certified robustness bounds than state-of-the-art approaches on large-scale datasets.

VULNERABILITIES TO SEMANTIC TRANSFORMATIONS Beyond adversarial ℓ_p perturbations, a realistic threat model is given by transformations that preserve the underlying semantics of a given test instance. Examples for these types of transformations include changes to contrast or brightness levels, or rotation of the entire image. These attacks have the following three characteristics in common:

1. The perturbation stemming from a successful semantic attack typically has higher ℓ_p norm compared to the attacks which are only constrained by bounded ℓ_p-norm. However, these attacks still preserve the underlying semantics: if an image of a car is rotated by 10° it remains an image of a car.

2. These attacks are governed by a low-dimensional parameter space. For example, the rotation attack chooses a one-dimensional rotation angle.

3. Some of these transformations lead to high interpolation errors (e. g., rotations), which incurs additional challenges for certification.

As has been shown [91, 98], these types of attacks can cause significant damage and pose realistic threats for practical ML applications such as autonomous driving [170]. We remark that our proposed framework can be extended to certify robustness against other attacks that share these characteristics even beyond the image domain, such as GAN-based attacks against ML based malware detection [100, 239], where a limited dimension of features of the malware can be manipulated in order to preserve the malicious functionalities. Such perturbations usually incur large ℓ_p differences for the generated instances.

7.1.2 *Contributions*

In this work, we present theoretical and empirical analyses to certify the ML robustness against a wide range of semantic transformations which go beyond ℓ_p-norm bounded

perturbations. The theoretical analysis studies different properties of the transformations and proposes transformation-specific certification techniques. Our empirical results set the new state-of-the-art robustness certification for a wide range of semantic transformations, exceeding existing work by a large margin. Our framework TSS is based on randomised smoothing [39, 128] and provides certified robustness for ML models against a wide range of adversarial transformations (Figure 8). Within our framework, we categorize semantic transformations as either *resolvable* or *differentially resolvable*. As a first building block, we provide certified robustness against *resolvable* transformations, which include brightness, contrast, translation, Gaussian blur, and their composition. In a second step, we build on the techniques for resolvable transformations and develop novel certification techniques for *differentially resolvable* transformations (e. g., rotation and scaling).

For resolvable transformations, we leverage the framework to jointly reason about (1) function smoothing under different smoothing distributions and (2) the properties inherent to each specific transformation. To our best knowledge, this is the first time that the interplay between smoothing distribution and semantic transformations has been analyzed. Indeed, existing work [39, 133, 264] that studies different smoothing distributions considers only ℓ_p perturbations. Based on this analysis, we find that against certain transformations such as Gaussian blur, exponential distribution is better than Gaussian smoothing, which is commonly used in the ℓ_p-case.

For differentially resolvable transformations, such as rotation, scaling, and their composition with other transformations, the common challenge is that they naturally induce interpolation errors. Existing work [8, 65] can provide robustness guarantees but it cannot rigorously certify robustness for ImageNet-scale data. We develop a collection of novel techniques, including stratified sampling and Lipschitz bound computation to provide a tighter and sound upper bound for the interpolation error. We integrate these novel techniques into TSS and further propose a progressive-sampling-based strategy to accelerate the robustness certification. We show that these techniques comprise a scalable and general framework for certifying robustness against differentially resolvable transformations.

We conduct extensive experiments to evaluate the proposed certification framework and show that our framework significantly outperforms the state-of-the-art on different datasets including the large-scale ImageNet against a series of practical semantic transformations. In summary, we make the following set of contributions:

- We propose a function smoothing framework, TSS, to certify ML robustness against general semantic transformations.
- We categorize common adversarial semantic transformations in the literature into *resolvable* and *differentially resolvable* transformations and show that our framework is general enough to certify both types of transformations.
- We theoretically explore different smoothing strategies by sampling from different distributions including non-isotropic Gaussian, uniform, and Laplace distributions. We show that for specific transformations, such as Gaussian blur, smoothing with exponential distribution is better.
- We propose a pipeline, TSS-DR, including a stratified sampling approach, an effective Lipschitz-based bounding technique, and a progressive sampling strategy to

provide rigorous, tight, and scalable robustness certification against differentially resolvable transformations such as rotation and scaling.

- We conduct extensive experiments and show that our framework TSS can provide significantly higher certified robustness compared with the state-of-the-art approaches, against a range of semantic transformations and their composition on MNIST, CIFAR-10, and ImageNet.
- We show that TSS also provides much higher empirical robustness against adaptive attacks and unforeseen corruptions such as CIFAR-10-C and ImageNet-C.

7.1.3 *Related Work*

CERTIFIED ROBUSTNESS AGAINST ℓ_p PERTURBATIONS With the seminal works that opened up research on adversarial vulnerability of neural networks [71, 212], there has emerged a rich body of research on evasion attacks [7, 29, 221, 254] and empirical defenses [146, 189, 197]. To provide robustness certification, different robustness training and verification approaches have been proposed. In particular, interval bound propagation [74, 274], linear relaxations [153, 246, 252, 253, 259], and semidefinite programming [45, 177] have been applied to certify the robustness of ML models. Recently, robustness certification based on randomized smoothing was shown to be scalable and leads to tight robustness guarantees [39, 128, 133]. With improvements on optimizing the smoothing distribution [56, 216, 264] and better training mechanisms [30, 106, 187, 271], the verified robustness of randomized smoothing is further improved. A recent survey summarizes certified robustness approaches [135].

SEMANTIC ATTACKS AGAINST NEURAL NETWORKS Recent work has shown that semantic transformations are able to mislead ML models [69, 98, 256]. For instance, image rotations and translations can attack ML models with 40% - 99% degradation on MNIST, CIFAR-10, and ImageNet on both vanilla models and models that are robust against ℓ_p-bounded perturbations [59]. Brightness/contrast attacks can achieve 91.6% attack success on CIFAR-10, and 71%-100% attack success rate on ImageNet [91]. Our evaluation on empirical robust accuracy (Table 5) for undefended models also confirms these observations. Moreover, brightness attacks have been shown to be of practical concern in autonomous driving [170]. Empirical defenses against semantic transformations have been investigated in [59, 91].

CERTIFIED ROBUSTNESS AGAINST SEMANTIC TRANSFORMATIONS While heuristic defenses against semantic attacks have been proposed, *provable* robustness requires further investigation. Existing certified robustness against transformations is based on heuristic enumeration, interval bound propagation, linear relaxation, or smoothing. Efficient enumeration in VeriVis [171] can handle only discrete transformations. Interval bound propagation has been used to certify common semantic transformations [8, 65, 202]. To tighten the interval bounds, linear relaxations are introduced. DeepG [8] optimizes linear relaxations for given semantic transformations, and Semantify-NN [156] encodes semantic transformations by neural networks and applies linear relaxations for neural networks [246, 274]. However, linear relaxations are loose and computationally intensive compared to our TSS. Recently, Fischer et al [65] have applied a smoothing

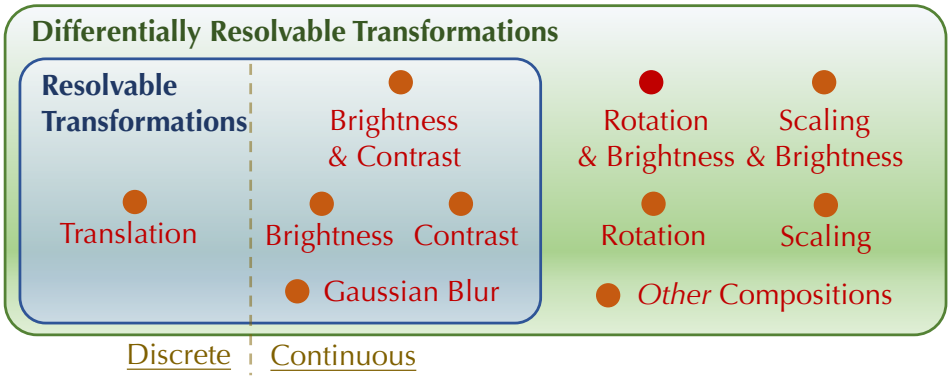


Figure 9: Overview of the categorization of resolvable and differentially resolvable transformations. These two categories cover common adversarial semantic transformations.

scheme to provide provable robustness against transformations but on the large ImageNet dataset, it can provide certification only against random attacks that draw transformation parameters from a pre-determined distribution. More details are available in Section D.6.4.

OUTLINE The remainder of this chapter is organized as follows. In Section 7.2, we introduce the threat model, certification goal, and provide an overview of the method presented in this chapter. We then introduce TSS, our general framework for certifying robustness against semantic transformations, in Section 7.3 before treating resolvable transformations in Section 7.4 and differentially resolvable transformations in Section 7.5. We validate our method experimentally in Section 7.6 and conclude in Section 7.7. Proofs for the main Theorems are provided in Section 7.8 and proofs for further result can be found in the supplementary materials (Appendix D).

7.2 METHOD OVERVIEW

In this section, we first introduce the notations. We then define the threat model, the defense goal and outline the challenges for certifying the robustness against semantic transformations. Finally, we provide a brief overview of the TSS certification framework.

NOTATION We denote the space of inputs as $\mathcal{X} \subseteq \mathbb{R}^d$ and the set of labels as $\mathcal{Y} = \{1, \ldots, C\}$ (where $C \geqslant 2$ is the number of classes). The set of transformation parameters is given by $\mathcal{Z} \subseteq \mathbb{R}^m$ (e.g., rotation angles). We use the notation $\mathbb{P}_X$ to denote the probability measure induced by the random variable X and write f_X for its probability density function. For a set S, we denote its probability by $\mathbb{P}_X(S)$. A classifier is defined to be a deterministic function h mapping inputs $x \in \mathcal{X}$ to classes $y \in \mathcal{Y}$. Formally, a classifier learns a conditional probability distribution $p(y|x)$ over labels and outputs the class that maximizes p, i.e., $h(x) = \arg\max_{y \in \mathcal{Y}} p(y|x)$.

7.2.1 *Threat Model*

SEMANTIC TRANSFORMATIONS We model semantic transformations as deterministic functions $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$, transforming an image $x \in \mathcal{X}$ with a $\mathcal{Z}$-valued parameter

α. For example, we use $\phi_R(x, \alpha)$ to model a rotation of the image x by α degrees counter-clockwise with bilinear interpolation. We further partition semantic transformations into two different categories, namely resolvable and differentially resolvable transformations. We will show that these two categories could cover commonly known semantic attacks. This categorization depends on whether or not it is possible to write the composition of the transformation ϕ with itself as applying the same transformation just once, but with a different parameter, i.e., whether for any α, $\beta \in \mathcal{Z}$ there exists γ such that $\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma)$. Precise definitions are given in Section 7.4 and Section 7.5. Figure 9 presents an overview of the transformations considered in this work.

THREAT MODEL We consider an adversary that launches a semantic attack, a type of data evasion attack, against a given classification model h by applying a semantic transformation ϕ with parameter α to an input image $x \to \phi(x, \alpha)$. We allow the attacker to choose an *arbitrary* parameter α within a predefined (attack) parameter space $\mathcal{S}$. For instance, a naïve adversary who randomly changes brightness from within $\pm 40\%$ is able to reduce the accuracy of a state-of-the-art ImageNet classifier from 74.4% to 21.8% (Table 5). While this attack is an example of a random adversarial attack, our threat model also covers other types of semantic attacks and we provide the first taxonomy for semantic attacks (i.e., resolvable and differentially resolvable) in detail in Section 7.4 and Section 7.5.

7.2.2 *Certification Goal*

Since the only degree of freedom that a semantic adversary has are the transformation parameters, our goal is to characterize a set of parameters for which the model under attack is guaranteed to be robust. Formally, we wish to find a set $\mathcal{S}_{adv} \subseteq \mathcal{Z}$ such that, for a classifier h and adversarial transformation ϕ, we have

$$h(x) = h(\phi(x, \alpha)) \quad \forall\, \alpha \in \mathcal{S}_{adv}. \tag{120}$$

CHALLENGES Certifying ML robustness against semantic transformations is nontrivial and requires careful analysis. We identify the following two main challenges that we aim to address in this work:

(C1) The absolute difference between semantically transformed images in terms of ℓ_p-norms is typically high. This factor causes existing certifiable defenses against ℓ_p bounded perturbations to be inapplicable [17, 83, 122, 188].

(C2) Certain semantic transformations incur additional *interpolation errors*. To derive a robustness certificate, it is required to bound these errors, an endeavour that has been proven to be hard both analytically and computationally. This challenge applies to transformations that involve interpolation, such as rotation and scaling.

We remark that it is in general not feasible to use brute-force approaches such as grid search to enumerate all possible transformation parameters (e.g., rotation angles) since the parameter spaces are typically continuous. Given that different transformations have their own unique properties, it is crucial to provide a *unified* framework that takes into account transformation-specific properties in a general way.

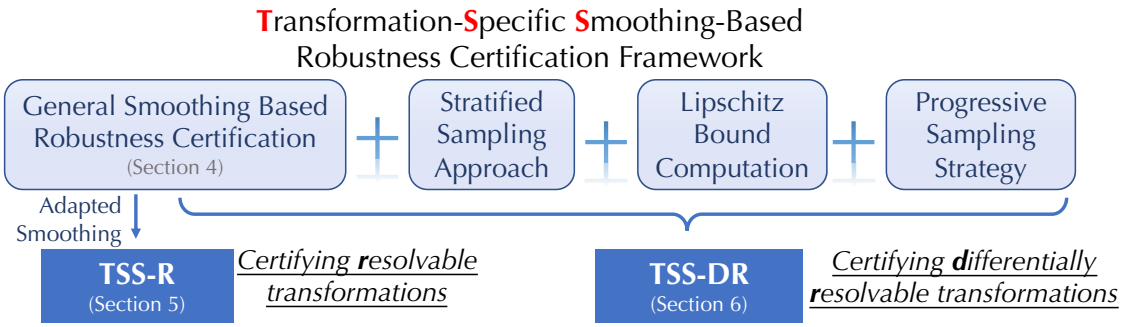


Figure 10: A conceptual overview of TSS.

To address these challenges, we generalize randomized smoothing via our proposed function smoothing framework to certify arbitrary input transformations via different smoothing distributions, paving the way to robustness certifications that go beyond ℓ_p perturbations. This result addresses challenge **(C1)** in a unified way. Based on this generalization and depending on specific transformation properties, we address challenge **(C2)** and propose a series of smoothing strategies and computing techniques that provide robustness certifications for a diverse range of transformations.

We next introduce our generalized function smoothing framework and show how it can be leveraged to certify semantic transformations. We then categorize transformations as either *resolvable* transformations (Section 7.4) such as Gaussian blur, or *differentially resolvable* transformations (Section 7.5) such as rotations.

7.2.3 *Framework Overview*

An overview of our proposed framework TSS is given in Figure 10. We propose the function smoothing framework, a generalization of randomized smoothing, to provide robustness certifications under general smoothing distributions (Section 7.3). This generalization enables us to smooth the model on specific transformation dimensions. We then consider two different types of transformation attacks. For *resolvable* transformations, using the function smoothing framework, we adapt different smoothing strategies for specific transformations and propose Transformation-Specific Smoothing for Resolvable Transformations (TSS-R) in Section 7.4. We show that some smoothing distributions are more suitable for certain transformations. For *differentially resolvable* transformations, to address the interpolation error, we combine function smoothing with the proposed stratified sampling approach and a novel technique for Lipschitz bound computation to compute a rigorous upper bound of the error. We then develop a progressive sampling strategy to accelerate the certification. This pipeline is termed TSS-DR, and we provide details and the theoretical groundwork in Section 7.5.

7.3 TSS: TRANSFORMATION SPECIFIC SMOOTHING

In this section, we extend randomized smoothing and propose TSS for certifying robustness against semantic transformations. This framework constitutes the main building block for TSS-R and TSS-DR against specific types of adversarial transformations.

Given an arbitrary base classifier h, we construct a smoothed classifier g by randomly transforming inputs with parameters sampled from a smoothing distribution. Specifi-

cally, given an input x, the smoothed classifier g predicts the class that h is most likely to return when the input is perturbed by some random transformation. We formalize this intuition in the following definition.

Definition 2 (ε-Smoothed Classifier)**.** *Let $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a transformation, $\varepsilon \sim \mathbb{P}_\varepsilon$ a random variable taking values in $\mathcal{Z}$ and let $h\colon \mathcal{X} \to \mathcal{Y}$ be a base classifier. We define the ε-smoothed classifier $g\colon \mathcal{X} \to \mathcal{Y}$ as $g(x;\varepsilon) = \arg\max_{y\in\mathcal{Y}} q(y|x;\varepsilon)$ where q is given by the expectation with respect to the smoothing distribution ε, i.e.,*

$$q(y|x;\,\varepsilon) := \mathbb{E}(p(y|\,\phi(x,\,\varepsilon))). \tag{121}$$

A key to certifying robustness against a specific transformation is the choice of transformation ϕ in the definition of the smoothed classifier (121). For example, if the goal is to certify the Gaussian blur transformation, a reasonable choice is to use the same transformation in the smoothed classifier. However, for other types of transformations this choice does not lead to the desired robustness certificate, and a different approach is required. In Section 7.4 and Section 7.5, we derive approaches to overcome this challenge and certify robustness against a broader family of semantic transformations.

GENERAL ROBUSTNESS CERTIFICATION Given an input $x \in \mathcal{X}$ and a random variable ε taking values in $\mathcal{Z}$, suppose that the base classifier h predicts $\phi(x,\varepsilon)$ to be of class y_A with probability at least p_A and the second most likely class with probability at most p_B (123). Our goal is to derive a robustness certificate for the ε-smoothed classifier g, i. e., we aim to find a set of perturbation parameters $\mathcal{S}_{\text{adv}}$ depending on p_A, p_B, and smoothing parameter ε such that, for all possible parameters $\alpha \in \mathcal{S}_{\text{adv}}$, it is guaranteed that

$$g(\phi(x,\,\alpha);\,\varepsilon) = g(x;\,\varepsilon) \tag{122}$$

In other words, the prediction of the smoothed classifier can never be changed by applying the transformation ϕ with parameters α that are in the robust set $\mathcal{S}_{\text{adv}}$. The following theorem provides a generic robustness condition that we will subsequently use to obtain conditions on transformation parameters. In particular, this result addresses the first challenge **(C1)** for certifying semantic transformations since this result allows to certify robustness beyond additive perturbations.

Theorem 5. *Let $\varepsilon_0 \sim \mathbb{P}_0$ and $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ on $\mathcal{Z}$ and let $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a semantic transformation. Suppose that $y_A = g(x;\,\varepsilon_0)$ and let $p_A, p_B \in [0,\,1]$ be bounds to the class probabilities, i.e.,*

$$q(y_A|x,\,\epsilon_0) \geqslant p_A > p_B \geqslant \max_{y\neq y_A} q(y|x,\,\epsilon_0). \tag{123}$$

For $t \geqslant 0$, let $\underline{S}_t, \overline{S}_t \subseteq \mathcal{Z}$ be the sets defined as $\underline{S}_t := \{f_1/f_0 < t\}$ and $\overline{S}_t := \{f_1/f_0 \leqslant t\}$ and define the function $\xi\colon [0,\,1] \to [0,\,1]$ by

$$\begin{aligned} &\xi(p) := \sup\{\mathbb{P}_1(S)\colon\ \underline{S}_{\tau_p} \subseteq S \subseteq \overline{S}_{\tau_p}\} \\ &\text{where}\quad \tau_p := \inf\{t \geqslant 0\colon\ \mathbb{P}_0(\overline{S}_t) \geqslant p\}. \end{aligned} \tag{124}$$

Then, if the condition

$$\xi(p_A) + \xi(1 - p_B) > 1 \tag{125}$$

is satisfied, then it is guaranteed that $g(x;\,\varepsilon_1) = g(x;\,\varepsilon_0)$.

A detailed proof for this statement is provided in Section 7.8.1. At a high level, the condition (123) defines a family of classifiers based on class probabilities obtained from smoothing the input x with the distribution ε_0. Based on the Neyman Pearson Lemma [162] from statistical hypothesis testing, shifting $\varepsilon_0 \to \varepsilon_1$ results in bounds to the class probabilities associated with smoothing x with ε_1. For class y_A, the lower bound is given by $\xi(p_A)$, while for any other class, the upper bound is given by $1 - \xi(1 - p_B)$, leading to the the robustness condition $\xi(p_A) > 1 - \xi(1 - p_B)$. It is a more general version of what is proved by Cohen et al. [39], and its generality allows us to analyze a larger family of threat models. Notice that it is not immediately clear how one can obtain the robustness guarantee (122) and deriving such a guarantee from Theorem 5 is nontrivial. We will therefore explain in detail how this result can be instantiated to certify semantic transformations in Section 7.4 and Section 7.5.

Remark 1. *We remark that this result is essentially an instantiation of Lemma 1 applied to the (parameter) smoothing distributions $\mathbb{P}_0$ and $\mathbb{P}_1$, and to functions of the form $h \circ \phi$, where h corresponds to the base classifier and ϕ is the semantic transform. In addition, in the formulation presented here, we use the level sets $\underline{S}_t$ and $\overline{S}_t$ corresponding to acceptance regions of likelihood ratio tests.*

7.4 TSS-R: RESOLVABLE TRANSFORMATIONS

In this section, we define resolvable transformations and then show how Theorem 5 is used to certify this class of semantic transformations. We then proceed to providing a robustness verification strategy for each specific transformation. In addition, we show how the generality of our framework allows us to reason about the best smoothing strategy for a given transformation, which is beyond the reach of related randomized smoothing based approaches [65, 264]. Intuitively, we call a semantic transformation resolvable if we can separate transformation parameters from inputs with a function that acts on parameters and satisfies certain regularity conditions.

Definition 3 (Resolvable transform)**.** *A transformation $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ is called resolvable if for any $\alpha \in \mathcal{Z}$ there exists a resolving function $\gamma_\alpha\colon \mathcal{Z} \to \mathcal{Z}$ that is injective, continuously differentiable, has non-vanishing Jacobian and for which*

$$\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma_\alpha(\beta)) \quad x \in \mathcal{X},\ \beta \in \mathcal{Z}. \tag{126}$$

Furthermore, we say that ϕ is additive, if $\gamma_\alpha(\beta) = \alpha + \beta$.

The following result provides a more intuitive view on Theorem 5, expressing the condition on probability distributions as a condition on the transformation parameters.

Corollary 3. *Suppose that the transformation ϕ in Theorem 5 is resolvable with resolving function γ_α. Let $\alpha \in \mathcal{Z}$ and set $\varepsilon_1 := \gamma_\alpha(\varepsilon_0)$ in the definition of the function ξ. Then, if α satisfies condition (125), it is guaranteed that $g(\phi(x, \alpha); \varepsilon_0) = g(x; \varepsilon_0)$.*

This corollary implies that for resolvable transformations, after we choose the smoothing distribution for the random variable ε_0, we can infer the distribution of $\varepsilon_1 = \gamma_\alpha(\varepsilon_0)$. Then, by using ε_0 and ε_1 in Theorem 5, we can derive an explicit robustness condition from (125) such that for any α satisfying this condition, we can certify the robustness. In particular, for additive transformations we have $\epsilon_1 = \gamma_\alpha(\epsilon_0) = \alpha + \epsilon_0$. For common

smoothing distributions ϵ_0 along with additive transformation, we derive the robustness conditions in Section D.1.

In the next section, we focus on specific resolvable transformations. For certain transformations, this result can be applied directly. However, for some transformations, e.g., the composition of brightness and contrast, more careful analysis is required. We remark that this corollary also serves as a stepping stone to certifying more complex transformations that are in general not resolvable, such as rotations as we will present in Section 7.5.

7.4.1 *Certifying Specific Transformations*

Here we build on our theoretical results from the previous section and provide approaches to certifying a range of different semantic transformations that are resolvable. We state all results here and provide proofs in Section D.2.

7.4.1.1 *Gaussian Blur*

This transformation is widely used in image processing to reduce noise and image detail. Mathematically, applying Gaussian blur amounts to convolving an image with a Gaussian function

$$G_\alpha(k) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-k^2/(2\alpha)\right) \tag{127}$$

where $\alpha > 0$ is the *squared* kernel radius. For $x \in \mathcal{X}$, we define Gaussian blur as the transformation $\phi_B \colon \mathcal{X} \times \mathbb{R}_{\geqslant 0} \to \mathcal{X}$ where

$$\phi_B(x, \alpha) = x * G_\alpha \tag{128}$$

and $*$ denotes the convolution operator. The following lemma shows that Gaussian blur is an *additive transform*. Thus, existing robustness conditions for additive transformations shown in D.1 are directly applicable.

Lemma 5. *The Gaussian blur transformation is additive, i.e., for any* $\alpha, \beta \geqslant 0$, *we have* $\phi_B(\phi_B(x, \alpha), \beta) = \phi_B(x, \alpha + \beta)$.

We notice that the Gaussian blur transformation uses only positive parameters. We therefore consider uniform noise on $[0, a]$ for $a > 0$, folded Gaussians and exponential distribution for smoothing.

7.4.1.2 *Brightness and Contrast*

This transformation first changes the brightness of an image by adding a constant value $b \in \mathbb{R}$ to every pixel, and then alters the contrast by multiplying each pixel with a positive factor e^k, for some $k \in \mathbb{R}$. We define the brightness and contrast transformation $\phi_{BC} \colon \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}$ as

$$(x, \alpha) \mapsto \phi_{BC}(x, \alpha) := e^k(x + b), \quad \alpha = (k, b)^T \tag{129}$$

where $k, b \in \mathbb{R}$ are contrast and brightness parameters, respectively. We remark that ϕ_{BC} is resolvable; however, it is not additive and applying Corollary 3 directly using the resolving function γ_α leads to analytically intractable expressions. On the other hand, if the parameters k and b follow independent Gaussian distributions, we can circumvent this difficulty as follows. Given $\varepsilon_0 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, \tau^2))$, we compute the bounds p_A and p_B to the class probabilities associated with the classifier $g(x; \varepsilon_0)$, i.e., smoothed with ε_0. In the next step, we identify a distribution ε_1 with the property that we can map any lower bound p of $q(y|x; \varepsilon_0)$ to a lower bound on $q(y|x; \varepsilon_1)$. Using ε_1 as a bridge, we then derive a robustness condition, which is based on Theorem 5, and obtain the guarantee that $g(\phi_{BC}(x, \alpha); \varepsilon_0) = g(x; \varepsilon_0)$ whenever the transformation parameters satisfy this condition. The next lemma shows that the distribution ε_1 with the desired property (lower bound to the classifier smoothed with ε_1) is given by a Gaussian with transformed covariance matrix.

Lemma 6. *Let $x \in \mathcal{X}$, $k \in \mathbb{R}$, and suppose that*

$$\varepsilon_0 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, \tau^2)) \quad \text{and} \quad \varepsilon_1 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, e^{-2k}\tau^2)). \tag{130}$$

Suppose that $q(y|x; \varepsilon_0) \geqslant p$ for some $p \in [0, 1]$ and $y \in \mathcal{Y}$. Let Φ be the cumulative density function of the standard Gaussian. Then

$$q(y|x; \varepsilon_1) \geqslant \begin{cases} 2\Phi\left(e^k \Phi^{-1}\left(\frac{1+p}{2}\right)\right) - 1 & k \leqslant 0 \\ 2\left(1 - \Phi\left(e^k \Phi^{-1}(1 - \frac{p}{2})\right)\right) & k > 0. \end{cases} \tag{131}$$

Now suppose that $g(\cdot; \varepsilon_0)$ makes the prediction y_A at x with probability at least p_A. Then, the preceding lemma tells us that the prediction confidence of $g(\cdot; \varepsilon_1)$ satisfies the lower bound (131) for the same class. Based on these confidence levels, we instantiate Theorem 5 with the random variables ε_1 and $\alpha + \varepsilon_1$ to get a robustness condition.

Lemma 7. *Let ε_0 and ε_1 be as in Lemma 6 and suppose that*

$$q(y_A|x; \varepsilon_1) \geqslant \tilde{p}_A > \tilde{p}_B \geqslant \max_{y \neq y_A} q(y|x; \varepsilon_1). \tag{132}$$

Then it is guaranteed that $y_A = g(\phi_{BC}(x, \alpha); \varepsilon_0)$ as long as $\alpha = (k, b)^T$ satisfies

$$\sqrt{(k/\sigma)^2 + (b/(e^{-k}\tau))^2} < \frac{1}{2}\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(\tilde{p}_B)\right). \tag{133}$$

In practice, we apply this lemma by replacing $\tilde{p}_A$ and $\tilde{p}_B$ in (133) with the bound computed from (131) based on the class probability bounds p_A and p_B associated with the classifier $g(x; \varepsilon_0)$. In addition, instead of certifying a single pair (k, b), in practice we certify the robustness against a set of transformation parameters

$$\mathcal{S}_{\text{adv}} = \{(k, b) | k \in [-k_0, k_0], b \in [-b_0, b_0]\}, \tag{134}$$

which stands for any contrast change within e^{k_0} and brightness change within b_0. Since it is not feasible to check every $(k, b) \in \mathcal{S}_{\text{adv}}$, we relax the robustness condition in Lemma 7 to

$$\sqrt{(k/\sigma)^2 + (b/(\min\{e^{-k}, 1\}\tau))^2} < \frac{1}{2}\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(\tilde{p}_B)\right). \tag{135}$$

Table 3: Summary of the robustness certification strategies for resolvable transformations. The confidence bounds p_A and p_B are computed using Monte-Carlo sampling.

Transformation	Strategy
Gaussian Blur	Apply Corollary 9 (Section D.1)
Brightness	Apply Corollary 8 (Section D.1)
Translation	Apply Corollary 8 (Section D.1)
Brightness and Contrast	Compute $\tilde{p}_A$ using Lemma 6, then apply Lemma 7
Gaussian Blur, Brightness, Contrast and Translation	Compute $\tilde{p}_A$ using Corollary 13, then apply Lemma 21 (Section D.2.4)

Thus, we only need to verify the condition (135) for (k_0, b_0) and $(-k_0, b_0)$ to certify the robustness for any (k, b) in (134). This is because the LHS of (135) is monotonically increasing w.r.t. $|k|$ and $|b|$, and the RHS of (135) is equal to $\Phi^{-1}(\tilde{p}_A)$ that is monotonically decreasing w.r.t. $|k|$. Throughout the experiments, we use this strategy for certification of brightness and contrast.

7.4.1.3 *Translation*

Let $\bar{\phi}_T : \mathcal{X} \times \mathbb{Z}^2 \to \mathcal{X}$ be the transformation moving an image k_1 pixels to the right and k_2 pixels to the bottom with reflection padding. In order to handle continuous noise distributions, we define the translation transformation $\phi_T : \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}$ as $\phi_T(x, \alpha) = \bar{\phi}_T(x, [\alpha])$ where $[\cdot]$ denotes rounding to the nearest integer, applied element-wise. We note that ϕ_T is an *additive transform*, allowing us to directly apply Corollary 3 and derive robustness conditions. We note that if we use black padding instead of reflection padding, the transformation is not additive. However, since the number of possible translations is finite, another possibility is to use a simple brute-force approach that can handle black padding, which has already been studied extensively [156, 171].

7.4.1.4 *Composition of Gaussian Blur, Brightness, Contrast, and Translation*

Interestingly, the composition of all these four transformations is still resolvable. Thus, we are able to derive the explicit robustness condition for this composition based on Corollary 3, as shown Section D.2.4. Based on this robustness condition, we compute practically meaningful robustness certificates as we will present in experiments in Section 7.6.

7.4.1.5 *Robustness Certification Strategies*

With these robustness conditions, for a given clean input x, a transformation ϕ, and a set of parameters $\mathcal{S}_{adv}$, we certify the robustness of the smoothed classifier g with two steps: 1) estimate p_A and p_B (see eq. (123)) with Monte-Carlo sampling and high-confidence bound following [39]; and 2) leverage the robustness conditions to obtain the certificate. A summary for each transformation including the used robustness conditions is shown in Table 3.

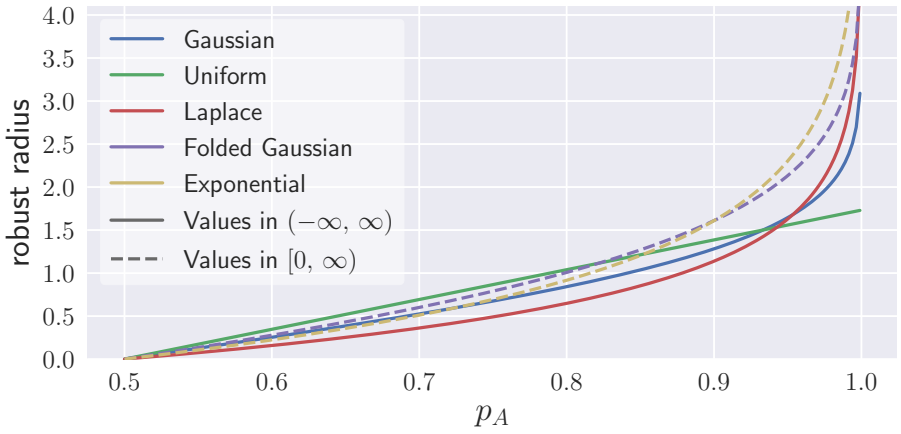


Figure 11: Robust radius comparison for different one-dimensional noise distributions, each with unit variance.

7.4.2 *Properties of Smoothing Distributions*

The robustness condition in Theorem 5 is generic and leaves a degree of freedom in regard to which smoothing distribution should be used. Previous work mainly provides results for cases in which this distribution is Gaussian [39, 271], while it is nontrivial to extend it to other distributions. Here, we aim to answer this question and provide results for a range of distributions, and discuss their differences. As we will see, *for different scenarios, different distributions behave differently and can certify different radii.* We instantiate Theorem 5 with an arbitrary transformation ϕ and with $\varepsilon_1 := \alpha + \varepsilon_0$ where ε_0 is the smoothing distribution and α is the transformation parameter. The robust radius is then derived by solving condition (125) for α.

Figure 11 illustrates robustness radii associated with different smoothing distributions, each scaled to have unit variance. The bounds are derived in Section D.1 and summarized in Table 4. We emphasize that the contribution of this work is not merely these results on different smoothing distributions but, more importantly, the *joint study between different smoothing mechanisms and different semantic transformations.* To compare the different radii for a fixed base classifier, we assume that *the smoothed classifier* $g(\cdot; \varepsilon)$ *always has the same confidence* p_A *for noise distributions with equal variance.* Finally, we provide the following conclusions and we will verify them empirically in Section 7.6.3.1.

1. *Exponential noise can provide larger robust radius.* We notice that smoothing with exponential noise generally allows for larger adversarial perturbations than other distributions. We also observe that, while all distributions behave similarly for low confidence levels, it is only non-uniform noise distributions that converge toward $+\infty$ when $p_A \to 1$ and exponential noise converges quickest.

2. *Additional knowledge can lead to larger robust radius.* When we have additional information on the transformation, e.g., all perturbations in Gaussian blur are positive, we can take advantage of this additional information and certify larger radii. For example, under this assumption, we can use folded Gaussian noise for smoothing instead of a standard Gaussian, resulting in a larger radius.

Table 4: Comparison of certification radii with $p_A + p_B = 1$. The variance and noise dimension are set to 1 for each distribution.

Distribution	Domain	Robust Radius
Gaussian$(0, 1)$	$\mathbb{R}$	$\Phi^{-1}(p_A)$
Laplace$(0, 1/\sqrt{2})$	$\mathbb{R}$	$-\log(2-2p_A)/\sqrt{2}$
Uniform$[-\sqrt{3}, -\sqrt{3}]$	$\mathbb{R}$	$2\sqrt{3}\cdot(p_A - 1/2)$
Exponential(1)	$\mathbb{R}_{\geqslant 0}$	$-\log(2-2p_A)$
FoldedGaussian$(0, \sqrt{\frac{\pi}{\pi-2}})$	$\mathbb{R}_{\geqslant 0}$	$\sqrt{\frac{\pi}{\pi-2}}\cdot\left(\Phi^{-1}\left(\frac{1+p_A}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right)$

7.5 TSS-DR: DIFFERENTIALLY RESOLVABLE TRANSFORMATIONS

As we have seen, our proposed function smoothing framework can directly deal with resolvable transformations. However, due to their use of interpolation, some important transformations do not fall into this category, including rotation, scaling, and their composition with resolvable transformations. In this section, we show that they belong to the more general class of *differentially resolvable transformations*. To address challenge **(C2)**, we propose TSS-DR to provide rigorous robustness certification using our function smoothing framework as a central building block.

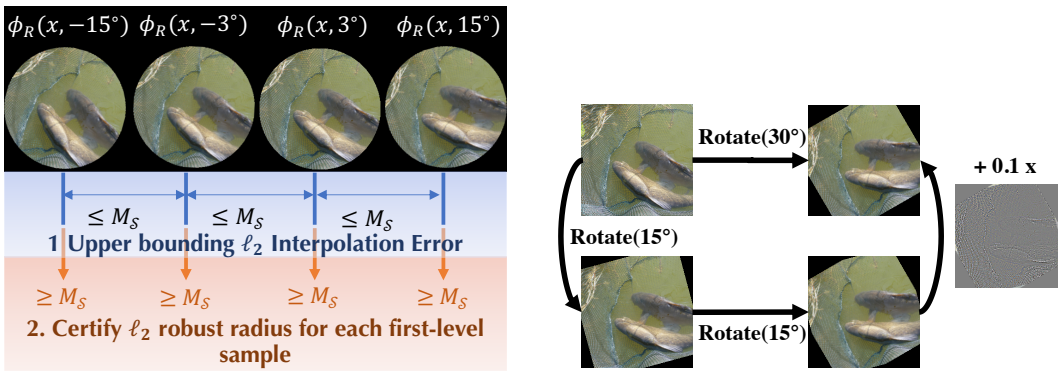
Common semantic transformations such as rotations and scaling do not fall into the category of resolvable transformations due to their use of interpolation. To see this issue, consider the rotation transformation denoted by ϕ_R. As shown in Figure 12b, despite being very similar, the image rotated by 30° is different from the image rotated separately by 15° and then again by 15°. The reason for this is the bilinear interpolation occurring during the rotation. Therefore, if the attacker inputs $\phi_R(x, 15)$, the smoothed classifier defined in Section 7.4 outputs

$$g(\phi_R(x, 15);\, \varepsilon) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}\left(p(y|\, \phi_R(\phi_R(x, 15), \varepsilon))\right), \tag{136}$$

which is a weighted average over the predictions of the base classifier on the randomly perturbed set $\{\phi_R(\phi_R(x, 15), \alpha) \colon \alpha \in \mathcal{Z}\}$. However, in order to use Corollary 3 and to reason about whether this prediction agrees with the prediction on the clean input (i.e., the average prediction on $\{\phi_R(x, \alpha) \colon \alpha \in \mathcal{Z}\}$), we need ϕ_R to be resolvable. As it turns out, this is not the case for transformations that involve interpolation such as rotation and scaling. To address these challenges, we define a transformation ϕ to be *differentially resolvable*, if it can be written in terms of a resolvable transformation ψ and a parameter mapping δ.

Definition 4 (Differentially resolvable transform)**.** *Let $\phi \colon \mathcal{X} \times \mathcal{Z}_\phi \to \mathcal{X}$ be a transformation with noise space $\mathcal{Z}_\phi$ and let $\psi \colon \mathcal{X} \times \mathcal{Z}_\psi \to \mathcal{X}$ be a resolvable transformation with noise space $\mathcal{Z}_\psi$. We say that ϕ can be resolved by ψ if for any $x \in \mathcal{X}$ there exists function $\delta_x \colon \mathcal{Z}_\phi \times \mathcal{Z}_\phi \to \mathcal{Z}_\psi$ such that for any $\alpha \in \mathcal{Z}_\phi$ and any $\beta \in \mathcal{Z}_\phi$,*

$$\phi(x, \alpha) = \psi(\phi(x, \beta), \delta_x(\alpha, \beta)). \tag{137}$$



(a) (b)

Figure 12: (a) High-level illustration of our robustness certification pipeline TSS-DR for differentially resolvable transformations; (b) interpolation error.

This definition leaves open a certain degree of freedom on the choice of the resolvable transformation ψ. For example, we can choose the resolvable transformation corresponding to additive noise

$$\psi\colon \mathcal{X}\times\mathcal{X}\to\mathcal{X},\ (x,\delta)\mapsto x+\delta, \tag{138}$$

which lets us write any transformation ϕ as $\phi(x,\alpha)=\phi(x,\beta)+(\phi(x,\alpha)-\phi(x,\beta))=\psi(\phi(x,\beta),\delta)$ with $\delta=(\phi(x,\alpha)-\phi(x,\beta))$. In other words, $\phi(x,\alpha)$ can be viewed as first being transformed to $\phi(x,\beta)$ and then to $\phi(x,\beta)+\delta$.

7.5.1 *Overview of TSS-DR*

Here, we derive a general robustness certification strategy for differentially resolvable transformations. Suppose that our goal is to certify the robustness against a transformation ϕ that can be resolved by ψ and for transformation parameters from the set $\mathcal{S}\subseteq\mathcal{Z}_\phi$. To that end, we first sample a set of parameters $\{\alpha_i\}_{i=1}^N\subseteq\mathcal{S}$, and transform the input, with those sampled parameters. This yields the set of transformed inputs $\{\phi(x,\alpha_i)\}_{i=1}^N$. In the second step, we compute the class probabilities for each transformed input $\phi(x,\alpha_i)$ with the classifier smoothed with the resolvable transformation ψ. Finally, the intuition is that, if every $\alpha\in\mathcal{S}$ is close enough to one of the sampled parameters, then the classifier is guaranteed to be robust against parameters from the set $\mathcal{S}$. In the next theorem, we show the existence of such a "proximity set" for general δ_x.

Theorem 6. *Let $\phi\colon\mathcal{X}\times\mathcal{Z}_\phi\to\mathcal{X}$ be a transformation that is resolved by $\psi\colon\mathcal{X}\times\mathcal{Z}_\psi\to\mathcal{X}$. Let $\varepsilon\sim\mathbb{P}_\varepsilon$ be a $\mathcal{Z}_\psi$-valued random variable and suppose that the smoothed classifier $g:\mathcal{X}\to\mathcal{Y}$ given by $q(y|x;\varepsilon)=\mathbb{E}(p(y|\psi(x,\varepsilon)))$ predicts $g(x;\varepsilon)=y_A=\arg\max_y q(y|x;\varepsilon)$. Let $\mathcal{S}\subseteq\mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N\subseteq\mathcal{S}$ be a set of transformation parameters such that for any i, the class probabilities satisfy*

$$q(y_A|\phi(x,\alpha_i);\varepsilon)\geqslant p_A^{(i)}\geqslant p_B^{(i)}\geqslant\max_{y\neq y_A}q(y|\phi(x,\alpha_i);\varepsilon). \tag{139}$$

Then there exists a set $\Delta^*\subseteq\mathcal{Z}_\psi$ with the property that, if for any $\alpha\in\mathcal{S}$, $\exists\alpha_i$ with $\delta_x(\alpha,\alpha_i)\in\Delta^*$, then it is guaranteed that

$$q(y_A|\phi(x,\alpha);\varepsilon)>\max_{y\neq y_A}q(y|\phi(x,\alpha);\varepsilon). \tag{140}$$

In Theorem 6, the smoothed classifier $g(\cdot;\varepsilon)$ is based on the resolvable transformation ψ that serves as a starting point to certify the target transformation ϕ. To certify ϕ over its parameter space $\mathcal{S}$, we input N transformed samples $\phi(x,\alpha_i)$ to the smoothed classifier $g(\cdot;\varepsilon)$. Then, we get Δ^*, the certified robust parameter set for the resolvable transformation ψ. This Δ^* means that for any $\phi(x,\alpha_i)$, if we apply the transformation ψ with any parameter $\delta\in\Delta^*$, the resulting instance $\psi(\phi(x,\alpha_i),\delta)$ is robust for $g(\cdot;\varepsilon)$. Since ϕ is resolvable by ψ, i.e., for any $\alpha\in\mathcal{S}$, there exists an α_i and $\delta\in\Delta^*$ such that $\phi(x,\alpha)=\psi(\phi(x,\alpha_i),\delta)$, we can assert that for any $\alpha\in\mathcal{S}$, the output of $g(\cdot;\varepsilon)$ on $\phi(x,\alpha)$ is robust.

The key of using this theorem for a specific transformation is to choose the resolvable transformation ψ that can enable a tight calculation of Δ^* under a specific way of sampling $\{\alpha_i\}_{i=1}^N$. First, we observe that a large family of transformations including rotation and scaling can be resolved by the additive transformation $\psi\colon \mathcal{X}\times\mathcal{X}\to\mathcal{X}$ defined by $(x,\delta)\mapsto x+\delta$. Indeed, any transformation whose pixel value changes are continuous (or with finite discontinuities) with respect to the parameter changes are differentially resolvable—they all can be resolved by the additive transformation. Choosing isotropic Gaussian noise $\varepsilon\sim\mathcal{N}(0,\sigma^2\mathbb{1}_d)$ as smoothing noise then leads to the condition that the maximum ℓ_2-interpolation error between the interval $\mathcal{S}=[a,b]$ (which is to be certified) and the sampled parameters α_i must be bounded by a radius R. This result is shown in the next corollary, which is derived from Theorem 6.

Corollary 4. *Let $\psi(x,\delta)=x+\delta$ and let $\varepsilon\sim\mathcal{N}(0,\sigma^2\mathbb{1}_d)$. Furthermore, let ϕ be a transformation with parameters in $\mathcal{Z}_\phi\subseteq\mathbb{R}^m$ and let $\mathcal{S}\subseteq\mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N\subseteq\mathcal{S}$. Let $y_A\in\mathcal{Y}$ and suppose that for any i, the ε-smoothed classifier defined by $q(y|x;\varepsilon):=\mathbb{E}(p(y|x+\varepsilon))$ has class probabilities that satisfy*

$$q(y_A|\phi(x,\alpha_i);\varepsilon)\geqslant p_A^{(i)}\geqslant p_B^{(i)}\geqslant\max_{y\neq y_A}q(y|\phi(x,\alpha_i);\varepsilon). \tag{141}$$

Then it is guaranteed that $\forall\alpha\in\mathcal{S}\colon\ y_A=\arg\max_y q(y|\phi(x,\alpha);\varepsilon)$ if the maximum interpolation error

$$M_{\mathcal{S}}:=\max_{\alpha\in\mathcal{S}}\min_{1\leqslant i\leqslant N}\|\phi(x,\alpha)-\phi(x,\alpha_i)\|_2 \tag{142}$$

satisfies

$$M_{\mathcal{S}}<R:=\frac{\sigma}{2}\min_{1\leqslant i\leqslant N}\left(\Phi^{-1}\left(p_A^{(i)}\right)-\Phi^{-1}\left(p_B^{(i)}\right)\right). \tag{143}$$

In a nutshell, this corollary shows that if the smoothed classifier classifies all samples of transformed inputs $\{\phi(x,\alpha_i)\}_{i=1}^N$ consistent with the original input and the smallest gap between confidence levels $p_A^{(i)}$ and $p_B^{(i)}$ is large enough, then it is guaranteed to make the same prediction on transformed inputs $\phi(x,\alpha)$ for *any* $\alpha\in\mathcal{S}$.

The main challenge now lies in computing a tight and scalable upper bound $M\geqslant M_{\mathcal{S}}$. Given this bound, a set of transformation parameters $\mathcal{S}$ can then be certified by computing R in (143) and checking that $R>M$. With this methodology, we address challenge **(C2)** and provide means to certify transformations that incur interpolation errors. Figure 12a illustrates this methodology on a high level for the rotation transformation as an example. In the following, we present the general methodology that provides an upper bound of the interpolation error $M_{\mathcal{S}}$ and provide closed form expressions for

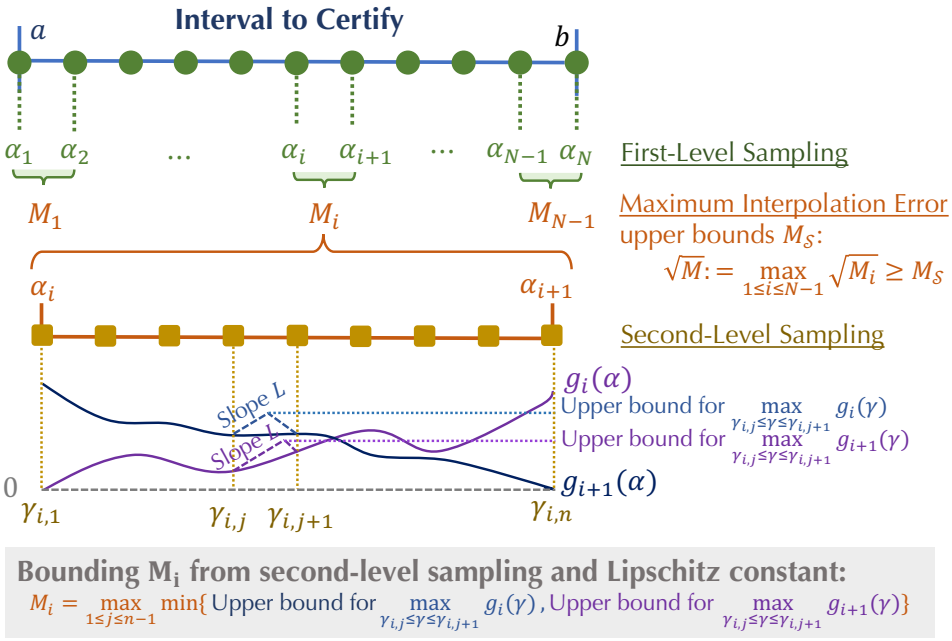


Figure 13: An overview of our interpolation error bounding technique based on stratified sampling and Lipschitz computation.

rotation and scaling. In Section D.3, we provide proofs and further extend this methodology to certify transformation compositions such as rotation, brightness change and ℓ_2 perturbations.

We remark that dealing with the interpolation error has already been tried before [8, 65]. However, these approaches either leverage explicit linear or interval bound propagation – techniques that are either not scalable or not tight enough. Therefore, on large datasets such as ImageNet, they can provide only limited certification (e.g., against certain random attack instead of any attack).

7.5.2 *Upper Bounding the Interpolation Error*

Here, we present the general methodology to compute a rigorous upper bound of the interpolation error introduced in Corollary 4. The methodology presented here is based on stratified sampling and is of a general nature; an explicit computation is shown for the case of rotation toward the end of this subsection.

Let $S = [a, b]$ be an interval of transformation parameters that we wish to certify and let $\{\alpha_i\}_{i=1}^N$ be parameters dividing S uniformly, i.e.,

$$\alpha_i = a + (b-a)\cdot\frac{i-1}{N-1}, \quad i = 1, \ldots, N. \tag{144}$$

The set of these parameters corresponds to the first-level samples in stratified sampling. With respect to these first-level samples, we define the functions $g_i\colon [a, b] \to \mathbb{R}_{\geqslant 0}$ as

$$\alpha \mapsto g_i(\alpha) := \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2^2 \tag{145}$$

corresponding to the squared ℓ_2 interpolation error between the image x transformed with α and α_i, respectively. For each first-level interval $[\alpha_i, \alpha_{i+1}]$ we look for an upper bound M_i such that

$$M_i \geqslant \max_{\alpha_i \leqslant \alpha \leqslant \alpha_{i+1}} \min\{g_i(\alpha), g_{i+1}(\alpha)\}. \tag{146}$$

It is easy to see that $\max_{1 \leqslant i \leqslant N-1} M_i \geqslant M_S^2$ and hence setting

$$\sqrt{M} := \max_{1 \leqslant i \leqslant N-1} \sqrt{M_i} \tag{147}$$

is a valid upper bound to M_S. The problem has thus reduced to computing the upper bounds M_i associated with each first-level interval $[\alpha_i, \alpha_{i+1}]$. To that end, we now continue with a second-level sampling within the interval $[\alpha_i, \alpha_{i+1}]$ for each i. Namely, let $\{\gamma_{i,j}\}_{j=1}^n$ be parameters dividing $[\alpha_i, \alpha_{i+1}]$ uniformly, i.e.,

$$\gamma_{i,j} = \alpha_i + (\alpha_{i+1} - \alpha_i) \cdot \frac{j-1}{n-1}, \quad j = 1, \ldots, n. \tag{148}$$

Now, suppose that L is a global Lipschitz constant for all functions $\{g_i\}_{i=1}^N$. By definition, for any $1 \leqslant i \leqslant N-1$, L satisfies

$$L \geqslant \max\left\{ \max_{c,d \in [\alpha_i, \alpha_{i+1}]} \left| \frac{g_i(c) - g_i(d)}{c-d} \right|, \max_{c,d \in [\alpha_i, \alpha_{i+1}]} \left| \frac{g_{i+1}(c) - g_{i+1}(d)}{c-d} \right| \right\}. \tag{149}$$

In the following, we will derive explicit expressions for L for rotation and scaling. Given the Lipschitz constant L, one can show the following closed-form expression for M_i:

$$\begin{aligned} M_i = \frac{1}{2} \max_{1 \leqslant j \leqslant n-1} \Big(& \min\big\{ g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}), \\ & g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1}) \big\} \Big) + L \cdot \frac{b-a}{(N-1)(n-1)}. \end{aligned} \tag{150}$$

An illustration of this bounding technique using stratified sampling is shown in Figure 13. We notice that, as the number N of first-level samples is increased, the interpolation error M_i becomes smaller by shrinking the sampling interval $[\alpha_i, \alpha_{i+1}]$; similarly, increasing the number of second-level samples n makes the upper bound of the interpolation error M_i tighter since the term $L(b-a)/((N-1)(n-1))$ decreases. Furthermore, it is easy to see that as $N \to \infty$ or $n \to \infty$ we have $M \to M_S^2$, i.e., our interpolation error estimation is *asymptotically tight*. Finally, this tendency also highlights an important advantage of our two-level sampling approach: without stratified sampling, it is required to sample $N \times n$ α_i's in order to achieve the same level of approximation accuracy. As a consequence, these $N \times n$ α_i's in turn require to evaluate the smoothed classifier in Corollary 4 $N \times n$ times, compared to just N times in our case.

It thus remains to find a way to efficiently compute the Lipschitz constant L for different transformations. In the following, we derive closed form expressions for rotation and scaling transformations.

7.5.3 *Computing the Lipschitz Constant*

Here, we derive a global Lipschitz constant L for the functions $\{g_i\}_{i=1}^N$ defined in (145), for rotation and scaling transformations. In the following, we define K-channel images

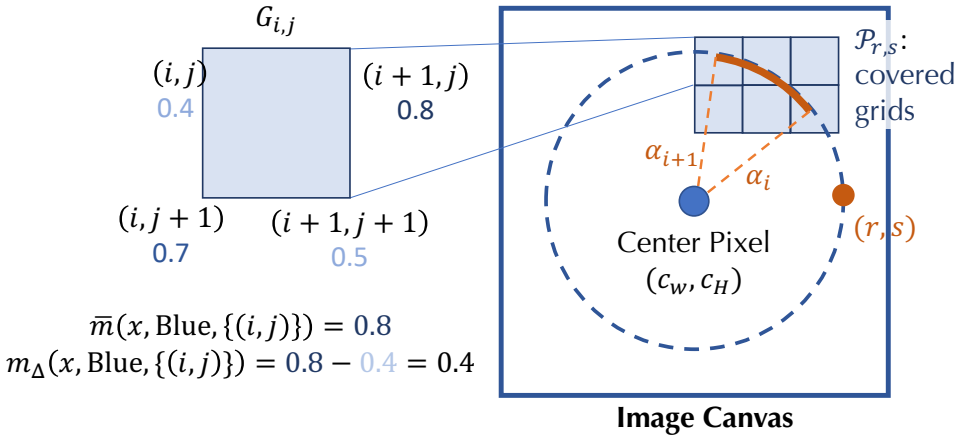


Figure 14: An illustration of the grid pixel generator $G_{i,j}$, color extractors $\bar{m}$ and m_Δ (take blue channel as example), and the set $\mathcal{P}_{r,s}$.

of width W and height H to be tensors $x \in \mathbb{R}^{K\times W\times H}$ and define the region of valid pixel indices as $\Omega := [0, W-1] \times [0, H-1] \cap \mathbb{N}^2$. Furthermore, for $(r, s) \in \Omega$, we define $d_{r,s}$ to be the ℓ_2-distance to the center of an image, i.e.,

$$d_{r,s} = \sqrt{(r-(W-1)/2)^2 + (s-(H-1)/2)^2}. \tag{151}$$

For ease of notation we make the following definitions that are illustrated in 14.

Definition 5 (Grid Pixel Generator). *For pixels $(i, j) \in \Omega$, we define the grid pixel generator G_{ij} as*

$$G_{ij} := \{(i, j), (i+1, j), (i, j+1), (i+1, j+1)\}. \tag{152}$$

Definition 6 (Max-Color Extractor). *We define the operator that extracts the channel-wise maximum pixel wise on a grid $S \subseteq \Omega$ as the map $\overline{m}\colon \mathbb{R}^{K\times W\times H} \times \{0, \dots, K-1\} \times 2^\Omega \to \mathbb{R}$ with*

$$\overline{m}(x, k, S) := \max_{(i,j)\in S} \Big(\max_{(r,s)\in G_{ij}} x_{k,r,s} \Big). \tag{153}$$

Definition 7 (Max-Color Difference Extractor). *We define the operator that extracts the channel-wise maximum change in color on a grid $S \subseteq \Omega$ as the map $m_\Delta\colon \mathbb{R}^{K\times W\times H} \times \{0, \dots, K-1\} \times 2^\Omega \to \mathbb{R}$ with*

$$m_\Delta(x, k, S) := \max_{(i,j)\in S} \Big(\max_{(r,s)\in G_{ij}} x_{k,r,s} - \min_{(r,s)\in G_{ij}} x_{k,r,s} \Big). \tag{154}$$

ROTATION The rotation transformation is defined as rotating an image by an angle α counter-clock wise, followed by bilinear interpolation I. Clearly, when rotating an image, some pixels may be padded that results in a sudden change of pixel colors. To mitigate this issue, we apply black padding to all pixels that are outside the largest centered circle in a given image (see Figure 12a for an illustration). We define the rotation transformation ϕ_R as the (raw) rotation $\tilde{\phi}_R$ followed by interpolation and the aforementioned preprocessing step P so that $\phi_R = P \circ I \circ \tilde{\phi}_R$ and refer the reader to Section D.4 for details. We remark that our certification is independent of different rotation padding mechanisms, since these padded pixels are all refilled by black padding during preprocessing. The following Lemma provides a closed form expression for L in (150) for rotation. A detailed proof is given in Section D.4.

Lemma 8. *Let* $x \in \mathbb{R}^{K\times W\times H}$ *be a* K*-channel image and let* $\phi_R = P \circ I \circ \tilde{\phi}_R$ *be the rotation transformation. Then, a global Lipschitz constant* L *for the functions* $\{g_i\}_{i=1}^{N}$ *is given by*

$$L_r = \max_{1\leqslant i\leqslant N-1} \sum_{k=0}^{K-1} \sum_{r,s\in V} 2d_{r,s} \cdot m_\Delta(x, k, \mathcal{P}_{r,s}^{(i)}) \cdot \overline{m}(x,\, k,\, \mathcal{P}_{r,s}^{(i)}) \tag{155}$$

where $V = \left\{(r,s) \in \mathbb{N}^2 |\, d_{r,s} < \frac{1}{2}(\min\{W,H\}-1)\right\}$. *The set* $\mathcal{P}_{r,s}^{(i)}$ *is given by all integer grid pixels that are covered by the trajectory of source pixels of* (r,s) *when rotating from angle* α_i *to* α_{i+1}.

SCALING Computing the Lipschitz bound for the scaling transformation is similar to rotations. We provide details for scaling and the certification technique in Section D.4.

COMPUTATIONAL COMPLEXITY We provide pseudo-code for computing the bound M in Section D.5. The algorithm is composed of two main parts, namely the computation of the Lipschitz constant L, and the computation of the interpolation error bound M based on L. The former is of computational complexity $\mathcal{O}(N \cdot KWH)$, and the latter is of $\mathcal{O}(NR \cdot KWH)$, for both scaling and rotation. We note that $\mathcal{P}_{r,s}$ contains only a constant number of pixels since each interval $[\alpha_i, \alpha_{i+1}]$ is small. Thus, the bulk of costs come from the transformation operation. We improve the speed by implementing a fast and fully-parallelized `C` kernel for rotation and scaling of images. As a result, on CIFAR-10, the algorithm takes less than 2 s on average with 10 processes for rotation with $N = 556$ and $n = 200$ and the time for scaling is faster. We refer readers to Section 7.6 for detailed experimental evaluation. Also, we remark that the algorithm is model-independent. Thus, we can precompute M for test set and reuse for any models that need a certification.

7.5.4 *Discussion*

Here, we briefly summarize the computation procedure of robustness certification, introduce an acceleration strategy—progressive sampling—and discuss the extensions beyond rotation and scaling.

7.5.4.1 *Computation of Robustness Certification*

With the methodology mentioned above, for differentially resolvable transformations such as rotation and scaling, computing robustness certification follows four steps:

(1) Compute the interpolation error bound M.

(2) Generate transformed samples $\{\phi(x, \alpha_i)\}_{i=1}^{N}$.

(3) Compute $p_A^{(i)}$ and $p_B^{(i)}$ for each sample i.

(4) Verify whether $M_S < R$ holds for each sample according to Corollary 4.

7.5.4.2 *Acceleration: Progressive Sampling*

In step (3) above, we need to estimate $p_A^{(i)}$ and $p_B^{(i)}$ for each sample $\phi(x, \alpha_i)$ to check whether $M_S < R$. In the brute-force approach, to obtain a high-confidence bound on

$p_A^{(i)}$ and $p_B^{(i)}$, we typically sample $n_s = 10,000$ or more [39] then apply the binomial statistical test. In total, we thus need to sample the classifier's prediction $N \times n_s$ times, which is computationally expensive.

To accelerate the computation, we design a *progressive sampling strategy* from the following two insights: (1) we only need to check whether $R > M_S$, but are not required to compute R precisely; (2) for any sample $\phi(x, \alpha_i)$ if the check fails, the model is not certifiably robust and there is no need to proceed. Based on (1), for the current $\phi(x, \alpha_i)$, we sample n_s samples in batches and maintain high-confidence lower bound of R based on existing estimation. Once the lower bound exceeds M_S we proceed to the next $\phi(x, \alpha_{i+1})$. Based on (2), we terminate early if the check $R > M_S$ for the current $\phi(x, \alpha_i)$ fails. More details are provided in Section D.5.

7.5.4.3 *Extension to More Transformations*

For other transformations that involve interpolation, we can similarly compute the interpolation error bound using intermediate results in our above lemmas. For the composition of transformation, we extend our certification pipeline for the composition of (1) rotation/scaling with brightness, and (2) rotation/scaling with brightness and ℓ_p-bounded additive perturbations. These compositions simulate an attacker who does not precisely perform the specified transformation. We present these extensions in Section D.3.2 and Section D.3.3 in detail. In Section D.4.4 we discuss possible new transformations and extend TSS to provide the certification.

7.6 EXPERIMENTS

Here, we validate our framework TSS by certifying the robustness over semantic transformations experimentally. We compare with state of the art for each transformation, highlight our main results, and present our findings and ablation studies.

7.6.1 *Experimental Setup*

7.6.1.1 *Dataset*

Our experiments are conducted on three classical image classification datasets: MNIST, CIFAR-10, and ImageNet. For all images, the pixel color is normalized to $[0, 1]$. We follow common practice to resize and center cropping the ImageNet images to 224×224 size [39, 106, 176, 264]. To our best knowledge, we are the *first* to provide rigorous certifiable robustness against semantic transformations on the large-scale *standard* ImageNet dataset.

7.6.1.2 *Model*

The undefended model is very vulnerable even under simple random semantic attacks. Therefore, we apply existing data augmentation training [39] combined with consistency regularization [106] to train the base classifiers. We then use the introduced smoothing strategies, to obtain the models for robustness certification. On MNIST and CIFAR-10, the models are trained from scratch while on ImageNet, we either finetune undefended models in `torchvision` library or finetune from state-of-the-art certifiably

robust models against ℓ_2 perturbations [187]. Details are available in Section D.6.1. We remark that our framework focuses on robustness certification and did not fully explore the training methods for improving the certified robustness or tune the hyperparameters.

7.6.1.3 *Implementation and Hardware*

We implement our framework TSS based on `PyTorch`. We improve the running efficiency by tensor parallelism and embedding `C` modules. Details are available in Section D.6.2. All experiments were run on 24-core Intel Xeon Platinum 8259CL CPU and one Tesla T4 GPU with 15 GB RAM.

7.6.1.4 *Evaluation Metric*

On each dataset, we uniformly pick 500 samples from the test set and evaluate all results on this *test subset* following Cohen et al [39]. In line with related work [39, 106, 187, 264], we report the certified robust accuracy which is defined as the fraction of samples (within the test subset) that are both certified robust *and* classified correctly, and set the certification confidence level to $p = 0.1\%$. We use $n_s = 10^5$ samples to obtain a confidence lower bound $\underline{p_A}$ for resolvable transformations, and $n_s = 10^4$ samples to obtain each $\underline{p_A}^{(i)}$ for differentially resolvable transformations. Due to progressive sampling (algorithm 4), the actual samples used for differentially resolvable transformations are usually far fewer than n_s. In addition, we report the benign accuracy in Section D.6.5.1 defined as the fraction of correctly classified samples when no attack is present, and the empirical robust accuracy, defined as the fraction of samples in the test subset that are classified correctly under either a simple random attack (following [8, 65]) or two adaptive attacks (namely Random+ Attack and PGD Attack). We introduce all these attacks in Section D.6.3 and provide a detailed comparison in Section D.6.5.3. Note that the empirical robust accuracy under any attacks is lower bounded by the certified accuracy.

7.6.1.5 *Notations for Robust Radii*

In the tables, we use these notations: α for squared kernel radius for Gaussian blur; $\sqrt{\Delta x^2 + \Delta y^2}$ for translation distance; b and c for brightness shift and contrast change respectively as in $x \mapsto (1+c)x + b$; r for rotation angle; s for size scaling ratio; and $\|\delta\|_2$ for ℓ_2 norm of additional perturbations.

7.6.1.6 *Vanilla Models and Baselines*

We compare with vanilla (undefended) models and baselines from related work. The vanilla models are trained to achieve high accuracy only on clean data. For fairness, on all datasets we use the same model architectures as in our approach. On the test subset, the *benign accuracy* of vanilla models is 98.6%/88.6%/74.4% on MNIST/CIFAR-10/ImageNet. We also report their empirical robust accuracy under attacks in Table 5. Since vanilla models are not smoothed, we cannot have certified robust accuracy for them. In terms of baselines, we consider the approaches that provide certification against semantic transformations: DeepG [8], Interval [202], VeriVis [171], Semantify-NN [156], and DistSPT [65]. In Section D.6.4, we provide more detailed discussion and

comparison with these baseline approaches, and list how we run these approaches for fair comparison.

7.6.2 *Main Results*

Here, we present our main results from five aspects: (1) certified robustness compared to baselines; (2) empirical robustness comparison; (3) certification time statistics; (4) empirical robustness under unforeseen physical attacks; (5) certified robustness under attacks exceeding the certified radii.

7.6.2.1 *Certified Robustness Compared to Baselines*

Our results are summarized in Table 5. For each transformation, we ensure that our setting is either the same as or strictly stronger than all other baselines.[1] When our setting is strictly stronger, the baseline setting is shown in corresponding parentheses, and our certified robust accuracy implies a higher or equal certified robust accuracy in the corresponding baseline setting. To our best knowledge, we are the first to provide certified robustness for Gaussian blur, brightness, composition of rotation and brightness, etc. Moreover, on the large-scale standard ImageNet dataset, we are the first to provide nontrivial certified robustness against certain semantic attacks. Note that DistSPT [65] is theoretically feasible to provide robustness certification for the ImageNet dataset. However, its certification is not tight enough to handle ImageNet and it provides robustness certification for only a certain random attack instead of arbitrarily worst-case attacks [65, Section 7.4]. We observe that, across transformations, our framework *significantly* outperforms the state of the art, if present, in terms of robust accuracy. For example, on the composition of contrast and brightness, we improve the certified robust accuracy from 74% to 97.6% on MNIST, from 0.0% (failing to certify) to 82.4% on CIFAR-10, and from 0% (absence of baseline) to 61.4% on ImageNet. On the rotation transformation, we improve the certified robust accuracy from 92.48% to 97.4% on MNIST, from 49.37% to 63.6% on CIFAR-10 (rotation angle within 30°), and from 16% against a certain random attack to 30.4% against arbitrary attacks on ImageNet. Some baselines are able to provide certification under other certification goals and the readers can refer to Section D.6.4 for a detailed discussion.

7.6.2.2 *Comparison of Empirical Robust Accuracy*

In Table 5, we report the empirical robust accuracy for both (undefended) vanilla models and trained TSS models. The empirical robust accuracy is either evaluated under random attack or two adaptive attacks–Random+ and PGD attack. When it is under adaptive attacks, we report the lower accuracy to evaluate against stronger attackers.

1. For almost all settings, TSS models have significantly higher *empirical robust accuracy*, which means that TSS models are also practical in terms of defending against existing attacks. The only exception is rotation and scaling on ImageNet. The reason is that a single rotation/scaling transformation is too weak to attack even an undefended model. At the same time, our robustness certification comes

1 The only exception is Semantify-NN [156] on brightness and contrast changes, where Semantify-NN considers these changes composed with clipping to $[0, 1]$ while we consider pure brightness and contrast changes to align with other baselines. We refer the reader to Section D.6.4 for a detailed discussion.

Table 5: Comparison of certified robust accuracy achieved by our framework TSS and other known baselines and empirical robust accuracy achieved by TSS and vanilla models under random and adaptive attacks. "-" denotes the settings where the baselines cannot support. The parentheses show the weaker baseline settings. For certified robust accuracy, the existing state of the art is **bolded**. For empirical robust accuracy, the higher accuracy under each setting are **bolded**.

Transformation	Type	Dataset	Attack Radius	Certified Robust Accuracy						Empirical Robust Accuracy			
										Random Attack		Adaptive Attacks	
				TSS	DeepG [8]	Interval [202]	VeriVis [171]	Semantify-NN [156]	DistSPT [65]	TSS	Vanilla	TSS	Vanilla
Gaussian Blur	Resolvable	MNIST	Squared Radius $\alpha \leqslant 36$	**90.6%**	-	-	-	-	-	**91.4%**	12.2%	**91.2%**	12.2%
		CIFAR-10	Squared Radius $\alpha \leqslant 16$	**63.6%**	-	-	-	-	-	**65.8%**	3.4%	**65.8%**	3.4%
		ImageNet	Squared Radius $\alpha \leqslant 36$	**51.6%**	-	-	-	-	-	**52.8%**	8.4%	**52.6%**	8.2%
Translation (Reflection Pad.)	Resolvable, Discrete	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 8$	**99.6%**	-	-	98.8%	98.8%	-	**99.6%**	0.0%	**99.6%**	0.0%
		CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 20$	**80.8%**	-	-	65.0%	65.0%	-	**86.2%**	4.4%	**86.0%**	4.2%
		ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 100$	**50.0%**	-	-	43.2%	43.2%	-	**69.2%**	46.6%	**69.2%**	46.2%
Brightness	Resolvable	MNIST	$b \pm 50\%$	**98.2%**	-	-	-	-	-	**98.2%**	96.6%	**98.2%**	96.6%
		CIFAR-10	$b \pm 40\%$	**87.0%**	-	-	-	-	-	**87.2%**	44.4%	**87.4%**	42.6%
		ImageNet	$b \pm 40\%$	**70.0%**	-	-	-	-	-	**70.4%**	19.6%	**70.4%**	18.4%
Contrast and Brightness	Resolvable, Composition	MNIST	$c \pm 50\%, b \pm 50\%$	**97.6%**	$\leqslant 0.4\%$ ($c, b \pm 30\%$)	0.0% ($c, b \pm 30\%$)	-	$\leqslant 74\%$ ($c \pm 5\%, b \pm 50\%$)	-	**98.0%**	94.6%	**98.0%**	93.2%
		CIFAR-10	$c \pm 40\%, b \pm 40\%$	**82.4%**	0.0% ($c, b \pm 30\%$)	0.0% ($c, b \pm 30\%$)	-	-	-	**86.0%**	21.0%	**85.8%**	9.6%
		ImageNet	$c \pm 40\%, b \pm 40\%$	**61.4%**	-	-	-	-	-	**68.4%**	1.2%	**68.4%**	0.0%
Gaussian Blur, Translation, Brightness, and Contrast	Resolvable, Composition	MNIST	$\alpha \leqslant 1, \sqrt{\Delta x^2 + \Delta y^2} \leqslant 5, c, b \pm 10\%$	**90.2%**	-	-	-	-	-	**97.2%**	0.4%	**97.0%**	0.4%
		CIFAR-10	$\alpha \leqslant 1, \sqrt{\Delta x^2 + \Delta y^2} \leqslant 5, c, b \pm 10\%$	**58.2%**	-	-	-	-	-	**67.6%**	9.6%	**67.8%**	5.6%
		ImageNet	$\alpha \leqslant 10, \sqrt{\Delta x^2 + \Delta y^2} \leqslant 10, c, b \pm 20\%$	**32.8%**	-	-	-	-	-	**48.8%**	9.4%	**47.4%**	4.0%
Rotation	Differentially Resolvable	MNIST	$r \pm 50°$	**97.4%**	$\leqslant 85.8\%$ ($r \pm 30°$)	$\leqslant 6.0\%$ ($r \pm 30°$)	-	$\leqslant 92.48\%$	82%	**98.4%**	12.2%	**98.2%**	11.0%
		CIFAR-10	$r \pm 10°$	**70.6%**	62.5%	20.2%	-	-	37%	**76.6%**	65.6%	**76.4%**	65.4%
			$r \pm 30°$	**63.6%**	10.6%	0.0%	-	$\leqslant 49.37\%$	22%	**69.2%**	21.6%	**69.4%**	21.4%
		ImageNet	$r \pm 30°$	**30.4%**	-	-	-	-	16% (rand. attack)	37.8%	**40.0%**	**37.8%**	37.0%
Scaling	Differentially Resolvable	MNIST	$s \pm 30\%$	**97.2%**	85.0%	16.4%	-	-	-	**99.2%**	90.2%	**99.2%**	89.2%
		CIFAR-10	$s \pm 30\%$	**58.8%**	0.0%	0.0%	-	-	-	**67.2%**	51.6%	**67.0%**	51.2%
		ImageNet	$s \pm 30\%$	**26.4%**	-	-	-	-	-	37.4%	**50.0%**	36.4%	**49.8%**
Rotation and Brightness	Differentially Resolvable, Composition	MNIST	$r \pm 50°, b \pm 20\%$	**97.0%**	-	-	-	-	-	**98.2%**	11.0%	**98.0%**	10.4%
		CIFAR-10	$r \pm 10°, b \pm 10\%$	**70.2%**	-	-	-	-	-	**76.6%**	59.4%	**76.0%**	56.8%
			$r \pm 30°, b \pm 20\%$	**61.4%**	-	-	-	-	-	**68.4%**	13.0%	**68.2%**	9.0%
		ImageNet	$r \pm 30°, b \pm 20\%$	**26.8%**	-	-	-	-	-	**37.4%**	22.4%	**36.8%**	21.2%
Scaling and Brightness	Differentially Resolvable, Composition	MNIST	$s \pm 50\%, b \pm 50\%$	**96.6%**	-	-	-	-	-	**97.8%**	24.8%	**97.8%**	15.6%
		CIFAR-10	$s \pm 30\%, b \pm 30\%$	**54.2%**	-	-	-	-	-	**67.2%**	17.4%	**66.8%**	11.6%
		ImageNet	$s \pm 30\%, b \pm 30\%$	**23.4%**	-	-	-	-	-	**36.4%**	16.0%	**36.0%**	8.8%
Rotation, Brightness, and ℓ_2	Differentially Resolvable, Composition	MNIST	$r \pm 50°, b \pm 20\%, \|\delta\|_2 \leqslant .05$	**96.6%**	-	-	-	-	-	**97.6%**	10.8%	**97.4%**	9.0%
		CIFAR-10	$r \pm 10°, b \pm 10\%, \|\delta\|_2 \leqslant .05$	**64.2%**	-	-	-	-	-	**71.6%**	31.8%	**71.2%**	29.6%
			$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leqslant .05$	**55.2%**	-	-	-	-	-	**65.2%**	0.8%	**64.0%**	0.4%
		ImageNet	$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leqslant .05$	**26.6%**	-	-	-	-	-	**37.0%**	17.6%	**36.4%**	14.0%
Scaling, Brightness, and ℓ_2	Differentially Resolvable, Composition	MNIST	$s \pm 50\%, b \pm 50\%, \|\delta\|_2 \leqslant .05$	**96.4%**	-	-	-	-	-	**97.6%**	22.2%	**97.6%**	12.2%
		CIFAR-10	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leqslant .05$	**51.2%**	-	-	-	-	-	**65.0%**	4.4%	**61.8%**	2.6%
		ImageNet	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leqslant .05$	**22.6%**	-	-	-	-	-	**36.0%**	7.4%	**35.6%**	4.8%

at the cost of benign accuracy, which also affects the empirical robust accuracy. This exception is eliminated when rotation and scaling are composed with other transformations.

2. Similar observations arise when comparing the *empirical robust accuracy of the vanilla model with the certified robust accuracy of ours*. Hence, even compared to *empirical* metrics, our *certified* robust accuracy is nontrivial and guarantees high accuracy.

3. Our *certified* robust accuracy is always lower or equal compared to the *empirical* one, verifying the validity of our robustness certification. The gaps range from $\sim 2\%$ on MNIST to $\sim 10\%$ - 20% on ImageNet. Since empirical robust accuracy is an upper bound of the certified accuracy, this implies that our certified bounds are usually tight, particularly on small datasets.

4. The adaptive attack decreases the empirical accuracy of TSS models *slightly*, while it decreases that of vanilla models significantly. Taking contrast and brightness on CIFAR-10 as example, TSS accuracy decreases from 86% to 85.8% while the vanilla model accuracy decreases from 21.0% to 9.6%. Thus, TSS is still robust against adaptive attacks. Indeed, TSS has robustness guarantee against any attack within the certified radius.

7.6.2.3 *Certification Time Statistics*

Our robustness certification time is usually less than 100 s on MNIST and 200 s on CIFAR-10; on ImageNet it is around 200 s - 2000 s. Compared to other baselines, ours is slightly faster and achieves much higher certified robustness. For fairness, we give 1000 s time limit per instance when running baselines on MNIST and CIFAR-10. Note that other baselines cannot scale up to ImageNet. Our approach is scalable due to the blackbox nature of smoothing-based certification, the tight interpolation error upper bound, and the efficient progressive sampling strategy. Details on hyperparameters including smoothing variance and average certification time are given in Section D.6.5.2.

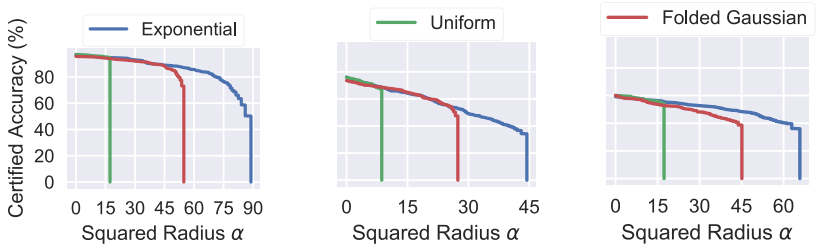
7.6.2.4 *Generalization to Unforeseen Common Corruptions*

Are TSS models still more robust when it comes to potential unforeseen physical attacks? To answer this question, we evaluate the robustness of TSS models on the realistic CIFAR-10-C and ImageNet-C datasets [91]. These two datasets are comprised of corrupted images from CIFAR-10 and ImageNet. They apply around 20 types of common corruptions to model *physical attacks*, such as fog, snow, and frost. We evaluate the *empirical robust accuracy* against the highest corruption level (level 5) to model the strongest physical attacker. We apply TSS models trained against a transformation composition attack, Gaussian blur + brightness + contrast + translation, to defend against these corruptions. We select two baselines: vanilla models and AugMix [92]. AugMix is the state of the art model on CIFAR-10-C and ImageNet-C [43].

The results are shown in Table 6. The answer is *yes*—TSS models are more robust than undefended vanilla models. It even exceeds the state of the art, AugMix, on CIFAR-10-C. On ImageNet-C, the empirical accuracy of the TSS model is between vanilla and AugMix. We emphasize that in contrast to TSS, both vanilla and AugMix fail to provide

Table 6: Comparison of empirical accuracy of different models under physical corruptions (CIFAR-10-C and ImageNet-C) and certified accuracy against composition of transformations. TSS achieves higher or comparable empirical accuracy against unforeseen corruptions and significantly higher certified accuracy (under attack radii in Table 5).

	CIFAR-10			ImageNet		
	Vanilla	AugMix [92]	TSS	Vanilla	AugMix [92]	TSS
Empirical Accuracy on CIFAR-10-C and ImageNet-C	53.9%	65.6%	**67.4%**	18.3%	**25.7%**	21.9%
Certified Accuracy against Composition of Gaussian Blur, Translation, Brightness, and Contrast	0.0%	0.4%	**58.2%**	0.0%	0.0%	**32.8%**



(a) MNIST (b) CIFAR-10 (c) ImageNet

Figure 15: Certified accuracy for different smoothing distributions for Gaussian blur. On MNIST/CIFAR-10/ImageNet the noise std. is 10/5/10.

robustness certification. Details on evaluation protocols and additional findings are in Section D.6.5.4.

7.6.2.5 *Evaluation on Attacks Beyond Certified Radii*

The semantic attacker in the physical world may not constrain itself to be within the specified attack radii. In Section D.6.5.5 we present a thorough evaluation of TSS's robustness when the attack radii go beyond the certified ones. We show, for example, for TSS model defending against ±40% brightness change on ImageNet, when the radius increases to 50%, the certified accuracy only slightly drops from 70.4% to 70.0%. In a nutshell, there is no significant or immediate degradation on both certified robust accuracy and empirical robust accuracy when the attack radii go beyond the certified ones.

7.6.3 *Ablation Studies*

Here, we provide two ablation studies: (1) Comparison of different smoothing distributions; (2) Comparison of different smoothing variances. In Section D.6.5.7, we present another ablation study on different numbers of samples for differentially resolvable transformations, which reveals a tightness-efficiency trade-off.

Table 7: Study of the impact of different smoothing variance levels on certified robust accuracy and benign accuracy on ImageNet for TSS. The attack radii are consistent with Table 5. `Dist.` refers to both training and smoothing distribution.

Transformation	Attack Radii	Certified Accuracy and Benign Accuracy under Different Variance Levels			
Gaussian Blur	$\alpha \leqslant 36$	Dist. of α	Exp(1/5)	Exp(1/10)	Exp(1/20)
		Cert. Rob. Acc.	0.0%	**51.6%**	48.4%
		Benign Acc.	**63.4%**	59.2%	53.2%
Translation (Reflection Pad.)	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 100$	Dist. of $(\Delta x, \Delta y)$	$\mathcal{N}(0, 20^2 I)$	$\mathcal{N}(0, 30^2 I)$	$\mathcal{N}(0, 40^2 I)$
		Cert. Rob. Acc.	0.0%	50.0%	**55.4%**
		Benign Acc.	70.0%	**72.6%**	70.0%
Brightness	$b \pm 40\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.3^2 I)$	$\mathcal{N}(0, 0.4^2 I)$	$\mathcal{N}(0, 0.5^2 I)$
		Cert. Rob. Acc.	**70.2%**	70.0%	67.6%
		Benign Acc.	**73.2%**	72.2%	69.4%
Contrast	$c \pm 40\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.3^2 I)$	$\mathcal{N}(0, 0.4^2 I)$	$\mathcal{N}(0, 0.5^2 I)$
		Cert. Rob. Acc.	58.4%	63.6%	**65.0%**
		Benign Acc.	**72.8%**	71.4%	68.6%
Rotation	$r \pm 30°$	Dist. of ϵ	$\mathcal{N}(0, 0.25^2 I)$	$\mathcal{N}(0, 0.50^2 I)$	$\mathcal{N}(0, 1.00^2 I)$
		Cert. Rob. Acc.	9.8%	**30.4%**	20.0%
		Benign Acc.	**55.6%**	46.2%	32.2%
Scaling	$s \pm 30\%$	Dist. of ϵ	$\mathcal{N}(0, 0.25^2 I)$	$\mathcal{N}(0, 0.50^2 I)$	$\mathcal{N}(0, 1.00^2 I)$
		Cert. Rob. Acc.	7.2%	**26.4%**	17.4%
		Benign Acc.	**58.8%**	50.8%	33.8%

7.6.3.1 *Comparison of Smoothing Distributions*

To study the effects of different smoothing distributions, we compare the certified robust accuracy for Gaussian blur when the model is smoothed by different smoothing distributions. We consider three smoothing distributions, namely exponential (blue line), uniform (green line), and folded Gaussian (red line). On each dataset, we adjust the distribution parameters such that each distribution has the same variance. All other hyperparameters are kept the same throughout training and certification. As shown in Figure 15, we notice that on all three datasets, the exponential distribution has the highest average certified radius. This observation is in line with our theoretical reasoning in Section 7.4.2.

7.6.3.2 *Comparison of Different Smoothing Variances*

The variance of the smoothing distribution is a hyperparameter that controls the accuracy-robustness trade-off. In Table 7, we evaluate different smoothing variances for several transformations on ImageNet and report both the certified accuracy and benign accuracy. The results on MNIST and CIFAR-10 and more discussions are in Section D.6.5.6. From these results, we observe that usually, when the smoothing variance increases, the benign accuracy drops and the certified robust accuracy first rises and then drops. This tendency is also observed in classical randomized smoothing [39, 264]. However, the range of acceptable variance is usually wide. Thus, even without carefully tuning

the smoothing variances, we are able to achieve high certified and benign accuracy as reported in Table 5 and Table 18.

7.7 CONCLUSION

In this chapter, in light of the second research question governing this thesis, we have provided a unified framework, TSS, to certify the robustness against semantic adversarial transformations which incur large ℓ_p-norm perturbations and can thus not be handled by standard additive smoothing approaches. In extensive experiments, we have shown that TSS significantly outperforms the state of the art or, if no previous work exists, sets new baselines. Since the approach is based on randomized smoothing, it incurs the known computational overhead for estimating expectation via Monte-Carlo sampling. In addition, while we have provided efficient implementations, the computation of the Lipschitz constant of transformations incurs additional computational burden. We believe that optimizing these approaches for efficiency and tightness is fruitful ground for future research. This could, for example, also involve the exploration of further transformation-specific smoothing strategies.

7.8 PROOFS

7.8.1 *Proof of Theorem 5*

Let us recall the following definition from Section 7.3:

Definition 2 (restated). *Let $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a transformation, $\varepsilon \sim \mathbb{P}_\varepsilon$ a random variable taking values in $\mathcal{Z}$ and let $h\colon \mathcal{X} \to \mathcal{Y}$ be a base classifier. We define the ε-smoothed classifier $g\colon \mathcal{X} \to \mathcal{Y}$ as $g(x;\varepsilon) = \arg\max_{y\in\mathcal{Y}} q(y|x;\varepsilon)$ where q is given by the expectation with respect to the smoothing distribution ε, i.e.,*

$$q(y|x;\varepsilon) := \mathbb{E}(p(y|\phi(x,\varepsilon))). \tag{156}$$

Here, we additionally define the notion of level sets separately. These sets originate from statistical hypothesis testing and correspond to rejection regions of likelihood ratio tests.

Definition 8 (Lower level sets). *Let $\varepsilon_0 \sim \mathbb{P}_0$, $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ. For $t \geqslant 0$ we define lower and strict lower level sets as*

$$\underline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) < t\}, \quad \overline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) \leqslant t\}, \ \text{where}\ \Lambda(z) := \frac{f_1(z)}{f_0(z)}. \tag{157}$$

Lemma 9. *Let ε_0 and ε_1 be random variables taking values in $\mathcal{Z}$ and with probability density functions f_0 and f_1 with respect to a measure μ. Let $h\colon \mathcal{Z} \to [0,1]$ be a determinstic function. Then, for any $t \geqslant 0$ the following implications hold:*

(i) For any $S \subseteq \mathcal{Z}$ with $\underline{S_t} \subseteq S \subseteq \overline{S_t}$ the following implication holds:

$$\mathbb{E}[h(\varepsilon_0)] \geqslant \mathbb{P}_0(S) \Rightarrow \mathbb{E}[h(\varepsilon_1)] \geqslant \mathbb{P}_1(S). \tag{158}$$

(i) For any $S \subseteq \mathcal{Z}$ with $\overline{S_t}^c \subseteq S \subseteq \underline{S_t}^c$ the following implication holds:

$$\mathbb{E}[h(\varepsilon_0)] \leqslant \mathbb{P}_0(S) \Rightarrow \mathbb{E}[h(\varepsilon_1)] \leqslant \mathbb{P}_1(S). \tag{159}$$

Proof. We first prove (i). For that purpose, consider

$$\begin{aligned}
\mathbb{E}[h(\varepsilon_1)] - \mathbb{P}_1(S) &= \int h f_1 \, d\mu - \int_S f_1 \, d\mu && (160)\\
&= \int_{S^c} h f_1 \, d\mu - \left(\int_S (1-h) f_1 \, d\mu \right) && (161)\\
&= \int_{S^c} h \wedge f_0 \, d\mu - \left(\int_S (1-h) \wedge f_0 \, d\mu \right) && (162)\\
&\geqslant t \cdot \int_{S^c} h f_0 \, d\mu - t \cdot \left(\int_S (1-h) f_0 \, d\mu \right) && (163)\\
&= t \cdot \left(\int h f_0 \, d\mu - \int_S f_0 \, d\mu \right) && (164)\\
&= t \cdot (\mathbb{E}[h(\varepsilon_0)] - \mathbb{P}_0(S)) \geqslant 0. && (165)
\end{aligned}$$

The inequality in (163) follows from the fact that whenever $z \in S^c$, then $f_1(z) \geqslant t \cdot f_0(z)$ and if $z \in S$, then $f_1(z) \leqslant t \cdot f_0(z)$ since S is a lower level set. Finally, the inequality in (165) follows from the assumption. The proof of (ii) is analogous and omitted here. □

Theorem 5 (restated). *Let $\varepsilon_0 \sim \mathbb{P}_0$ and $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ on $\mathcal{Z}$ and let $\phi \colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a semantic transformation. Suppose that $y_A = g(x; \varepsilon_0)$ and let $p_A, p_B \in [0, 1]$ be bounds to the class probabilities, i.e.,*

$$q(y_A | x, \epsilon_0) \geqslant p_A > p_B \geqslant \max_{y \neq y_A} q(y | x, \epsilon_0). \tag{166}$$

For $t \geqslant 0$, let $\underline{S}_t, \overline{S}_t \subseteq \mathcal{Z}$ be the sets defined as $\underline{S}_t := \{f_1/f_0 < t\}$ and $\overline{S}_t := \{f_1/f_0 \leqslant t\}$ and define the function $\xi \colon [0, 1] \to [0, 1]$ by

$$\begin{aligned}
&\xi(p) := \sup\{\mathbb{P}_1(S) \colon \underline{S}_{\tau_p} \subseteq S \subseteq \overline{S}_{\tau_p}\}\\
&\text{where} \quad \tau_p := \inf\{t \geqslant 0 \colon \mathbb{P}_0(\overline{S}_t) \geqslant p\}.
\end{aligned} \tag{167}$$

Then, if the condition

$$\xi(p_A) + \xi(1 - p_B) > 1 \tag{168}$$

is satisfied, then it is guaranteed that $g(x; \varepsilon_1) = g(x; \varepsilon_0)$.

Proof. For ease of notation, let ζ be the function defined by

$$t \mapsto \zeta(t) := \mathbb{P}_0(\overline{S}_t) \tag{169}$$

and notice that $\tau_p = \zeta^{-1}(p)$ where ζ^{-1} denotes the generalized inverse of ζ. Furthermore, let $\tau_A := \tau_{p_A}$, $\tau_B := \tau_{1-p_B}$, $\underline{S}_A := \underline{S}(\tau_A)$, $\underline{S}_B := \underline{S}(\tau_B)$, $\overline{S}_A := \overline{S}(\tau_A)$ and $\overline{S}_B := \overline{S}(\tau_B)$. We first show that $q(y_A | x, \varepsilon_1)$ is lower bounded by $\xi(\tau_A)$. For that purpose, note that by Lemma 12 we have that $\zeta(\tau_A) = \mathbb{P}_0(\overline{S}_A) \geqslant p_A \geqslant \mathbb{P}_0(\underline{S}_A)$. Thus, the collection of sets

$$\mathcal{S}_A := \{S \subseteq \mathcal{Z} \colon \underline{S}_A \subseteq S \subseteq \overline{S}_A, \mathbb{P}_0(S) \leqslant p_A\} \tag{170}$$

is not empty. Pick some $A \in \mathcal{S}_A$ arbitrary and note that, since by assumption $g(\cdot; \varepsilon_0)$ is (p_A, p_B)-confident at x it holds that

$$\mathbb{E}(p(y_A|\phi(x, \varepsilon_0))) = q(y_A|x; \varepsilon_0) \geqslant p_A \geqslant \mathbb{P}_0(A). \tag{171}$$

Since $\underline{S}_A \subseteq A \subseteq \overline{S}_A$ we can apply part (i) of Lemma 9 and obtain the lower bound

$$q(y_A|x, \varepsilon_1) = \mathbb{E}(p(y_A|\phi(x, \varepsilon_1))) \geqslant \mathbb{P}_1(A). \tag{172}$$

Since $A \in \mathcal{S}_A$ was arbitrary, we take the sup over all $A \in \mathcal{S}_A$ and obtain

$$q(y_A|x; \varepsilon_1) \geqslant \sup_{A \in \mathcal{S}_A} \mathbb{P}_1(A) = \xi(p_A) \tag{173}$$

We now show that for any $y \neq y_A$ the prediction $q(y|x; \varepsilon_1)$ is upper bounded by $1 - \xi(1 - p_B)$. For that purpose, note that by Lemma 12 we have that $\zeta(\tau_B) = \mathbb{P}_0(\overline{S}_A) \geqslant 1 - p_B \geqslant \mathbb{P}_0(\underline{S}_B)$. Thus, the collection of sets

$$\mathcal{S}_B := \{S \subseteq \mathcal{Z} : \underline{S}_B \subseteq S \subseteq \overline{S}_B, \mathbb{P}_0(S) \leqslant 1 - p_B\} \tag{174}$$

is not empty. Pick some $B \in \mathcal{S}_A$ arbitrary and note that, since by assumption $g(\cdot; \varepsilon_0)$ is (p_A, p_B)-confident at x it holds that

$$\begin{aligned} \mathbb{E}(p(y|\phi(x, \varepsilon_0))) &= q(y|x; \varepsilon_0) \leqslant p_B \\ &= 1 - (1 - p_B) \leqslant 1 - \mathbb{P}_0(B). \end{aligned} \tag{175}$$

Since $\underline{S}_B^c \subseteq B^c \subseteq \overline{S}_B^c$ we can apply part (ii) of Lemma 9 and obtain the upper bound

$$q(y|x; \varepsilon_1) = \mathbb{E}(p(y|\phi(x, \varepsilon_1))) \leqslant 1 - \mathbb{P}_1(B). \tag{176}$$

Since $B \in \mathcal{S}_B$ was arbitrary, we take the inf over all $B \in \mathcal{S}_B$ and obtain

$$q(y|x; \varepsilon_1) \leqslant \inf_{B \in \mathcal{S}_B} (1 - \mathbb{P}_1(B)) = 1 - \xi(1 - p_B). \tag{177}$$

Combining together (177) and (173), we find that, whenever

$$\xi(p_A) + \xi(1 - p_B) > 1 \tag{178}$$

it is guaranteed that

$$q(y_A|x; \varepsilon_1) > \max_{y \neq y_A} q(y_|x; \varepsilon_1) \tag{179}$$

what concludes the proof. □

7.8.2 *Proof of Theorem 6*

Here we provide the proof for Theorem 6, which is the foundation for the techniques to certify differentially resolvable transformations. First, let us recall the definition of differentially resolvable transformations.

Definition 4 (restated). *Let $\phi : \mathcal{X} \times \mathcal{Z}_\phi \to \mathcal{X}$ be a transformation with noise space $\mathcal{Z}_\phi$ and let $\psi : \mathcal{X} \times \mathcal{Z}_\psi \to \mathcal{X}$ be a resolvable transformation with noise space $\mathcal{Z}_\psi$. We say that ϕ can be resolved by ψ if for any $x \in \mathcal{X}$ there exists function $\delta_x : \mathcal{Z}_\phi \times \mathcal{Z}_\phi \to \mathcal{Z}_\psi$ such that for any $\beta \in \mathcal{Z}_\phi$*

$$\phi(x, \alpha) = \psi(\phi(x, \beta), \delta_x(\alpha, \beta)). \tag{180}$$

Theorem 6 (restated). *Let $\phi\colon \mathcal{X} \times \mathcal{Z}_\phi \to \mathcal{X}$ be a transformation that is resolved by $\psi\colon \mathcal{X} \times \mathcal{Z}_\psi \to \mathcal{X}$. Let $\varepsilon \sim \mathbb{P}_\varepsilon$ be a $\mathcal{Z}_\psi$-valued random variable and suppose that the smoothed classifier $g : \mathcal{X} \to \mathcal{Y}$ given by $q(y|x;\varepsilon) = \mathbb{E}(p(y|\psi(x,\varepsilon)))$ predicts $g(x;\varepsilon) = y_A = \arg\max_y q(y|x;\varepsilon)$. Let $\mathcal{S} \subseteq \mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$ be a set of transformation parameters such that for any i, the class probabilities satisfy*

$$q(y_A|\phi(x,\alpha_i);\varepsilon) \geqslant p_A^{(i)} \geqslant p_B^{(i)} \geqslant \max_{y \neq y_A} q(y|\phi(x,\alpha_i);\varepsilon). \tag{181}$$

Then there exists a set $\Delta^* \subseteq \mathcal{Z}_\psi$ with the property that, if for any $\alpha \in \mathcal{S}$, $\exists \alpha_i$ with $\delta_x(\alpha,\alpha_i) \in \Delta^*$, then it is guaranteed that

$$q(y_A|\phi(x,\alpha);\varepsilon) > \max_{y \neq y_A} q(y|\phi(x,\alpha);\varepsilon). \tag{182}$$

Proof. We prove the theorem by explicitly constructing a region Δ^* with the desired property by applying Theorem 5. For that purpose let $\delta \in \mathcal{Z}_\psi$ and denote by $\gamma_\delta\colon \mathcal{Z}_\psi \to \mathcal{Z}_\psi$ the resolving function of ψ, i.e.,

$$\psi(\psi(x,\delta),\delta') = \psi(x,\gamma_\delta(\delta')). \tag{183}$$

Let $\mathbb{P}_\gamma$ be the distribution of the random variable $\gamma := \gamma_\delta(\varepsilon)$ with density function f_γ and let

$$\underline{S}_t = \{z \in \mathcal{Z}_\psi\colon \Lambda(z) < t\}, \quad \overline{S}_t = \{z \in \mathcal{Z}_\psi\colon \Lambda(z) \leqslant t\} \tag{184}$$

$$\text{where} \quad \Lambda(z) = \frac{f_\gamma(z)}{f_\varepsilon(z)}. \tag{185}$$

Furthermore, recall the definition of the function $\zeta\colon \mathbb{R}_{\geqslant 0} \to [0,1]$ that is given by $t \mapsto \zeta(t) := \mathbb{P}_\varepsilon(\overline{S}_t)$ with generalized inverse $\zeta^{-1}(p) := \inf\{t \geqslant 0\colon \zeta(t) \geqslant p\}$. For $t \geqslant 0$ and the function $\xi\colon [0,1] \to [0,1]$ is given by by

$$\xi(p) := \sup\{\mathbb{P}_\gamma(S)\colon \underline{S}_{\zeta^{-1}(p)} \subseteq S \subseteq \overline{S}_\zeta^{-1}(p), \mathbb{P}_\varepsilon(S) \leqslant p\}. \tag{186}$$

By assumption, for every $i = 1,\ldots,n$, the ε-smoothed classifier g is $(p_A^{(i)}, p_B^{(i)})$-confident at $\phi(x,\alpha_i)$. Identify $\Delta_i \subseteq \mathcal{Z}_\psi$ with the set of perturbations that satisfy the robustness condition (125) in Theorem 5, i.e.,

$$\Delta_i \equiv \{\delta \in \mathcal{Z}_\psi\colon 1 - \xi(1-p_B^{(i)}) < \xi(p_A^{(i)})\}. \tag{187}$$

Thus, by Theorem 5, we have that for any $\delta \in \Delta_i$

$$q(y_A|\psi(\phi(x,\alpha_i),\delta);\varepsilon) > \max_{y \neq y_A} q(y|\psi(\phi(x,\alpha_i),\delta);\varepsilon). \tag{188}$$

Finally, note that for the set

$$\Delta^* \equiv \bigcap_{i=1}^N \Delta_i \tag{189}$$

it holds that, if for $\alpha \in \mathcal{S}$ there exists α_i with $\delta_x(\alpha,\alpha_i) \in \Delta^*$, then in particular $\delta_x(\alpha,\alpha_i) \in \Delta_i$ and hence, by Theorem 5 it is guaranteed that

$$q(y_A|\phi(x,\alpha);\varepsilon) = q(y_A|\psi(\phi(x,\alpha_i),\delta_x(\alpha,\alpha_i));\varepsilon) \tag{190}$$

$$> \max_{y \neq y_A} q(y|\psi(\phi(x,\alpha_i),\delta_x(\alpha,\alpha_i));\varepsilon) \tag{191}$$

$$= \max_{y \neq y_A} q(y|\phi(x,\alpha);\varepsilon) \tag{192}$$

what concludes the proof. □

8 CERTIFYING OUT-OF-DOMAIN GENERALIZATION

In the preceding two chapters, we have taken an instance-level view on robustness certification and presented techniques to certify the robustness for specific test instances. In the case of backdoor certification (Chapter 6) we considered an instance to be composed of a training set and a test sample and showed a robustness certificate based on the magnitude of the backdoor pattern. For the certification of perturbations arising from semantic transformations (Chapter 7), an instance is considered to be a single test sample that has been perturbed using transformations such as rotations, Gaussian blur, and others.

In this chapter, we take a different view and certify the distributional robustness of ML models. In other words, in contrast to certifying the robustness for a specific test sample x, here we are interested in how performance measures based on the entire population of possible inputs behave when the underlying distribution changes. In its most general form, here we are interested in studying the behaviour of expectation values $\mathbb{E}_{Z\sim P}[\ell(Z)]$ under shifts to the distribution $P \to Q$. In the context of ML, this shift can for example model changes in training and testing distributions, or changes in the distribution of subpopulations. Here we focus on the former and develop theoretical tools to certify out-of-domain generalization.

8.1 INTRODUCTION

8.1.1 *Overview*

The wide application of machine learning models in the real world brings an emerging challenge of understanding the performance of an ML model under different data distributions — ML systems operating autonomous vehicles which are trained based on data collected in the northern hemisphere might fail when deployed in desert-like environments or under different weather conditions [44, 229], while recognition systems have been shown to fail when deployed in new environments [9]. Similar concerns also apply to many mission-critical applications such as medicine and cyber-security [3, 78, 113]. In all these applications, it is imperative to have a sound understanding of the model's robustness and possible failure cases in the presence of a shift in the data distribution, and to have corresponding guarantees on the performance.

Recently, this problem has attracted intensive interest under the umbrella of *distributional robustness* [10, 16, 52, 66, 121, 190]. Specifically, let P be a joint data distribution over features $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y}$, and let $h_\theta : \mathcal{X} \to \mathcal{Y}$ be an ML model parameterized by θ. For a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we hope to compute

$$\mathcal{R}_\theta(\mathcal{U}_P) := \sup_{Q \in \mathcal{U}_P} \mathbb{E}_{(X,Y)\sim Q}[\ell(h_\theta(X), Y)] \tag{193}$$

where $\mathcal{U}_P \subseteq \mathcal{P}(\mathcal{Z})$ is a set of probability distributions on $\mathcal{Z}$, called the *uncertainty set*. Intuitively, this measures the *worst-case risk* of h_θ when the data distribution drifts from P to another distribution in $\mathcal{U}_P$.

Table 8: Current landscape of certified distributional robustness.

Ref.	Assumptions on ℓ	Assumption on h	Distance	Largest Dataset
Gao et al. [66]	Generalised Lipschitz Continuity		Wasserstein	–
Sinha et al. [203]	Bounded, Smoothness	Smoothness	Wasserstein	MNIST
Staib et al. [205]	Bounded, Continuous	Kernel Methods	MMD	–
Shafieezadeh-Abadeh et al. [198]	Lipschitz Continuity		Wasserstein	–
Blanchet et al. [16]	Bounded, Smoothness	Smoothness	Wasserstein	–
Cranko et al. [42]	Generalised Lipschitz Continuity		Wasserstein	–
Ours	Bounded	*any* Blackbox	Hellinger	ImageNet

Providing a technical solution to this problem has gained increased attention over the years, as summarized in Table 8. However, most — if not all — existing approaches impose strong constraints such as bounded Lipschitz gradients on both h and ℓ and rely on expensive certification methods such as direct minimax optimization. As a result, these methods have been applied only to small-scale datasets and ML models.

8.1.2 *Contributions*

Here, we consider the case that both h *and* ℓ *can be non-convex and non-smooth* — h can be a full-fledged neural network, e.g., ImageNet-scale EfficientNet-B7 [214], and ℓ can be a general non-smooth loss function such as the 0-1 loss. We provide, to our best knowledge, the first practical method for blackbox functions that scales to real-world, ImageNet-scale neural networks and datasets. Our key innovation is a novel framework that arises from bounding inner products between elements of a suitable Hilbert space. Specifically, we can characterize the upper bound of the performance of h on any Q within the uncertainty set as a function of the Hellinger distance, a specific type of f-divergence, and the expectation and variance of the loss of h on P.

We then apply our framework to the problem of certifying the out-of-domain generalization performance of a given classifier, taking advantage of its scalability and flexibility. Specifically, let P be the *in-domain distribution*, and h_θ a classifier. Then, to reason about the performance of h_θ on shifted distributions Q, we provide a certificate in the following form:

$$\forall Q\colon \mathtt{dist}(Q,P) \leqslant \rho \implies \mathbb{E}_{(X,Y)\sim Q}[\ell(h(X),Y)] \leqslant C_\ell(\rho, P) \tag{194}$$

where C_ℓ is a bound which depends on the distance ρ and the distribution P. This requires several nontrivial instantiations of our framework with careful practical considerations. To this end, we first develop a certification algorithm that relies only on a finite set of samples from the in-domain distribution P. Moreover, we also instantiate it with different domain drifting models such as label drifting and covariate drifting, connecting the general Hellinger distance to the degree of domain drifting specific to these scenarios. We then consider a diverse range of loss functions, including JSD loss, 0-1 loss, and AUC. To the best of our knowledge, we provide the first certificate for such diverse realistic scenarios, which is able to scale to large problems.

Last but not least, we conduct extensive experiments verifying the efficiency and effectiveness of our result. Our method is able to scale to datasets and neural networks as large as ImageNet and full-fledged models like EfficientNet-B7 and BERT. We further

apply our method on smaller-scale datasets, in order to compare with strong, state-of-the-art methods. We show that our method provides much tighter certificates.

Our contributions can be summarized as follows:

- We present a novel framework which provides a non-vacuous, computationally tractable bound to the distributionally robust worst-case risk $\mathcal{R}_\theta(\mathcal{U}_P)$ for general bounded loss functions ℓ and models h.
- We apply this framework to the problem of certifying out-of-domain generalization for blackbox functions and provide a means to certify distributional robustness in specific scenarios such as label and covariate drifts.
- We provide an extensive experimental study of our approach on a wide range of datasets including the large scale ImageNet [185] dataset, as well as NLP datasets with complex models.

8.1.3 *Related Work*

Distributionally robust optimization first appeared in the context of inventory management [190] and has since been discovered by the machine learning community as a useful tool to train machine learning models which generalize better to new distributions [10, 66, 198]. The uncertainty set occurring in the distributionally robust loss has been studied in terms of Wasserstein balls in [16, 37, 42, 66, 121, 129, 198, 203], and f-divergence balls in [10, 52–54, 126]. From a more general viewpoint, [103] connects integral probability metrics with distributional robustness in general and provides links with generative adversarial networks. In another vein, maximum mean discrepancy measures have been investigated in [205] for generalization in Kernel methods. [203] propose a method to certify generalization by using the dual formulation of the Wasserstein worst-case risk. However, their approach requires the model and loss function to be smooth and relies on an estimate of the Lipschitz constant of gradients, which quickly becomes vacuous for large problem sizes. Related techniques based on Wasserstein distances [16, 42, 66, 121, 198] make similarly prohibitive assumptions and generally fail to provide scalable alternatives. In contrast, we study uncertainty sets expressed as Hellinger balls and provide a model-specific distributional robustness guarantee which only makes minimal assumptions on the loss (namely, boundedness) and thus scales to large problems. The authors in [208] consider distributionally robust optimization under fine-grained shifts in the marginal distributions, and reason about the worst-case risk on subpopulations in the data distribution. Orthogonal to our work is the topic of certified robustness at the instance level [27, 39, 56, 128, 213, 253]. This line of research seeks to reason about robustness against adversarial attacks at the instance level, while here we aim to bound the worst-case population-level risk over a set of distributions.

8.2 DISTRIBUTIONAL ROBUSTNESS FOR BLACKBOX FUNCTIONS

In this section, we present the main results of this chapter, namely, a computationally tractable upper bound to the worst-case risk (193) for uncertainty sets expressed in terms of Hellinger balls around the data-generating distribution P. The result follows as a direct consequence of Theorem 1 where we have shown lower and upper bounds

to expectation values using properties of Gram matrices in an appropriately chosen Hilbert space structure. Here we instantiate this generic Theorem in the context of bounding loss and score functions of ML models under distribution shifts.

For the remainder of this section, to simplify notation and maintain generality, we consider loss functions $\ell\colon \mathcal{Z} \to \mathbb{R}_+$ which include the model h and take inputs from an input space $\mathcal{Z}$. For example in the context of supervised learning, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ can be the product space of features and labels and the loss $\ell(z) = \tilde{\ell}(h_\theta(x), y)$ can be seen as a composition of the loss function $\tilde{\ell}$ and the model h_θ. We denote the set of probability measures on the space $\mathcal{Z}$ by $\mathcal{P}(\mathcal{Z})$. For two measures μ, ν on $\mathcal{Z}$, we say that ν is absolutely continuous with respect to μ, denoted $\nu \ll \mu$, if $\mu(A) = 0$ implies that $\nu(A) = 0$ for any measurable set $A \subseteq \mathcal{Z}$. Among the plethora of distances between probability measures, such as total variation and Wasserstein distance, a particularly popular choice is the family of f-divergences which has been extensively studied in the context of distributionally robust optimization [10, 52, 54, 126]. Here we focus on the Hellinger distance, a particular f-divergence, as it emerges naturally from the derivation of the bounds.

Definition 9 (Hellinger-distance). *Let $P, Q \in \mathcal{P}(\mathcal{Z})$ be probability measures on $\mathcal{Z}$ that are absolutely continuous with respect to a reference measure μ with $P, Q \ll \mu$. The Hellinger distance between P and Q is defined as*

$$H(P, Q) := \sqrt{\frac{1}{2}\int_{\mathcal{Z}} \left(\sqrt{p(z)} - \sqrt{q(z)}\right)^2 \, d\mu(z)} \tag{195}$$

where $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ are the Radon-Nikodym derivatives of P and Q with respect to μ. The Hellinger distance is independent of the choice of the reference measure μ.

The Hellinger distance is bounded with values in $[0, 1]$, with $H(P, Q) = 0$ if and only if $P = Q$ and the maximum value of 1 attained when P and Q have disjoint support. Furthermore, H defines a metric on the space of probability measures and hence satisfies the triangle inequality. In the following Theorem, we state our main result as an upper bound to the worst-case risk (193) and refer the reader to Section E.2 for the analogous lower bound.

Theorem 7. *Let $\ell\colon \mathcal{Z} \to \mathbb{R}_+$ be a loss function and suppose that $\sup_{z\in\mathcal{Z}} |\ell(z)| \leqslant M$ for some $M > 0$. Then, for any probability measure P on $\mathcal{Z}$ and $\rho > 0$ we have*

$$\begin{aligned}\sup_{Q\in B_\rho(P)} \mathbb{E}_Q[\ell(Z)] \leqslant \mathbb{E}_P[\ell(Z)] &+ 2\sqrt{C_\rho(1-C_\rho)\mathbb{V}_P[\ell(Z)]} \\ &+ C_\rho\left[\frac{\mathbb{E}_P[M-\ell(Z)]^2 - \mathbb{V}_P[\ell(Z)]}{M - \mathbb{E}_P[\ell(Z)]}\right]\end{aligned} \tag{196}$$

where $C_\rho = \rho^2(2-\rho^2)$ and $B_\rho(P) = \{Q \in \mathcal{P}(\mathcal{Z})\colon H(P, Q) \leqslant \rho\}$ is the Hellinger ball of radius ρ centered at P. The radius ρ is required to be small enough such that

$$\rho^2 \leqslant 1 - \sqrt{\frac{\mathbb{V}_P[\ell(Z)]}{\mathbb{V}_P[\ell(Z)] + \mathbb{E}_P[M-\ell(Z)]^2}}. \tag{197}$$

Proof. For ease of notation, let $m = \mathbb{E}_P[\ell(Z)]$ and $v = \mathbb{V}_P[\ell(Z)]$. Let $\rho \geqslant 0$ such that

$$\rho^2 \leqslant 1 - \sqrt{\frac{v}{v + (M-m)^2}}. \tag{198}$$

Then, for arbitrary $Q \in B_\rho(P)$, it follows as a direct consequence of Theorem 1 that

$$\mathbb{E}_{X\sim Q}[h(X)] \leqslant m + 2\sqrt{C_\rho(1-C_\rho)v} + C_\rho\frac{(M-m)^2-v}{M-m}. \tag{199}$$

Since the right hand side of (199) does not depend on the choice of Q, we can take the supremum over the left hand side. This completes the proof. □

We now make some general observations about this result. The bound (196) presents a *pointwise* guarantee in the sense that it upper bounds the distributional worst-case risk for a particular model $\ell(\cdot)$. This is in contrast to bounds which hold uniformly for an entire model class and introduce complexity measures such as covering numbers and VC-dimension which are hard to compute for many practical problems. Other techniques which yield a pointwise robustness certificate of the form (196), typically express the uncertainty set as a Wasserstein ball around the distribution P [16, 42, 198, 203], and require the model ℓ to be sufficiently smooth. For example, the certificate presented in [203] can only be tractably computed for small neural networks for which one can upper bound their smoothness by bounding the Lipschitz constant of their gradients. For more general and large-scale neural networks, these bounds quickly become intractable and/or lead to vacuous certificates. For example, it is known that computing the Lipschitz constant of neural networks with ReLU activations is NP-hard [227]. Secondly, we emphasize that our bound (196) is "faithful", in the sense that, as the radius approaches zero, $\rho \to 0$, the bound converges towards the true expectation $\mathbb{E}_P[\ell(Z)]$. This is of course desirable for any such bound as it indicates that any intrinsic gap vanishes as the covered distributions become increasingly closer to the reference distribution P. A third observation is that the bound (196) is *monotonically increasing* in the variance, indicating that low-variance models exhibit better generalization properties, which can be seen in light of the bias-variance trade-off. More specifically, from the form our bound (196) takes, we see that minimizing the variance-regularized objective $\mathcal{L}(\theta) = \mathbb{E}_{Z\sim P}[\ell_\theta(Z)] + \lambda\mathbb{V}_{Z\sim P}\ell_\theta(Z)$, effectively amounts to minimizing an upper bound on the worst-case risk. Indeed, various recent works have highlighted the connection between variance regularization and generalization [54, 72, 126, 148] and our result provides further evidence for this observation.

8.3 CERTIFYING OUT-OF-DOMAIN GENERALIZATION

Taking advantage of our weak assumptions on the loss functions and models, we now apply our framework to the problem of certifying the out-of-domain generalization performance of a given classifier, when measured in terms of different loss functions. In practice, one is typically only given a finite sample $Z_1, \ldots, Z_n$ from the in-domain distribution P and the bound (196) needs to be estimated empirically. To address this problem, our next step is to present a finite sampling version of the bound (196) which holds with arbitrarily high probability over the distribution P. Second, we instantiate our results with specific distribution shifts, namely, shifts in the label distribution, and shifts which only affect the covariates. Finally, we highlight specific loss and score functions and show how our result can be applied to certify the out-of-domain generalization of these functions.

8.3.1 *Finite Sample Results*

Let $Z_1, \ldots, Z_n \overset{iid}{\sim} P$ be an independent and identically distributed sample from the in-domain distribution P. One immediate way to use our bound would be to construct the empirical distribution $\hat{P}_n$ and consider the worst-case risk over distributions $Q \in B_\rho(\hat{P}_n)$, while computing the bound on the right hand side of (196) with the empirical mean and unbiased sample variance. However, for $\rho < 1$, the Hellinger ball $B_\rho(\hat{P}_n)$ will in general only contain distributions with discrete support since any continuous distribution Q has distance 1 from $\hat{P}_n$. We therefore seek another path and make use of concentration inequalities for the population variance and mean, in order to get statistically sound guarantees which hold with arbitrarily high probability. To achieve this, we bound the expectation value via Hoeffding's inequality [94], and the population variance via a bound presented in [148]. In the second step, we use the union bound as a means to bound both variance and expectation simultaneously with high probability. We leave the derivation and proof to Section E.1. These ingredients lead to the finite sampling-based version of Theorem 7, which we state in the following Corollary.

Corollary 5 (Finite-sampling bound). *Let $Z_1, \ldots, Z_n$ be independent random variables drawn from P and taking values in $\mathcal{Z}$. For a loss function $\ell\colon \mathcal{Z} \to [0, M]$, let $\hat{L}_n := \frac{1}{n}\sum_{i=1}^n \ell(Z_i)$ be the empirical mean and $S_n^2 := \frac{1}{n(n-1)}\sum_{1\leqslant i<j\leqslant n}^n (\ell(Z_i) - \ell(Z_j))^2$ be the unbiased estimator of the variance of the random variable $\ell(Z)$, $Z \sim P$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\begin{aligned}\sup_{Q\in B_\rho(P)} \mathbb{E}_Q[\ell(Z)] \leqslant \hat{L}_n + 2\sqrt{C_\rho(1-C_\rho)S_n^2} + \Delta_{n,\rho} \\ + C_\rho\left[M - \hat{L}_n + \frac{S_n^2 + 2M\sqrt{S_n^2\epsilon_{n,\delta}} + M^2\epsilon_{n,\delta}}{\hat{L}_n - M\left(1-\sqrt{\frac{n-1}{4n}\epsilon_{n,\delta}}\right)}\right]\end{aligned} \tag{200}$$

where

$$\epsilon_{n,\delta} = \frac{2\ln 2/\delta}{n-1}, \quad \Delta_{n,\rho} = M\sqrt{\epsilon_{n,\delta}}\left(2\sqrt{C_\rho(1-C_\rho)} - \frac{C_\rho}{2}\sqrt{\frac{n-1}{n}}\right) \tag{201}$$

and $C_\rho = \rho^2(2-\rho^2)$. The radius ρ is required to be small enough such that

$$\rho^2 \leqslant 1 - \left[1 + \left(\frac{\hat{L}_n - M\left(1-\sqrt{\frac{\ln 2/\delta}{2n}}\right)}{\sqrt{S_n^2} + M\sqrt{\frac{2\ln 1/\delta}{n-1}}}\right)^2\right]^{-1/2}. \tag{202}$$

Thus, we have derived a certificate for out-of-domain generalization for general bounded loss functions and models h which can be efficiently estimated from finite data sampled from the distribution P.

8.3.2 *Specific Distribution Shifts*

We now consider specific distribution shifts and discuss our main results in light of shifts in the distributions of labels and covariates.

8.3.2.1 *Label Distribution Shifts*

Shifts in the label distribution occur when, during deployment, an ML-system operates in an environment where the relative frequency of certain classes increases or decreases, compared to the training environment, or, as is common in practical applications, instances of previously unseen classes appear. This can potentially harm the model performance dramatically and can have severe implications, in particular in the context of fairness and ethics in machine learning. To investigate this type of distribution shift, we follow the common practice to assume that the distribution over covariates, conditioned on the labels, stays constant. Formally, here, we consider the distribution shift $P \to Q$ expressed via

$$p(x, y) = \pi(x|y)p(y) \mapsto q(x, y) = \pi(x|y)q(y) \tag{203}$$

where $\pi(x|y)$ is given by a fixed distribution over covariates, conditioned on labels. In this case, it can be shown that the Hellinger distance is equal to the L_2 norm between the square roots of the (label) probability vectors $p = (p(1), \ldots, p(K))^T \in \mathbb{R}^K$ and $q = (q(1), \ldots, q(K))^T \in \mathbb{R}^K$ where K is the number of classes, so that

$$H(P, Q) = \frac{1}{\sqrt{2}} \|\sqrt{p} - \sqrt{q}\|_2 \tag{204}$$

where the square root is applied to each element in the respective probability vector.

8.3.2.2 *Covariate Distribution Shifts*

In contrast to label distribution shifts, here we consider shifts to the distribution of covariates. This models scenarios where the relative frequency of labels stays constant, but environments change, for example the shift from day to night in autonomous driving or wildlife surveillance. Formally, we consider the shift $P \to Q$ with

$$p(x, y) = \pi(y|x)p(x) \mapsto q(x, y) = \pi(y|x)q(x) \tag{205}$$

where $\pi(y|x)$ is given by a fixed distribution on labels, conditioned on the covariates. In this scenario, the Hellinger distance between P and Q reduces to the distance between the marginals

$$H(P, Q) = \sqrt{\frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx}. \tag{206}$$

In principle, this quantity could be estimated from unlabeled samples of a target distribution Q, enabling one to reason about distributional robustness of a given model, by evaluating our bounds from Theorem 7 and Corollary 5. However, in practice, it is generally difficult to estimate f-divergences, and in particular the Hellinger distance, from data for practically relevant problem instances. Although first steps in this direction have been made [163, 164, 204], it remains largely an open problem and a potential solution would give our approach additional ounces of practical significance. We view this problem as orthogonal to certifying out-of-domain generalization and believe that research efforts towards such an end-to-end solution pose an exciting future research direction.

DISCUSSION. We notice that when considering label- or covariate distribution shifts, we are effectively interested in a subset of all probability distributions with a given predefined Hellinger distance. In other words, if the shift $P \to Q$ models the label distribution shift with distance $H(P, Q) \leqslant \rho$, then applying the certificate (196) with radius ρ also covers every other type of distribution shift bounded by ρ and hence gives a more conservative view than desired. This is because, in general,

$$\sup_{\substack{Q : H(P,Q) \leqslant \rho \\ q(\cdot|y) \equiv p(\cdot|y)}} \mathbb{E}_Q[\ell(Z)] \leqslant \sup_{Q : H(P,Q) \leqslant \rho} \mathbb{E}_Q[\ell(Z)] \tag{207}$$

arising from the additional constraint that $q(x|y) = p(x|y)$ for all $x \in \mathcal{X}$. A similar argument can be made for covariate shifts. Naturally, this leads to an intrinsic gap between the actual and certified robustness, which we also observe in our experiments. Finally, it is worth pointing out the connection with generalization from finite amounts of data which can be seen as a specific instantiation of the worst-case risk (193) where the in-domain distribution corresponds to the empirical distribution $\hat{P}_n$ and the radius ρ decays as $\mathcal{O}(1/n)$. In this sense, the distribution shift originates from the transition from the empirical to the true data distribution. This type of distribution shift has been analyzed in [54] where further links to variance-based regularization have been established.

8.3.3 *Specific Loss and Score Functions*

We now turn our attention to specific loss functions and discuss, in particular, the Jensen-Shannon divergence loss, the classification error, and the AUC score.

8.3.3.1 *Jensen-Shannon Divergence*

The Jensen-Shannon Divergence is a particular type of loss function for classification models, and serves as a symmetric alternative to other common losses such as cross entropy. It has been observed that the JSD loss and its generalizations have favorable properties compared to the standard cross entropy loss, as it is bounded, symmetric, and its square root is a distance and hence satisfies the triangle inequality. In [58] it has been observed that JSD loss can be seen as an interpolation between cross-entropy and mean absolute error and is particularly well suited for classification problems with noisy labels. Formally, the Jensen-Shannon divergence is defined as

$$D_{JS}(P, Q) := \frac{1}{2}\Big(D_{KL}(P \| \mu) + D_{KL}(Q \| \mu)\Big) \tag{208}$$

where D_{KL} is the Kullback-Leibler divergence and $\mu = \frac{1}{2}(P + Q)$. Since it is a bounded loss function, it is straightforward to apply our results to certify the out-of-domain generalization for the JSD loss and, due to its smoothness, allows for a principled comparison between our bound and the Wasserstein distance certificates proposed in [42, 203].

8.3.3.2 *Classification Error*

The classification error is among the most popular choices for measuring the performance of classification models and serves as a means to assess how accurate a classifier

is on a given data distribution. As it is a non-smooth function, existing approaches cannot in general certify distributional robustness for this function. In contrast, one can immediately instantiate our Theorem 7 (or the finite sampling version from Corollary 5) with this loss. Indeed, for a fixed model $h\colon \mathcal{X} \to \mathcal{Y}$, let $\epsilon_P := \mathbb{P}_{(X,Y)\sim P}[h(X) \neq Y]$ and analogously ϵ_Q. Then, in the infinite sampling regime, we immediately get an upper bound on the worst-case classification error from Theorem 7. Namely, for a sufficiently small radius $\rho^2 \leqslant 1 - \sqrt{\epsilon_P}$, we have

$$\sup_{Q\in B_\rho(P)} \epsilon_Q \leqslant \epsilon_P + 2\sqrt{C_\rho(1-C_\rho)\epsilon_P(1-\epsilon_P)} + C_\rho(1-2\epsilon_P) \tag{209}$$

where $C_\rho = \rho^2(2-\rho^2)$.

8.3.3.3 *AUC Score*

Among other uses, the Area under the ROC (AUC) score [38, 79] is a metric to measure the performance of binary classification models. Unlike the classification error, which captures the ability to classify a single randomly chosen instance, the AUC score provides a means to quantify the ability to correctly assigning to any positive instance a higher score than to a randomly chosen negative instance. For a binary classification model $h\colon \mathcal{X} \to \mathbb{R}$ that outputs the score of the positive class, the AUC score is defined as

$$\mathrm{AUC}(h) = \mathbb{P}\left[h(X) \geqslant h(X') | Y = 1, Y' = -1\right] \tag{210}$$

where (X, Y) and (X', Y') are independent and identically distributed according to P. By introducing the notation $X_\pm := X|\, Y = \pm 1$, we can equivalently write the AUC score as an expectation value over the joint (conditional) distribution of $Z := (X_+, X_-)$

$$\mathrm{AUC}(h) = \mathbb{E}_{(X_+,X_-)\sim P_Z}[\mathbb{1}_{\{h(X_+)\geqslant h(X_-)\}}]. \tag{211}$$

We notice that only distribution shifts on the covariates have an impact on the AUC score. For this reason, we consider a setting similar to the covariate shift setting of Section 8.3.2.2, although we consider shifts in the conditional distribution $p(x|y) \mapsto q(x|y)$ for each $y \in \{\pm 1\}$ in contrast to shifts in the marginals. Due to independence, the probability density function of $Z \sim P_Z$ can be written as

$$p_Z(x_+, x_-) = p(x|y = +1)p(x|y = -1) \tag{212}$$

and similarly for the shifted distribution Q. Thus, assuming that for both negative and positive samples a distribution drift with $H(P_{X|Y=y}, Q_{X|Y=y}) \leqslant \rho$ occurs, the squared Hellinger distance between P_Z and Q_Z is bounded by

$$H^2(P_Z, Q_Z) \leqslant \rho^2(2-\rho^2). \tag{213}$$

Thus, for certifying out-of-domain generalization for the AUC score, we can apply our bound by instantiating it with Hellinger distance $\sqrt{\rho^2(2-\rho^2)}$. We remark that for the AUC score, one is typically interested in *lower* bounding it under distribution shifts. To that end, we present a lower bound version of our Theorem 7 in Section E.2.

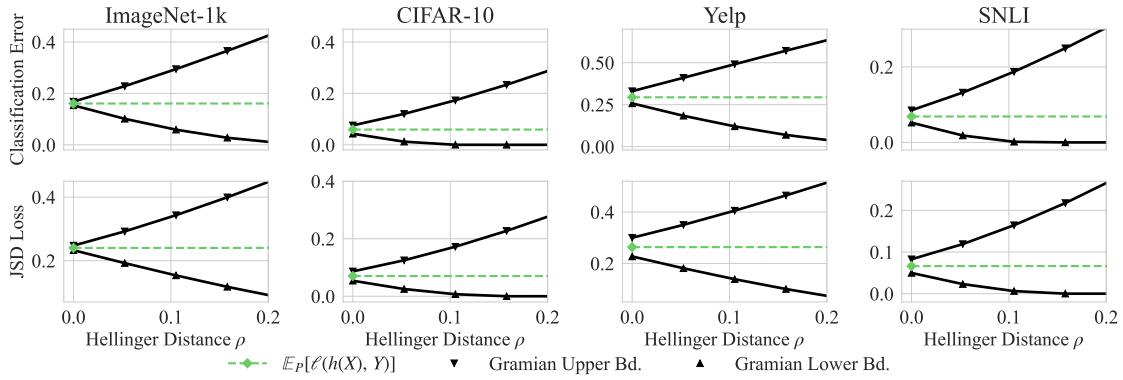


Figure 16: Distributional robustness certificates for generic distribution shifts on vision and NLP datasets for JSD loss and classification error.

8.4 EXPERIMENTS

We now experimentally validate our theoretical findings on a diverse collection of datasets and scenarios. We first provide certificates considering generic distribution shifts $P \to Q$ and then provide detailed analysis on the two specific scenarios described in Section 8.3.2.1 and Section 8.3.2.2, namely, shifts in the label and in the covariate distributions. Finally, we construct a synthetic example that allows for a fair comparison of our bounds with the Wasserstein certificate of [203], which indicates that in addition to favorable scalability properties, our bounds are also considerably tighter. We remark that all our bounds are computed using the finite sampling bounds presented in Corollary 5 and hold with 99% probability ($\delta = 0.01$).[1]

DATASETS We certify out-of-domain generalization on two standard vision datasets: ImageNet-1k [185] containing objects of 1,000 different classes; and CIFAR-10 [119], which contains natural images of 10 different classes. We also conduct experiments on the standard natural language processing (NLP) datasets Yelp [31] and SNLI [19]. We follow Lin et al. [137] to sample 2,000 examples for the Yelp test set and 10,000 examples for the SNLI test set.

MODELS For classification on ImageNet-1k, we use the EfficientNet-B7 [214] architecture which we initialize with pre-trained weights; we use DenseNet-121 [101] for CIFAR-10. On Yelp, we use BERT [48] and on SNLI we use a DeBERTa architecture [86].

SETTINGS FOR AUC SCORES When we consider AUC scores, we further constrain all multiclass datasets into a binary version. To this end, on ImageNet, we randomly choose two classes and train a ResNet-152 architecture to discriminate between the two Synsets *n01601694* and *n04330267* (corresponding to the classes 'water ouzel' and 'stove'). Similarly, on CIFAR-10 we also pick two classes at random and train a ResNet-110 classifier for the two classes *'bird'* and *'horse'*.

1 Our code is publicly available at https://github.com/DS3Lab/certified-generalization.

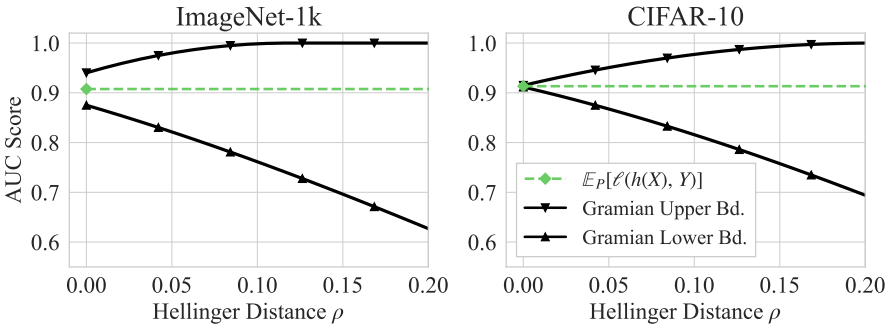


Figure 17: Certified Generalization for AUC against generic distribution shifts on binary ImageNet and CIFAR datasets.

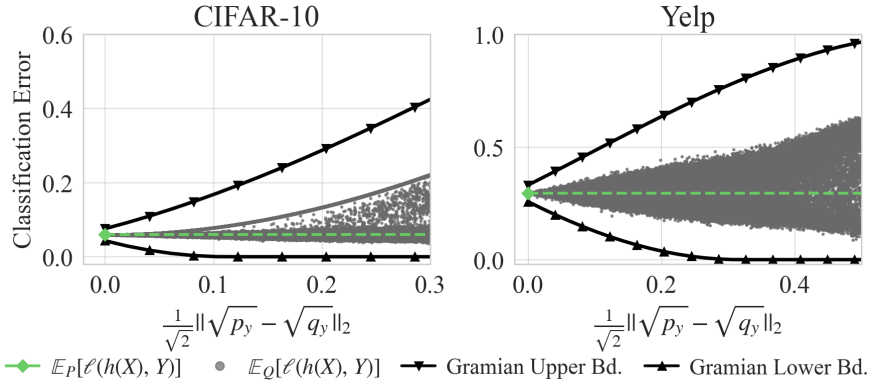


Figure 18: Certified Generalization for label distribution shifts. Each gray point corresponds to a randomly sampled label distribution with corresponding Hellinger distance and empirical loss.

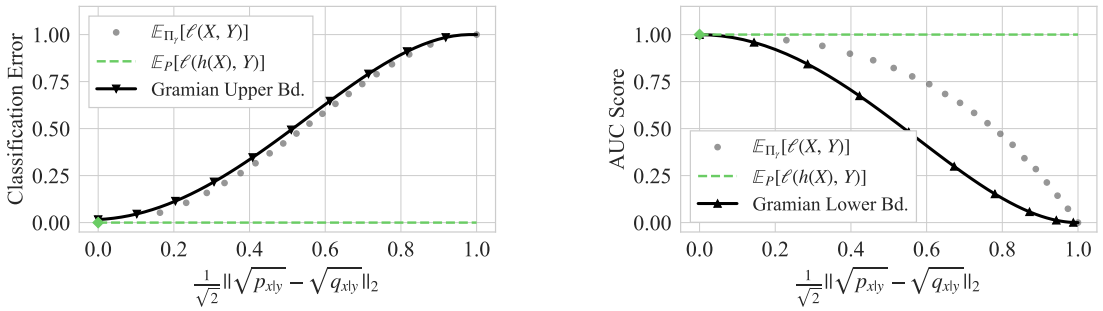
8.4.1 *Certifying specific Distribution Shifts*

Figure 16 and Figure 17 illustrate the certificates that we provide on a diverse range of datasets, considering three different scores: classification error, JSD loss, and AUC score. In all these figures, the x-axis corresponds to the degree of distribution drift, and the *Gramian Certificate* curves correspond to the lower and upper bound of these scores under distribution drifts. To our best knowledge, this is the first time that nonvacuous certificates are obtained on this diverse range of datasets, scores, and large-scale models.

8.4.1.1 *Label Distribution Shifts*

To get a better indication of how well our certificates capture the true risk under label distribution shifts, we randomly generate 100,000 shifted class distributions on the CIFAR-10 and Yelp datasets by 1) subsampling existing classes, 2) removing the counts of existing classes, and 3) including new "unseen" classes. This allows us to empirically compute both the classification error and the Hellinger distance and enables us to compare the certificates to the actual loss on the shifted distribution. We can see from Figure 18 that our certificates indeed provide a valid upper and lower bound. Note that, given that all shifted class distributions are randomly sampled, we might not hit the true worst-case scenario, explaining the clear gap between the generalization certificates and the scores obtained from the randomly generated label distributions. Another

reason for the gap can be attributed to the intrinsic gap for label and covariate shifts, discussed in Section 8.3.2. We refer the reader to Section E.6 for analogous figures with a larger set of model architectures on the CIFAR-10 dataset. Finally, we point out the difficulty in sampling these class distributions for datasets with a large number of classes and include analogous figures for ImageNet and the SNLI dataset in Section E.6.



(a) Classification Error

(b) AUC Score

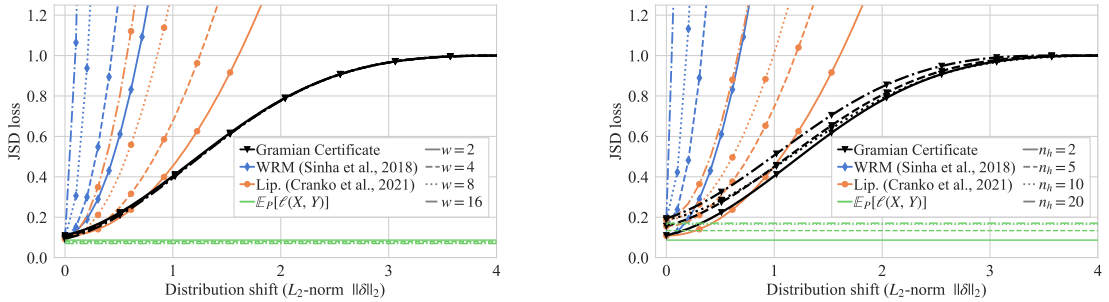
Figure 19: Certified Generalization for covariate shift on colored MNIST.

8.4.1.2 *Covariate Distribution Shifts.*

We now investigate our certificates in light of changes in the distribution of the covariates and consider the scenario described in Section 8.3.2.2. In this experiment, we use the binary Colored MNIST dataset [4, 112], which is constructed from the MNIST dataset by coloring the digits 0-4 in green and 5-9 in red for the training set, while flipping the coloring in the test set. The classifier is then trained to classify the digits into the two groups $\{0, 1, 2, 3, 4\}$ and $\{5, 6, 7, 8, 9\}$. In this setting, the classifier learns to perfectly distinguish the two classes in the training set, but fails on the testing set since the color is a stronger predictor than the shape of the digits. To investigate the space between these two extreme cases, we generate mixture distributions between training and test distribution in the following way. We set P to be the training distribution and Q the testing distribution (containing digits with flipped colors). Guided by a mixing parameter γ, we mix P and Q to obtain the mixture distribution $\Pi_\gamma := \gamma \cdot P + (1-\gamma) \cdot Q$. Since P and Q have disjoint support, we compute the Hellinger distance between P and Π_γ as $H(P, \Pi_\gamma) = \sqrt{1-\sqrt{\gamma}}$ as shown in Section E.5. Figure 19 illustrates our robustness certificates for the 0-1 loss and the AUC Score, as well as the empirical losses $\mathbb{E}_{\Pi_\gamma}[\ell(X, Y)]$ for different values of the mixture parameter γ. We see from the figure that our technique provides quite tight certificates for both classification error and AUC score.

8.4.2 *Comparison with Wasserstein Certificates*

We now construct a synthetic example that enables a fair comparison with two baseline certificates based on the Wasserstein distance. Namely, we compare our approach with 1) the certificate which uses the Lipschitz constant of the ML model, presented in [42]; and 2) with the pointwise robustness certificate derived in [203] from the dual formulation of the worst-case risk. We remark that these certificates cannot be applied to our previous examples because of their prohibitive assumptions. To make the three



(a) Robustness certificates for varying network widths w and $n_h = 2$ hidden layers.

(b) Robustness certificates for varying numbers of hidden layers n_h and fixed width $w = 4$.

Figure 20: Comparison of our approach with the Wasserstein-based certificates from [42, 203] for varying levels of model complexity.

techniques comparable, we consider a Gaussian mixture model and certify the Jensen-Shannon divergence loss, while modeling distribution shifts as dislocations, $X \mapsto X + \delta$ for a fixed perturbation vector δ. This allows us to parameterize the distribution shift via the L_2-norm of δ and obtain a one-to-one correspondence between our Hellinger distance and the Wasserstein distance, and enables a principled comparison. We describe the details of this synthetic dataset in Section E.3. To investigate how the techniques scale with increased model complexity, we use fully connected feedforward neural networks with varying depths and widths. In addition, to accommodate [203]'s assumptions on smoothness, we use ELU activation functions on all layers. We remark that the bound in [203] requires one to solve a complex maximization problem, which requires the composition of the loss function and the network to be sufficiently smooth. Furthermore, the concavity of the maximization problem hinges on knowledge of the Lipschitz constant of the gradient. For small examples, this Lipschitz constant can be obtained, as we show in Section E.4 for the JSD loss function. As can be seen in Figure 20, all bounds converge to the expected loss $\mathbb{E}_P[\ell(X, Y)]$ as the perturbation goes to zero, $\|\delta\|_2 \to 0$. However, the certificate from [203] quickly becomes vacuous as the perturbation magnitude increases. In addition, both baseline bounds become loose with increasing model complexity, while our bound is virtually agnostic to the model architecture as it only depends on the variance and expected loss on the distribution P.

8.5 CONCLUSION

In this chapter, we have focused on a further aspect of the robustness during the deployment stage of an ML pipeline and studied the problem of certifying the out-of-domain generalization while only allowing for blackbox access to the model. To that end, we have presented bounds on the worst-case population risk over an uncertainty set of probability distributions given by a Hellinger ball. In contrast to existing approaches, our framework is scalable since it treats the loss function together with the model as a blackbox and thus requires virtually no knowledge about the internals of, e.g., neural networks. We have provided experimental evidence that our technique can handle large models and datasets and provides, to the best of our knowledge, the first non-vacuous out-of-domain generalization bounds for problems as large as ImageNet with a full-fledged EfficientNet-B7. While our techniques provide a means to certify robustness

against *general* distribution shifts, future research directions can potentially extensively study more specific distribution shifts in order to close the gap stemming from the general nature of the uncertainty set. Perhaps the most significant limitation from a practical point of view, is the difficulty of estimating the Hellinger distance from general real-life data samples. We believe that investigating problems where such an estimate can be efficiently obtained, is fruitful grounds for future research as it provides our guarantees with further operational significance.

Part IV

QUANTUM MACHINE LEARNING

9 CERTIFIED ROBUSTNESS VIA QUANTUM HYPOTHESIS TESTING

In this part of the thesis, Part iv, we switch our attention to a different model of computation. In the part on classical machine learning, Part iii, information is stored as bits, and the state of the system of computation is deterministically known. Here we move away from this model of computation and consider information stored in qubits, resulting in an inherently probabilistic way in which computations are executed and information is extracted from a system. In this chapter, we first focus on certifying the robustness of quantum classification models by generalizing the robustness guarantees based on the Neyman-Pearson Lemma [39] to the quantum domain. Next to robustness against adversarial attacks, quantum ML models are just as vulnerable to natural noise occurring in quantum computing systems, and we unify these two notions by viewing adversarial attacks as a worst-case form of noise. In the subsequent chapter, we take a further step and consider not only classification models, but more general algorithms whose output is computed via the estimation of expectation values.

9.1 INTRODUCTION

9.1.1 *Overview*

The flourishing interplay between quantum computation and machine learning has inspired a wealth of algorithmic invention in recent years [14, 55, 192]. Among the most promising proposals are quantum classification algorithms which aspire to leverage the exponentially large Hilbert space uniquely accessible to quantum algorithms to either drastically speed up computational bottlenecks in classical protocols [41, 63, 179, 278], or to construct quantum-enhanced kernels that are prohibitive to compute classically [81, 140, 191]. Although these quantum classifiers are recognised as having the potential to offer quantum speedup or superior predictive accuracy, they are shown to be just as vulnerable to input perturbations as their classical counterparts [71, 138, 141, 211]. These perturbations can occur either due to imperfect implementation which is prevalent in the NISQ era [174], or, more menacingly, due to adversarial attacks where a malicious party aims to fool a classifier by carefully crafting practically undetectable noise patterns which trick a model into misclassifying a given input.

In order to address these short-comings in reliability and security of QML, several protocols in the setting of adversarial quantum learning, i.e. learning under the worst-case noise scenario, have been developed [50, 77, 138, 141, 248]. More recently, data encoding schemes are linked to robustness properties of classifiers with respect to different noise models in [125]. The connection between provable robustness and quantum differential privacy is investigated in [50], where naturally occurring noise in quantum systems is leveraged to increase robustness against adversaries. A further step towards robustness guarantees is made in [77] where a bound is derived from elementary properties of the trace distance. These advances, though having accumulated considerable momentum toward a coherent strategy for protecting QML algorithms against adversar-

ial input perturbations, have not yet provided an adequate framework for deriving a tight robustness condition for any given quantum classifier. In other words, the known robustness conditions are sufficient but not, in general, necessary.

Thus, a major open problem remains which is significant on both the conceptual and practical levels. Conceptually, adversarial robustness, being an intrinsic property of the classification algorithms under consideration, is only accurately quantified by a tight bound, the absence of which renders the direct robustness comparison between different quantum classifiers implausible. Practically, an optimal robustness certification protocol, in the sense of being capable of faithfully reporting the noise tolerance and resilience of a quantum algorithm can only arise from a robustness condition which is both sufficient and necessary. Here we set out to confront both aspects of this open problem by generalising the state-of-the-art classical wisdom on certifiable adversarial robustness into the quantum realm.

The pressing demand for robustness against adversarial attacks is arguably even more self-evident under the classical setting in the present era of wide-spread industrial adaptation of machine learning [61, 71, 211]. Many heuristic defence strategies have been proposed but have subsequently been shown to fail against suitably powerful adversaries [6, 28]. In response, provable defence mechanisms that provide robustness guarantees have been developed. One line of work, interval bound propagation, uses interval arithmetic [73, 153] to certify neural networks. Another approach makes use of randomizing inputs and adopts techniques from differential privacy [128] and, to our particular interest, statistical hypothesis testing [39, 240] which has a natural counter-part in the quantum domain. Since the pioneering works by Helstrom [88] and Holevo [95], the task of QHT has been well-studied and regarded as one of the foundational tasks in quantum information, with profound linkages with topics ranging from quantum communication [147, 235], estimation theory [90], to quantum illumination [139, 250].

9.1.2 *Contributions*

In this work, we lay bare a fundamental connection between quantum hypothesis testing and the robustness of quantum classifiers against unknown noise sources. The methods of QHT enable us to derive a robustness condition which, in contrast to other methods, is both *sufficient and necessary* and puts constraints on the amount of noise that a classifier can tolerate. Due to tightness, these constraints allow for an accurate description of noise-tolerance. Absence of tightness, on the other hand, would underestimate the true degree of such noise tolerance. Based on these theoretical findings, we provide (1) an optimal robustness certification protocol to assess the degree of tolerance against input perturbations (independent of whether these occur due to natural or adversarial noise), (2) a protocol to verify whether classifying a perturbed (noisy) input has had the same outcome as classifying the clean (noiseless) input, without requiring access to the latter, and (3) tight robustness conditions on parameters for amplitude and phase damping noise. In addition, we will also consider randomizing quantum inputs, what can be seen as a quantum generalisation to randomized smoothing, a technique that has recently been applied to certify the robustness of classical machine learning models [39]. The conceptual foundation of our approach is rooted in the inherently probabilistic nature of quantum classifiers. Intuitively, while QHT is concerned with the question of

Table 9: Summary of Results. Robustness conditions where the tightness result from Theorem 9 applies are highlighted in **bold font**.

	Input States	Quantum Differential Privacy	Hölder Duality	Quantum Hypothesis Testing			
				SDP Problem[a]	Fidelity	Bures Metric	Trace Distance
No Smoothing	Pure	—	Lemma 11[b]	**Theorem 8**	**Theorem 10**	**Eq. (240)**	**Eq. (236)**
	Mixed						Lemma 11
Depolarization Smoothing	Pure	Lemma 2 in [50]	Eq. (266)	**Theorem 8**	—	—	**Corollary 7 (single-qubit)**
	Mixed						—

[a] Robustness condition expressed in terms of type-II error probabilities β^* associated with an optimal quantum hypothesis test.
[b] Independently discovered in [77].

how to optimally discriminate between two given states, certifying adversarial robustness aims at giving a guarantee for which two states can *not* be discriminated. These two seemingly contrasting notions go hand in hand and, as we will see, give rise to optimal robustness conditions fully expressible in the language of QHT. Furthermore, while we focus on robustness in a worst-case scenario, our results naturally cover narrower classes of *known* noise sources and can potentially be put in context with other areas such as error mitigation and error tolerance in the NISQ era. Finally, while we treat robustness in the context of QML, our results in principle do not require the decision function to be learned from data. Rather, our results naturally cover a larger class of quantum algorithms whose outcomes are determined by the most likely measurement outcome. Our robustness conditions on quantum states are then simply conditions under which the given measurement outcome remains the most likely outcome.

The remainder of this chapter is organized as follows. We introduce the notations and terminologies in Section 9.2, where we also formally define quantum classifiers and the assumptions on the threat model. In Section 9.3, we present our main results on provable robustness from quantum hypothesis testing. Additionally, in Section 9.4, these results are demonstrated and visualised with a simple toy example. In Section 9.5 we present an algorithm for robustness certification and study the robustness against specific noise models in detail. We conclude this chapter in Section 9.6 with a higher-level view on our findings and layout several related open problems with an outlook for future research.

9.2 PRELIMINARIES

NOTATION Let $\mathcal{H}$ be a Hilbert space of finite dimension $d := \dim(\mathcal{H}) < \infty$ corresponding to the quantum system of interest. The space of linear operators acting on $\mathcal{H}$ is denoted by $\mathcal{L}(\mathcal{H})$ and the identity operator on $\mathcal{H}$ is written as $\mathbb{1}$. If not clear from context, the dimensionality is explicitly indicated through the notation $\mathbb{1}_d$. The set of density operators (i.e. positive semi-definite trace-one Hermitian matrices) acting on $\mathcal{H}$, is denoted by $\mathcal{S}(\mathcal{H})$ and elements of $\mathcal{S}(\mathcal{H})$ are written in lowercase Greek letters. The Dirac notation will be adopted whereby Hilbert space vectors are written as $|\psi\rangle$ and their dual as $\langle\psi|$. We will use the terminology density operator and quantum state interchangeably. For two Hermitian operators $A, B \in \mathcal{L}(\mathcal{H})$ we write $A > B$ ($A \geqslant B$) if $A - B$ is positive (semi-)definite and $A < B$ ($A \leqslant B$) if $A - B$ is negative (semi-)definite. For a Hermitian operator $A \in \mathcal{L}(\mathcal{H})$. with spectral decomposition $A = \sum_i \lambda_i P_i$, we write $\{A > 0\} := \sum_{i:\, \lambda_i > 0} P_i$ (and analogously $\{A < 0\} := \sum_{i:\, \lambda_i < 0} P_i$) for the projection onto the eigenspace of A associated with positive (negative) eigenvalues. The Hermitian transpose of an operator A is written as $A^\dagger$ and the complex conjugate of a complex

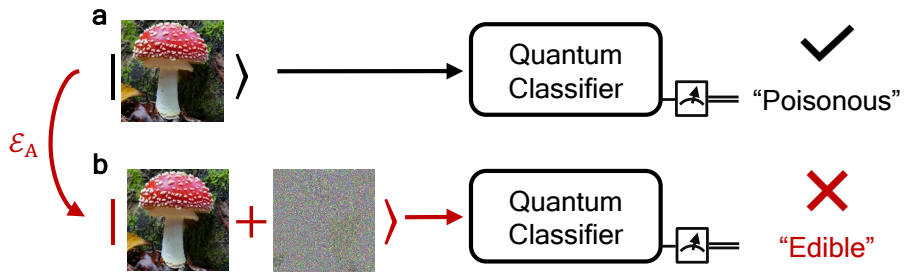


Figure 21: Adversarial attack. **a** A quantum classifier correctly classifies the (toxic) mushroom as `poisonous`. **b** An adversary perturbs the image to fool the classifier into believing that the mushroom is `edible`.

number $z \in \mathbb{C}$ as $\bar{z}$. For two density operators ρ and σ, the trace distance is defined as $T(\rho, \sigma) := \frac{1}{2}\|\rho - \sigma\|_1$ where $\|\cdot\|_1$ is the Schatten 1-norm defined on $\mathcal{L}(\mathcal{H})$ and given by $\|A\|_1 := \mathrm{Tr}[|A|]$ with $|A| = \sqrt{A^\dagger A}$. The Uhlmann fidelity between density operators ρ and σ is denoted by F and defined as $F(\rho, \sigma) := \mathrm{Tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]^2$ which for pure states reduces to the squared overlap $F(|\psi\rangle, |\phi\rangle) = |\langle\psi|\phi\rangle|^2$. Finally, the Bures metric is denoted by d_B and is closely related to the Uhlmann fidelity via $d_B(\rho, \sigma) = [2(1 - \sqrt{F(\rho, \sigma)})]^{\frac{1}{2}}$.

QUANTUM ADVERSARIAL MACHINE LEARNING Adversarial examples are attacks on classification models where an adversary aims to induce a wrong prediction using imperceptible modifications of a benign input example. Specifically, given a quantum classifier $\mathcal{A}$, as defined in Section 3.2, and a benign input state σ, an adversary can craft a small perturbation $\sigma \to \rho$ which results in a flipped prediction, i.e. $\mathcal{A}(\rho) \neq \mathcal{A}(\sigma)$. An illustration for this threat scenario is given in Figure 21. In this work, we seek a worst-case robustness guarantee against *any* possible attack: as long as ρ does not differ from σ by more than a certain amount, we aim to guarantee that $\mathcal{A}(\sigma) = \mathcal{A}(\rho)$, independently of how the adversarial state ρ has been crafted. Formally, suppose the quantum classifier $\mathcal{A}$ takes as input a *benign* quantum state $\sigma \in \mathcal{S}(\mathcal{H})$ and produces a measurement outcome denoted by the class $k \in \mathcal{C}$ with probability $\mathbf{y}_k(\sigma) = \mathrm{Tr}[\Pi_k \mathcal{E}(\sigma)]$. Recall that the prediction of $\mathcal{A}$ is taken to be the most likely class $k_A = \arg\max_k \mathbf{y}_k(\sigma)$. An adversary aims to alter the output probability distribution in order to change the most likely class by applying a quantum operation $\mathcal{E}_A \colon \mathcal{S}(\mathcal{H}) \to \mathcal{S}(\mathcal{H})$ to σ which results in the *adversarial state* $\rho = \mathcal{E}_A(\sigma)$. Finally, we say that the classifier $\mathbf{y}$ is provably robust around σ with respect to the robustness condition $\mathcal{R}$, if for any ρ which satisfies $\mathcal{R}$, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$. It is the goal of this work to determine a tight robustness condition $\mathcal{R}$.

9.3 CERTIFIED ROBUSTNESS VIA QUANTUM HYPOTHESIS TESTING

In this section, we derive a robustness condition $\mathcal{R}$ for quantum classifiers by using the QHT bounds on expectation values presented in Lemma 2, which provides a guarantee for the outcome of a computation being unaffected by the worst-case input noise or perturbation under a given set of constraints. In the regime where the most likely class is measured with probability lower bounded by $p_A > 1/2$ and the runner up class is less likely than $p_B = 1 - p_A$, we prove tightness of the robustness bound, hence demonstrating that the QHT condition is at least partially optimal. The QHT robustness

condition, in its full generality, has a formulation as an SDP problem in terms of the optimal type-II error probabilities. Based on the closed form solution of the optimal type-II error probability presented in Lemma 3, we then simplify this condition and derive robustness bounds in terms of Uhlmann fidelity, Bures metric and trace distance between benign and adversarial inputs. The robustness bounds expressed in terms of fidelity and Bures metric are shown to be sufficient and necessary for general states and in the same regime where the QHT formulation is proven to be tight. In the case of trace distance, this can be claimed for pure states, while the bound for mixed states occurs to be weaker. These results are then compared with an alternative approach which directly applies Hölder duality to trace distances to obtain a sufficient robustness condition. The different robustness bounds and robustness conditions are summarized in Table 9.

Let us first recall that quantum hypothesis testing is concerned with the question of finding measurements that optimally discriminate between two states. A measurement is said to be optimal if it minimizes the probabilities of identifying the quantum system to be in the state σ, corresponding to the null hypothesis, when in fact it is in the alternative state ρ, and vice versa. When considering robustness guarantees, on the other hand, one aims to find a neighbourhood around a benign state σ where the class which is most likely to be measured is constant or, expressed differently, where the classifier can not discriminate between states. It becomes thus clear that quantum hypothesis testing and QML robustness aim to achieve a similar goal, although viewed from opposite directions. Indeed, as it turns out, QHT determines the robust region around σ to be the set of states (i.e. alternative hypotheses) for which the optimal type-II error probability β^* is larger than $1/2$.

To establish this connection more formally, we identify the benign state with the null hypothesis σ and the adversarial state with the alternative ρ. We note that, in the Heisenberg picture, we can identify the score function $\mathbf{y}$ of a classifier $\mathcal{A}$ with a POVM $\{\Pi_k\}_k$. For $k_A = \mathcal{A}(\sigma)$, the operator $\mathbb{1} - \Pi_{k_A}$ (and thus the the classifier $\mathcal{A}$) can be viewed as a hypothesis test discriminating between σ and ρ. Note that, for $p_A \in [0, 1]$ with $\mathbf{y}_{k_A}(\sigma) = \operatorname{Tr}[\Pi_{k_A}\sigma] \geqslant p_A$, the operator $\mathbb{1}_d - \Pi_{k_A}$ is feasible for the SDP problem $\beta^*(1 - p_A : \sigma, \rho)$ in (41) and hence

$$\mathbf{y}_{k_A}(\rho) = \beta(\mathbb{1}_d - \Pi_{k_A}; \rho) \geqslant \beta^*(1 - p_A; \sigma, \rho). \tag{214}$$

Thus, it is guaranteed that $k_A = \mathcal{A}(\rho)$ for any ρ with $\beta^*(1 - p_A; \sigma, \rho) > 1/2$. The following theorem makes this reasoning concise and extends to the setting where the probability of measuring the second most likely class is upper-bounded by p_B.

Theorem 8 (QHT robustness bound)**.** *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H})$ be benign and adversarial quantum states and let $\mathcal{A}$ be a quantum classifier with score function $\mathbf{y}$. Suppose that for $k_A \in \mathcal{C}$ and $p_A, p_B \in [0, 1]$, the score function $\mathbf{y}$ satisfies*

$$\mathbf{y}_{k_A}(\sigma) \geqslant p_A > p_B \geqslant \max_{k \neq k_A} \mathbf{y}_k(\sigma). \tag{215}$$

Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with

$$\beta^*(1 - p_A; \sigma, \rho) + \beta^*(p_B; \sigma, \rho) > 1 \tag{216}$$

To get a more intuitive understanding of Theorem 8, we first note that for $p_B = 1 - p_A$, the robustness condition (216) simplifies to

$$\beta^*(1 - p_A; \sigma, \rho) > 1/2 \tag{217}$$

With this, the relation between quantum hypothesis testing and robustness becomes more evident: if the *optimal* hypothesis test performs poorly when discriminating the two states, then a classifier will predict both states to belong to the same class. In other words, viewing a classifier as a hypothesis test between the benign input σ and the adversarial ρ, the optimality of the Helstrom operators implies that the classifier $\mathbf{y}$ is a worse discriminator and will also not distinguish the states, or, phrased differently, it is robust. This result formalizes the intuitive connection between quantum hypothesis testing and robustness of quantum classifiers. While the former is concerned with finding operators that are optimal for discriminating two states, the latter is concerned with finding conditions on states for which a classifier does *not* discriminate. In the following, we state the full proof for Theorem 8.

Proof. We prove this theorem by utilizing the bounds on expectation values presented in Lemma 2. Note that, in the Heisenberg picture, we can write the score function $\mathbf{y}$ of the classifier $\mathcal{A}$ as

$$\mathbf{y}_k(\sigma) = \operatorname{Tr}\left[\mathcal{E}^\dagger(\Pi_k)\,\sigma\right] = \operatorname{Tr}\left[E_k\sigma\right] \tag{218}$$

where $E_k := \mathcal{E}^\dagger(\Pi_k)$. Since $\mathcal{E}$ is a CPTP map, its dual is completely positive and unital and thus $0 \leqslant E_k \leqslant \mathbb{1}$ and

$$\sum_k E_k = \sum_k \mathcal{E}^\dagger(\Pi_k) = \mathcal{E}^\dagger(\mathbb{1}) = \mathbb{1}. \tag{219}$$

Applying Lemma 2 to the operator E_k and setting $\underline{m} \equiv p_A$, we have the lower bound

$$\mathbf{y}_{k_A}(\rho) = \langle E_k\rangle_\rho \geqslant \beta^*(1-p_A;\,\sigma,\,\rho). \tag{220}$$

Similarly, for any $k \neq k_A$, we apply Lemma 2 to the operator E_k and set $\overline{m} \equiv p_B$ and find the upper bound

$$\mathbf{y}_k(\rho) = \langle E_k\rangle_\rho \leqslant 1-\beta^*(p_B;\,\sigma,\,\rho). \tag{221}$$

Since the RHS does not depend on k, we can take the maximum over all $k \neq k_A$ in the LHS and find that

$$k_A = \arg\max_k \mathbf{y}_k(\rho) \tag{222}$$

whenever

$$\beta^*(1-p_A;\,\sigma,\,\rho) + \beta^*(p_B;\,\sigma,\,\rho) > 1. \tag{223}$$

This concludes the proof. □

9.3.1 *Optimality*

The robustness condition (216) from QHT is provably optimal in the regime of $p_A + p_B = 1$, which covers binary classifications in full generality and multi-class classification where the most likely class is measured with probability larger than $p_A > \frac{1}{2}$. The robustness condition is tight in the sense that, whenever condition (216) is violated, then there exists a classifier $\mathcal{A}^\star$ which is consistent with the class probabilities (215) on the benign input but which will classify the adversarial input differently from the benign input. The following theorem demonstrates this notion of tightness by explicitly constructing the worst-case classifier $\mathcal{A}^\star$.

Theorem 9 (Tightness). *Suppose that* $p_A + p_B = 1$. *Then, if the adversarial state* ρ *violates condition* (216), *there exists a quantum classifier* $\mathcal{A}^\star$ *that is consistent with the class probabilities* (215) *and for which* $\mathcal{A}^\star(\rho) \neq \mathcal{A}^\star(\sigma)$.

The main idea of the proof relies on the explicit construction of a "worst-case" classifier with Helstrom operators and which classifies ρ differently from σ while still being consistent with the class probabilities (215). In the following, we formalize this intuition and provide the proof of Theorem 9.

Proof. Note that, since $p_B = 1 - p_A$ by assumption, the robustness condition (216) reads

$$\beta^*(1 - p_A; \sigma, \rho) > 1/2. \tag{224}$$

Let $M_A^\star$ be an optimizer of the corresponding SDP problem (41) such that $\alpha(M_A^\star; \sigma) = 1 - p_A$ and

$$\beta(M_A^\star; \rho) = \beta^*(1 - p_A, \sigma, \rho). \tag{225}$$

Consider the classifier $\mathcal{A}^\star$ with score function $\mathbf{y}^\star$ defined by the POVM $\{\mathbb{1} - M_A^\star, M_A^\star, 0\}$ where the number of 0 operators is such that $\mathbf{y}$ has the desired number of classes. The score function $\mathbf{y}^\star$ is consistent with the class probabilities (215) since

$$\mathbf{y}^\star_{k_A}(\sigma) = \alpha(\mathbb{1} - M_A^\star; \sigma) = p_A \tag{226}$$

and

$$\mathbf{y}^\star_{k_B}(\sigma) = \alpha(M_A^\star; \sigma) = 1 - p_A = p_B. \tag{227}$$

Furthermore, if ρ violates (224), then we have

$$\mathbf{y}_{k_A}(\rho) = \beta(M_A^\star; \rho) \leqslant 1/2 \tag{228}$$

and thus, in particular $\mathcal{A}^\star(\rho) \neq k_A = \mathcal{A}^\star(\sigma)$. □

Whether or not the QHT robustness condition is tight for $p_A + p_B < 1$ is an interesting open question for future research. It turns out that a worst-case classifier which is consistent with p_A and p_B for benign input but leads to a different classification on adversarial input upon violating condition (216), if exists, is more challenging to construct for these cases. If such a tightness result for all class probability regimes would be proven, there would be a complete characterization for the robustness of quantum classifiers.

9.3.2 *Closed form robustness conditions*

Although Theorem 8 provides a general condition for robustness with provable tightness, it is formulated as a semidefinite program in terms of type-II error probabilities of QHT. To get a more intuitive and operationally convenient perspective, we wish to derive a condition for robustness in terms of a meaningful notion of difference between quantum states. Specifically, based on Theorem 8, and using the closed form type-II error probability presented in Lemma 3, here we derive robustness conditions expressed in terms of Uhlmann's fidelity F, Bures distance d_B and the trace distance T. To that

end, we first focus on pure state inputs and will then extend these bounds to mixed states. Finally, we show that expressing robustness in terms of fidelity or Bures distance results in a tight bound for both pure and mixed states, while for trace distance, the same can only be claimed in the case of pure states.

9.3.2.1 *Pure States*

We first assume that both the benign and the adversarial states are pure. This assumption allows us to first write the optimal type-II error probabilities $\beta^*_\alpha(\rho, \sigma)$ as a function of α and the fidelity between ρ and σ. This leads to a robustness bound on the fidelity and subsequently to a bound on the trace distance and on the Bures distance. Finally, since these conditions are equivalent to the QHT robustness condition (216), Theorem 9 implies tightness of these bounds. We formalize this result in the following Lemma.

Lemma 10. *Let $|\psi_\sigma\rangle, |\psi_\rho\rangle \in \mathcal{H}$ and let $\mathcal{A}$ be a quantum classifier. Suppose that for $k_A \in \mathcal{C}$ and $p_A, p_B \in [0, 1]$, we have $k_A = \mathcal{A}(\psi_\sigma)$ and suppose that the score function $\mathbf{y}$ satisfies (215). Then, it is guaranteed that $\mathcal{A}(\psi_\rho) = \mathcal{A}(\psi_\sigma)$ for any ψ_ρ with*

$$|\langle\psi_\sigma|\psi_\rho\rangle|^2 > \frac{1}{2}\left(1 + \sqrt{g(p_A, p_B)}\right), \tag{229}$$

where the function g is given by

$$g(p_A, p_B) = 1 - p_B - p_A(1 - 2p_B) + 2\sqrt{p_A p_B (1 - p_A)(1 - p_B)}. \tag{230}$$

This condition is equivalent to (216) and is hence both sufficient and necessary whenever $p_A + p_B = 1$.

Proof. It follows from Theorem 8 that $k_A = \mathcal{A}(\psi_\rho)$ as long as

$$\beta^*(1 - p_A; \sigma, \rho) + \beta^*(p_B; \sigma, \rho) > 1. \tag{231}$$

Since here both ρ and σ are pure states, it follows from Lemma 3 that

$$\beta^*(1 - p_A; \sigma, \rho) = (1 - p_A)(1 - 2\gamma) + \gamma - 2\sqrt{\gamma(1 - \gamma)p_A(1 - p_A)} \tag{232}$$

where we have set $\gamma = |\langle\psi_\sigma|\psi_\rho\rangle|^2$. Similarly, we have

$$\beta^*(p_B; \sigma, \rho) = p_B(1 - 2\gamma) + \gamma - 2\sqrt{\gamma(1 - \gamma)p_B(1 - p_B)} \tag{233}$$

The claim now follows directly by solving (231) for γ. □

Lemma 10 thus provides a closed form robustness bound which is equivalent to the SDP formulation in condition (216) and is hence sufficient and necessary in the regime $p_A + p_B = 1$. We remark that, under this assumption, the robustness bound (229) has the compact form

$$|\langle\psi_\sigma|\psi_\rho\rangle|^2 > \frac{1}{2} + \sqrt{p_A(1 - p_A)}. \tag{234}$$

Due to its relation with the Uhlmann fidelity, it is straight forward to obtain a robustness condition in terms of Bures metric. Namely, the condition

$$d_B(|\psi_\rho\rangle, |\psi_\sigma\rangle) < \left[2 - \sqrt{2(1 + \sqrt{g(p_A, p_B)})}\right]^{\frac{1}{2}} \tag{235}$$

is equivalent to (216). Furthermore, since the states are pure, we can directly link (229) to a bound in terms of the trace distance via the relation $T(|\psi_\rho\rangle, |\psi_\sigma\rangle)^2 = 1 - |\langle\psi_\sigma|\psi_\rho\rangle|^2$, so that

$$T(|\psi_\rho\rangle, |\psi_\sigma\rangle) < \left[\frac{1}{2}\left(1 - \sqrt{g(p_A, p_B)}\right)\right]^{\frac{1}{2}} \tag{236}$$

is equivalent to (216). Due to the equivalence of these bounds to (216), Theorem 9 applies and it follows that both bounds are sufficient and necessary in the regime where $p_A + p_B = 1$. In the following, we will extend these results to mixed states and show that both the fidelity and Bures metric bounds are tight.

9.3.2.2 *Mixed States*

Reasoning about the robustness of a classifier if the input states are mixed, rather than just for pure states, is practically relevant for a number of reasons. Firstly, in a realistic scenario, the assumption that an adversary can only produce pure states is too restrictive and gives an incomplete picture. Secondly, if we wish to reason about the resilience of a classifier against a given noise model (e. g., amplitude damping), then the robustness condition needs to be valid for mixed states as these noise models typically produce mixed states. Finally, in the case where we wish to certify whether a classification on a noisy input has had the same outcome as on the noiseless input, a robustness condition for mixed states is also required. For these reasons, and having established closed form robustness bounds which are both sufficient and necessary for pure states, here we aim to extend these results to the mixed state setting. The following theorem extends the fidelity bound (229) for mixed states. As for pure states, it is then straight forward to obtain a bound in terms of the Bures metric.

Theorem 10. *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H})$ and let $\mathcal{A}$ be a quantum classifier. Suppose that for $k_A \in \mathcal{C}$ and $p_A, p_B \in [0, 1]$, we have $k_A = \mathcal{A}(\sigma)$ and suppose that the score function $\mathbf{y}$ satisfies (215). Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with*

$$F(\rho, \sigma) > \frac{1}{2}\left(1 + \sqrt{g(p_A, p_B)}\right) =: r_F \tag{237}$$

where g is defined as in (230). This condition is both sufficient and necessary if $p_A + p_B = 1$.

Proof. To show sufficiency of (237), we notice that $\mathbf{y}$ can be rewritten as

$$\begin{aligned} \mathbf{y}_k(\sigma) &= \mathrm{Tr}\left[\Pi_k \mathcal{E}(\sigma)\right] && (238)\\ &= \mathrm{Tr}\left[\Pi_k (\mathcal{E} \circ \mathrm{Tr}_E)(|\psi_\sigma\rangle\langle\psi_\sigma|)\right] && (239)\end{aligned}$$

where $|\psi_\sigma\rangle$ is a purification of σ with purifying system E and Tr_E denotes the partial trace over E. We can thus view $\mathbf{y}$ as a score function on the larger Hilbert space which admits the same class probabilities for σ and any purification of σ (and equally for ρ). It follows from Uhlmann's Theorem that there exist purifications $|\psi_\sigma\rangle$ and $|\psi_\rho\rangle$ such that $F(\rho, \sigma) = |\langle\psi_\sigma|\psi_\rho\rangle|^2$. Robustness at ρ then follows from (237) by (238) and Lemma 10. To see that the bound is necessary when $p_A + p_B = 1$, suppose that there exists some $\tilde{r}_F < r_F$ such that $F(\sigma, \rho) > \tilde{r}_F$ implies that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$. Since pure states are a subset of mixed states, this bound must also hold for pure states. In particular, suppose $|\psi_\rho\rangle$

is such that $\tilde{r}_F < |\langle\psi_\rho|\psi_\sigma\rangle|^2 \leqslant r_F$. However, this is a contradiction, since $|\langle\psi_\rho|\psi_\sigma\rangle|^2 \geqslant r_F$ is both sufficient and necessary in the given regime, i.e. by Theorem 9, there exists a classifier $\mathcal{A}^\star$ whose score function satisfies (215) and for which $\mathcal{A}^\star(\psi_\sigma) \neq \mathcal{A}^\star(\psi_\rho)$. It follows that $\tilde{r}_F \geqslant r_F$ and hence the claim of the theorem. □

Due to the close relation between Uhlmann fidelity and the Bures metric, we arrive at a robustness condition for mixed states in terms of d_B, namely

$$d_B(\rho, \sigma) < \left[2 - \sqrt{2(1 + \sqrt{g(p_A, p_B)})}\right]^{\frac{1}{2}} \tag{240}$$

which inherits the tightness properties of the fidelity bound (237). In contrast to the pure state case, here it is less straight forward to obtain a robustness bound in terms of trace distance. However, we can still build on Lemma 10 and the trace distance bound for pure states (236) to obtain a sufficient robustness condition. Namely, when assuming that the benign state is pure, but the adversarial state is allowed to be mixed we have the following result.

Corollary 6 (Pure Benign & Mixed Adversarial States). *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H})$ and suppose that $\sigma = |\psi_\sigma\rangle\langle\psi_\sigma|$ is pure. Let $\mathcal{A}$ be a quantum classifier and suppose that for $k_A \in \mathcal{C}$ and $p_A, p_B \in [0, 1]$, we have $k_A = \mathcal{A}(\sigma)$ and suppose that the score function $\mathbf{y}$ satisfies (215). Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with*

$$T(\rho, \sigma) < \delta(p_A, p_B)\left(1 - \sqrt{1 - \delta(p_A, p_B)^2}\right) \tag{241}$$

where $\delta(p_A, p_B) = [\frac{1}{2}(1 - g(p_A, p_B))]^{\frac{1}{2}}$.

We provide a proof for this result in Section F.1.1. Intuitively, condition (241) is derived by noting that any convex mixture of robust pure states must also be robust, thus membership of the set of mixed states enclosed by the convex hull of robust pure states (certified by equation (236) is a natural sufficient condition for robustness. As such, the corresponding robustness radius in condition (241) is obtained by lower-bounding, with triangle inequalities, the radius of the maximal sphere centered at σ within the convex hull. However, the generalization from Lemma 10 and equation (236) to Corollary 6, mediated by the above geometrical argument, results in a sacrifice of tightness. How or to what extent such loosening of the explicit bound in the cases of mixed states may be avoided or ameliorated remains an open question. In the following, we compare the trace distance bounds from QHT with a robustness condition derived from an entirely different technique.

We note that a sufficient condition can be obtained from a straightforward application of Hölder duality for trace norms:

Lemma 11 (Hölder duality bound). *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H})$ be arbitrary quantum states and let $\mathcal{A}$ be a quantum classifier. Suppose that for $k_A \in \mathcal{C}$ and $p_A, p_B \in [0, 1]$, we have $k_A = \mathcal{A}(\sigma)$ and the score function $\mathbf{y}$ satisfies (215). Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with*

$$\frac{1}{2}\|\rho - \sigma\|_1 < \frac{p_A - p_B}{2}. \tag{242}$$

Proof. Let $\delta := \frac{1}{2}\|\rho - \sigma\|_1 = \sup_{0 \leqslant P \leqslant \mathbb{I}} \mathrm{Tr}\,[P(\rho - \sigma)]$, which follows from Hölder duality. We have that $\mathbf{y}_{k_A}(\sigma) - \mathbf{y}_{k_A}(\rho) \leqslant \delta$ and that $\mathbf{y}_{k_A}(\sigma) \geqslant p_A$, hence $\mathbf{y}_{k_A}(\rho) \geqslant p_A - \delta$. We

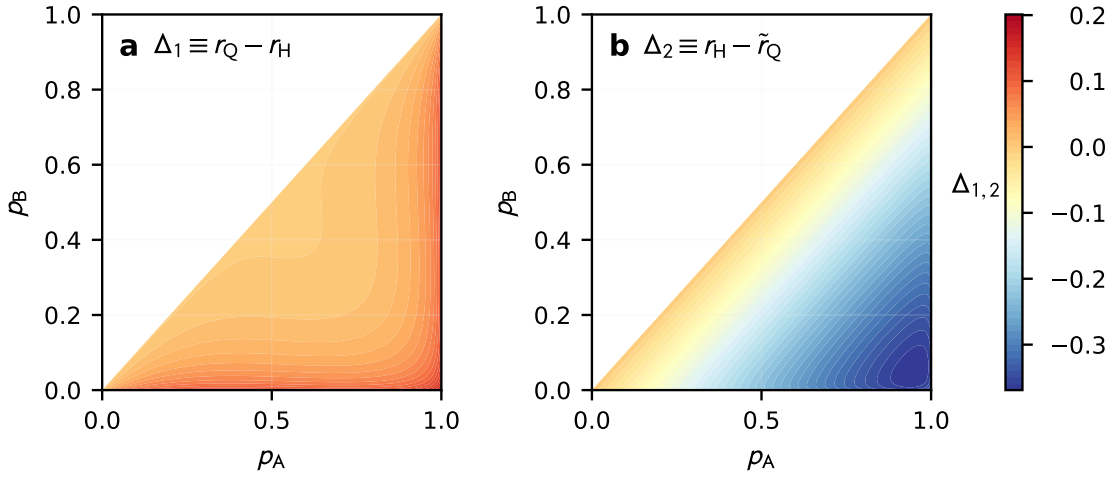


Figure 22: Comparison between robustness bounds in terms of trace distance. **a** Difference $r_Q - r_H$ between the pure state bound derived from QHT r_Q, given in Eq. (236) and the Hölder duality bound r_H from Lemma 11. **b** Difference $r_H - \tilde{r}_Q$ between the Hölder duality bound r_H and the bound $\tilde{r}_Q$ derived from the convex hull approximation to the QHT robustness condition from Theorem 8 for mixed adversarial states. It can be seen that the pure state bound r_Q is always larger than r_H which in turn is always larger than the convex hull approximation bound $\tilde{r}_Q$.

also have, for k' such that $\mathbf{y}_{k'}(\rho) = \max_{k \neq k_A} \mathbf{y}_k(\rho)$, that $\mathbf{y}_{k'}(\rho) - \mathbf{y}_{k'}(\sigma) \leqslant \delta$ and that $\mathbf{y}_{k'}(\sigma) \leqslant p_B$, hence $\max_{k \neq k_A} \mathbf{y}_k(\rho) \leqslant p_B + \delta$. Thus $\frac{1}{2}\|\rho - \sigma\|_1 < \frac{p_A - p_B}{2} \iff p_A - \delta > p_B + \delta \implies \mathbf{y}_{k_A}(\rho) > \max_{k \neq k_A} \mathbf{y}_k(\rho)$. □

We acknowledge the above robustness bound from Hölder duality was independently discovered in Lemma 1 in [77]. For intuitive insights, it is worth remarking that the condition (242) stems from comparing the maximum probability of distinguishing σ and ρ with the optimal measurement (Hölder measurement) with the gap between the first two class probabilities on σ. Since no classifier can distinguish σ and ρ better than the Hölder measurement by definition, (242) is clearly a sufficient condition. However, the Hölder measurement on σ does not necessarily result in class probabilities consistent with equation (215). Without additional constraints on desired class probabilities on the benign input, the robustness condition (242) from Hölder duality is stronger than necessary. In contrast, the QHT bound from Theorem 8, albeit implicitly written in the language of hypothesis testing, naturally incorporates such desired constraints. Hence, as expected, this gives rise to a tighter robustness condition.

In summary, the closed form solutions in terms of fidelity and Bures metric completely inherit the tightness of Theorem 8, while for trace distance, tightness is inherited for pure states, but partially lost in Corollary 6 for mixed adversarial states. The numerical comparison between the trace distance bounds from QHT and the Hölder duality bound is shown in a contour plot in Figure 22.

9.4 TOY EXAMPLE WITH SINGLE-QUBIT PURE STATES

We now present a simple example to highlight the connection between quantum hypothesis testing and classification robustness. We consider a single-qubit system which is prepared either in the state σ or ρ described by

$$|\sigma\rangle = |0\rangle, \tag{243}$$
$$|\rho\rangle = \cos(\theta_0/2)|0\rangle + \sin(\theta_0/2)e^{i\phi_0}|1\rangle \tag{244}$$

with $\theta_0 \in [0, \pi)$ and $\phi_0 \in [0, 2\pi)$. The state σ corresponds to the null hypothesis in the QHT setting and to the benign state in the classification setting. Similarly, ρ corresponds to the alternative hypothesis and adversarial state. The operators which are central to both QHT and robustness are the Helstrom operators (42) which are derived from the projection operators onto the eigenspaces associated with the non-negative eigenvalues of the operator $\rho - t\sigma$. For this example, the eigenvalues are functions of $t \geqslant 0$, and we have

$$\eta_1 = \frac{1}{2}(1-t) + R > 0, \tag{245}$$
$$\eta_2 = \frac{1}{2}(1-t) - R \leqslant 0 \tag{246}$$
$$R = \frac{1}{2}\sqrt{(1-t)^2 + 4t(1-|\gamma|^2)} \tag{247}$$

where γ is the overlap between σ and ρ and given by $\gamma = \cos(\theta_0/2)$. For $t > 0$, the Helstrom operators are then given by the projection onto the eigenspace associated with the eigenvalue $\eta_1 > 0$. The projection operator is given by $M_t = |\eta_1\rangle\langle\eta_1|$ with

$$|\eta_1\rangle = (1-\eta_1)A_1|0\rangle - \gamma A_1|\rho\rangle \tag{248}$$
$$|A_1|^{-2} = 2R\left|\eta_1 - \sin^2(\theta_0/2)\right| \tag{249}$$

where A_1 is a normalization constant ensuring that $\langle\eta_1|\eta_1\rangle = 1$. Given a preassigned probability α_0 for the maximal allowed type-I error probability, we determine t such that $\alpha(M_t; \sigma) = \alpha_0$.

Hypothesis testing view. In QHT, we are given a specific alternative hypothesis ρ and error probability α_0 and are interested in finding the minimal type-II error probability. In this example, we pick $\theta_0 = \pi/3$, $\phi_0 = \pi/6$ for the alternative state and set the type-I error probability to $\alpha_0 = 1 - p_A = 0.1$. These states are graphically represented on the Bloch sphere in Figure 23. We note that, for this choice of states, we obtain an expression for the eigenvector $|\eta_1\rangle$ given by

$$|\eta_1\rangle = \frac{9-\sqrt{3}}{\sqrt{30}}|0\rangle - 3\sqrt{\frac{2}{5}}|\rho\rangle. \tag{250}$$

which yields the type-II error probability

$$\beta^*(1-p_A; \sigma, \rho) = \beta(M_t; \rho) = 1 - |\langle\eta_1|\rho\rangle|^2 \approx 0.44 < 1/2. \tag{251}$$

We thus see that the optimal hypothesis test can discriminate σ and ρ with error probabilities less than $1/2$ since on the Bloch sphere they are located far enough apart.

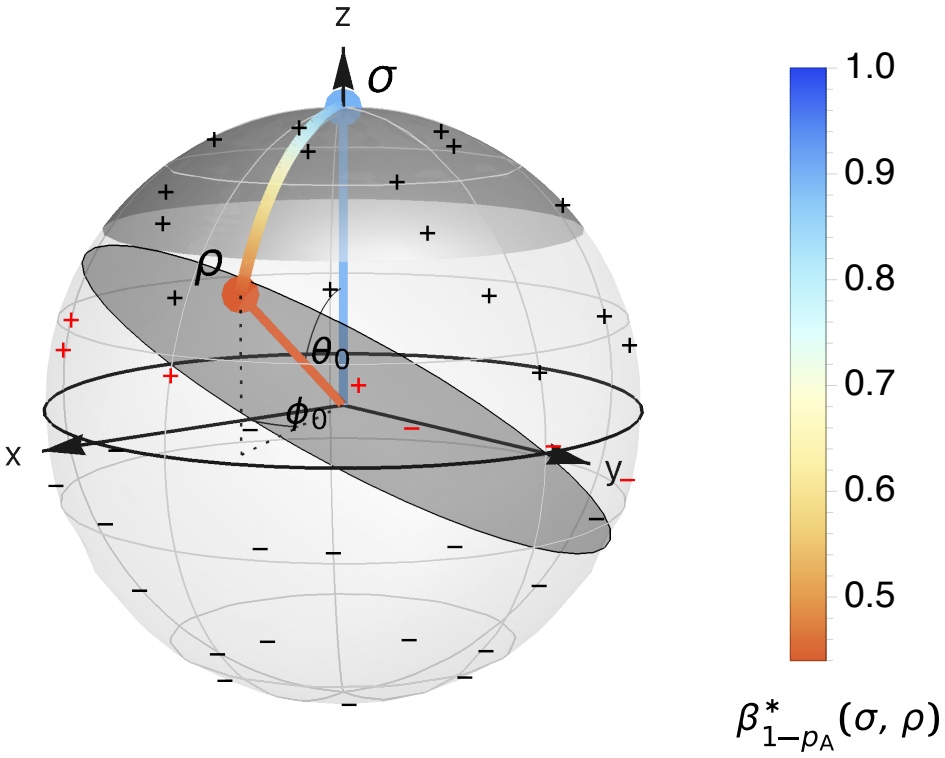


Figure 23: Example classifier for single-qubit quantum states. The decision boundary is represented by the grey disk passing through the origin of the Bloch sphere. The robust region around σ is indicated by the dark spherical cap. States belonging to different classes are marked with $+$ and $-$ and are color red if not classified correctly. The colorbar indicates different values for the optimal type-II error probability $\beta^*(1-p_A; \sigma, \rho)$. We see that, for the given classifier, the state ρ is not contained in the robust region around σ since the optimal type-II error probability is less than $1/2$ as indicated by the colorbar. The state ρ is thus not guaranteed to be classified correctly by every classifier with the same class probabilities. In the asymmetric hypothesis testing view, an optimal discriminator which admits 0.1 type-I error probability for testing σ against ρ has type-II error probability 0.44.

However, since $\beta(M_t; \rho) \not> 1/2$, Theorem 8 implies that ρ is not guaranteed to be classified equally as σ by a classifier which makes a prediction on σ with confidence at least 0.9. In other words, the two states are far enough apart to be easily discriminated by the optimal hypothesis test but too far apart to be guaranteed to be robust.

Classification robustness view. In this scenario, in contrast to the QHT view, we are not given a specific adversarial state ρ, but rather aim to find a condition on a generic ρ such that the classifier is robust for all configurations of ρ that satisfy this condition. Theorem 8 provides a necessary and sufficient condition for robustness, expressed in terms of β^*, which, for $p_B = 1 - p_A$ and $p_A > 1/2$, reads

$$\beta^*(1-p_A; \sigma, \rho) > 1/2 \tag{252}$$

Recall that the probability and $p_A > 1/2$ is a lower bound to the probability of the most likely class and in this case we set $p_B = 1 - p_A$ to be the upper bound to the

probability of the second most likely class. For example, as the QHT view shows, for $\alpha_0 = 1 - p_A = 0.1$ we have that $\beta^*(1 - p_A; \sigma, \rho) \approx 0.44 < 1/2$ for a state ρ with $\theta_0 = \pi/3$. We thus see that it is not guaranteed that *every* quantum classifier, which predicts σ to be of class k_A with probability at least 0.9, classifies ρ to be of the same class. Now, we would like to find the maximum θ_0, for which every classifier with confidence greater than p_A is guaranteed to classify ρ and σ equally. Using the fidelity bound (237), we find the robustness condition on θ_0

$$\begin{aligned} |\langle\rho|\sigma\rangle|^2 = \cos^2(\theta_0/2) > \frac{1}{2} + \sqrt{p_A(1-p_A)} \\ \iff \theta_0 < 2 \cdot \arccos\sqrt{\frac{1}{2} + \sqrt{p_A(1-p_A)}}. \end{aligned} \tag{253}$$

In particular, if $p_A = 0.9$, we find that angles $\theta_0 < 2 \cdot \arccos(\sqrt{0.8}) \approx 0.93 < \pi/3$ are certified. Figure 23 illustrates this scenario: the dark region around σ contains all states ρ for which is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any classifier $\mathcal{A}$ with confidence at least 0.9.

Classifier example. We consider a binary quantum classifier $\mathcal{A}$ which discriminates single-qubit states on the upper half of the Bloch sphere (class $+$) from states on the lower half (class $-$). Specifically, we consider the dichotomic POVM $\{\Pi_{\theta,\phi}, \mathbb{1}_2 - \Pi_{\theta,\phi}\}$ defined by the projection operator $\Pi_{\theta,\phi} = |\psi_{\theta,\phi}\rangle\langle\psi_{\theta,\phi}|$ where

$$|\psi_{\theta,\phi}\rangle := \cos(\theta/2)|0\rangle + \sin(\theta/2)e^{i\phi}|1\rangle \tag{254}$$

with $\theta = 2 \cdot \arccos(\sqrt{0.9}) \approx 0.644$ and $\phi = \pi/2$. Furthermore, for the rest of this section, we assume that $p_A + p_B = 1$ so that p_B is determined by p_A via $p_B = 1 - p_A$. An illustration of this classification problem is given in Figure 23, where the decision boundary of $\mathcal{A}$ is represented by the grey disk crossing the origin of the Bloch sphere. The states marked with a black $+$ correspond to $+$ states which have been classified correctly, states marked with a black $-$ sign correspond to data points correctly classified as $-$ and red states are misclassified by $\mathcal{A}$. It can be seen that, since the state ρ has been shown to violate the robustness condition (i.e. $\beta^*(1 - p_A; \sigma, \rho) \approx 0.44 < 1/2$), it is not guaranteed that ρ and σ are classified equally. In particular, for the example classifier $\mathcal{A}$ we have $\mathcal{A}(\rho) \neq \mathcal{A}(\sigma)$.

In summary, as $p_A \to \frac{1}{2}$, the robust radius approaches 0. In the QHT view, this can be interpreted in the sense that if the type-I error probability α_0 approaches $1/2$, then all alternative states can be discriminated from σ with type-II error probability less than $1/2$. As $p_A \to 1$, the robust radius approaches $\pi/2$. In this regime, the QHT view says that if the type-I error probability α_0 approaches 0, then the optimal type-II error probability is smaller than $1/2$ only for states in the lower half of the Bloch sphere.

9.5 ROBUSTNESS CERTIFICATION

The theoretical results in Section 9.3 provide conditions under which it is guaranteed that the output of a classification remains unaffected if the adversarial (noisy) state and the benign state are close enough, measured in terms of the fidelity, Bures metric, or trace distance. Here, we show how this result can be put to work and make concrete examples of scenarios where reasoning about the robustness is relevant. Specifically,

we first present a protocol to assess how resilient a quantum classifier is against input perturbations. Secondly, in a scenario where one is provided with a potentially noisy or adversarial input, we wish to obtain a statement as to whether the classification of the noisy input is guaranteed to be the same as the classification of a clean input without requiring access to the latter. Thirdly, we analyse the robustness of quantum classifiers against known noise models, namely phase and amplitude damping.

9.5.1 *Robustness against Adversarial Inputs*

In security critical applications, such as for example the classification of medical data or home surveillance systems, it is critical to assess the degree of resilience that machine learning systems exhibit against actions of malicious third parties. In other words, the goal is to estimate the expected classification accuracy, under perturbations of an input state within $1-\varepsilon$ fidelity. In the classical machine learning literature, this quantity is called the *certified test set accuracy* at radius r, where distance is typically measured in terms of ℓ_p-norms, and is defined as the fraction of samples in a test set which has been classified correctly and with a robust radius of at least r (i.e. an adversary can not change the prediction with a perturbation of magnitude less than r). We can adapt this notion to the quantum domain and, given a test set consisting of pairs of labelled samples $\mathcal{T} = \{(\sigma_i, y_i)\}_{i=1}^{|\mathcal{T}|}$, the *certified test set accuracy* at fidelity $1-\varepsilon$ is given by

$$\frac{1}{|\mathcal{T}|} \sum_{(\sigma, y)\in\mathcal{T}} \mathbb{1}\{\mathcal{A}(\sigma) = y \wedge r_F(\sigma) \leqslant 1-\varepsilon\} \tag{255}$$

where $r_F(\sigma)$ is the minimum robust fidelity (237) for sample σ and $\mathbb{1}$ denotes the indicator function. To evaluate this quantity, we need to obtain the prediction and to calculate the minimum robust fidelity for each sample $\sigma \in \mathcal{T}$ as a function of the class probabilities $\mathbf{y}_k(\sigma)$. In practice, in the finite sampling regime, we have to estimate these quantities by sampling the quantum circuit N times. To that end, we use Hoeffding's inequality so that the bounds hold with probability at least $1-\alpha$. Specifically, we run the following steps to certify the robustness for a given sample σ:

1. Apply the quantum circuit N times to σ and perform the $|\mathcal{C}|$-outcome measurement $\{\Pi_k\}_{k=1}^{|\mathcal{C}|}$ each time. Store the outcomes in variables n_k for every $k \in \mathcal{C}$.
2. Determine the most frequent measurement outcome k_A and set $\hat{p}_A = n_{k_A}/N - \sqrt{-\ln(\alpha)/2N}$.
3. If $\hat{p}_A > 1/2$, set $\hat{p}_B = 1-\hat{p}_A$ and calculate the minimum robust fidelity r_F according to (237) and return (k_A, r_F); otherwise abstain from certification.

Executing these steps for a given sample σ returns the true minimum robust fidelity with probability $1-\alpha$, which follows from Hoeffding's inequality

$$\Pr\left[\frac{n_k}{N} - \langle\Lambda_k\rangle_\sigma \geqslant \delta\right] \leqslant \exp\{-2N\delta^2\} \tag{256}$$

with $\Lambda_k = \mathcal{E}^\dagger(\Pi_k)$ and setting $\delta = \sqrt{-\ln(\alpha)/2N}$. In Section F.2, this intuition is shown in detail in Algorithm 5.

9.5.2 *Certifying Robustness for noisy Inputs*

In practice, inputs to quantum classifiers are typically noisy. This noise can occur either due to imperfect implementation of the state preparation device, or due to an adversary which interferes with state or gate preparation. Under the assumption that we know that the state has been prepared with fidelity at least $1-\varepsilon$ to the noiseless state, we would like to know whether this noise has altered our prediction, *without having access to the noiseless state*. Specifically, given the classification result, which is based on the *noisy* input, we would like to have the guarantee that the classifier would have predicted the same class, had it been given the noiseless input state. This would allow the conclusion that the result obtained from the noisy state has not been altered by the presence of noise. To obtain this guarantee, we leverage Theorem 10 in the following protocol. Let ρ be a noisy input with $F(\rho, \sigma) > 1-\varepsilon$ where σ is the noiseless state and let $\mathcal{A}$ be a quantum classifier with quantum channel $\mathcal{E}$ and POVM $\{\Pi_k\}_k$. Similar to the previous protocol, we again need to take into account that in practice we can sample the quantum circuit only a finite number of times. Thus, we again use Hoeffding's inequality to obtain estimates for the class probability p_A which holds with probability at least $1-\alpha$. The protocol then consists of the following steps:

1. Apply the quantum circuit N times to the (noisy) state ρ and perform the $|\mathcal{C}|$-outcome measurement $\{\Pi_k\}_{k=1}^{|\mathcal{C}|}$ each time. Store the outcomes in variables n_k for every $k \in \mathcal{C}$.
2. Determine the most frequent measurement outcome k_A and set $\hat{p}_A = n_{k_A}/N - \sqrt{-\ln(\alpha)/2N}$.
3. If $\hat{p}_A > 1/2$, set $\hat{p}_B = 1-\hat{p}_A$ and calculate the minimum robust fidelity r_F according to (237) using $\hat{p}_A$; otherwise, abstain from certification.
4. If $1-\varepsilon > r_F$, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$.

Running these steps, along with a classification, allows to certify that the classification has not been affected by the noise, i.e. that the same classification outcome would have been obtained on the noiseless input state.

9.5.3 *Robustness for known Noise Models*

Now, we analyse the robustness of a quantum classifier against known noise models which are parametrized by a noise parameter γ. Specifically, we investigate robustness against phase damping and amplitude damping. Using Theorem 10, we calculate the fidelity between the clean input σ and the noisy input $\mathcal{N}_\gamma(\sigma)$ and rearrange the robustness condition (237) such that it yields a bound on the maximal noise which the classifier tolerates.

9.5.3.1 *Phase Damping*

Phase damping describes the loss of quantum information without loosing energy. For example, it describes how electronic states in an atom are perturbed upon interacting

with distant electrical charges. The quantum channel corresponding to this noise model can be expressed in terms of Kraus operators which are given by

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}, \quad K_1 = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{\gamma} \end{pmatrix} \tag{257}$$

where γ is the noise parameter. From this description alone, we can see that a system which is in the $|0\rangle$ or $|1\rangle$ state is always robust against all noise parameters in this model as it acts trivially on $|0\rangle$ and $|1\rangle$. Any such behaviour should hence be reflected in the tight robustness condition we derive from QHT. Indeed, for a pure state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, Theorem 10 leads to the robustness condition $\gamma \leqslant 1$ if $\alpha = 0$ or $\beta = 0$ and, for any $\alpha, \beta \neq 0$,

$$\gamma < 1 - \left(\max\left\{ 0, 1 + \frac{r_F - 1}{2\,|\alpha|^2\,|\beta|^2} \right\} \right)^2 \tag{258}$$

where $r_F = \frac{1}{2}(1 + \sqrt{g(p_A, p_B)})$ is the fidelity bound from Theorem 10 and p_A, p_B are the corresponding class probability bounds. This bound is illustrated in Figure 24 as a function of $|\alpha|^2$ and p_A. The expected behaviour towards the boundaries can be seen in the plot, namely that when $|\alpha|^2 \to \{0, 1\}$, then the classifier is robust under all noise parameters $\gamma \leqslant 1$.

9.5.3.2 *Amplitude Damping*

Amplitude damping models effects which occur due to the loss of energy from a quantum system (energy dissipation). For example, it can be used to model the dynamics of an atom which spontaneously emits a photon. The quantum channel corresponding to this noise model can be written in terms of Kraus operators

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}, \quad K_1 = \begin{pmatrix} 0 & \sqrt{\gamma} \\ 0 & 0 \end{pmatrix}, \tag{259}$$

where γ is the noise parameter and can be interpreted as the probability of losing a photon. It is clear from the Kraus decomposition that the $|0\rangle$ state remains unaffected. This again needs to be reflected by a tight robustness condition. For a pure state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, Theorem 10 leads to the robustness condition $\gamma \leqslant 1$ if $|\alpha| = 1$ and, for any $\alpha, \beta \neq 0$,

$$\gamma < 1 - \left[\frac{|\alpha|^2}{|\alpha|^2 - |\beta|^2} \cdot \left(1 - \sqrt{1 - \frac{|\alpha|^2 - |\beta|^2}{|\alpha|^2\,|\beta|^2} \cdot \frac{\max\{0, r_F - |\alpha|^2\}}{|\alpha|^2}} \right) \right]^2 \tag{260}$$

where again $r_F = \frac{1}{2}(1 + \sqrt{g(p_A, p_B)})$ is the fidelity bound from Theorem 10. This bound is illustrated in Figure 24 as a function of $|\alpha|^2$ and p_A. It can be seen again that the bound shows the expected behaviour, namely that when $|\alpha|^2 \to 1$, then the classifier is robust under all noise parameters $\gamma \leqslant 1$.

We remark that, in contrast to the previous protocol, here we assume access to the noiseless state σ and we compute the robustness condition on the noise parameter based on the classification of this noiseless state. This can be used in a scenario where a quantum classifier is developed and tested on one device, but deployed on a different device with different noise sources.

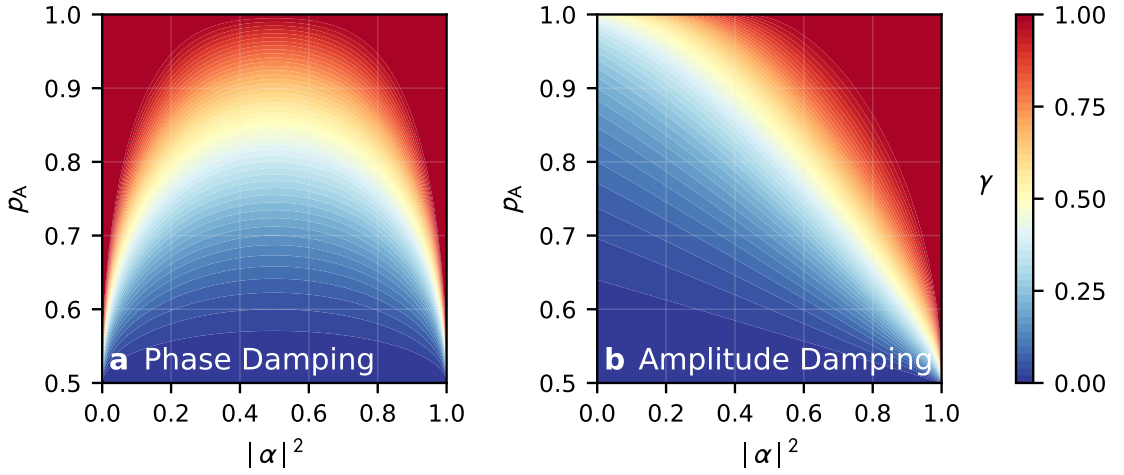


Figure 24: Robustness against known noise models. Both plots show the maximal noise parameter γ for which the classifier $\mathcal{A}$ is still guaranteed to be robust, for **a** phase damping and **b** amplitude damping, when classifying a pure state input $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. In **a**, we can see that for states $|0\rangle$ and $|1\rangle$, the classifier is robust against any $\gamma \leqslant 1$, while for **b** the same holds if the input state is $|1\rangle$.

9.5.4 *Randomized inputs with depolarization smoothing*

In the previous section, we looked at robustness of quantum classifiers against certain types of noise, either with respect to a known noise model, or with respect to unknown, potentially adversarial, noise. Here we take a different viewpoint, and investigate how robustness against unknown noise sources can be enhanced by harnessing depolarization noise. This is led by the intuition that noise can be exploited to increase robustness and privacy. We first present provable robustness in terms of trace distance which is equivalent to the robustness condition (216) from Theorem 8 but with depolarized inputs. The bound is then compared numerically with the Hölder duality bound from Lemma 11 and with a result obtained recently from quantum differential privacy [50].

Quantum channel smoothing: depolarization. Consider depolarization noise which maps a state σ onto a linear combination of itself and the maximally mixed state

$$\sigma \mapsto \mathcal{E}_p^{\mathrm{dep}}(\sigma) := (1-p)\sigma + \frac{p}{d}\mathbb{1}_d \tag{261}$$

where $p \in (0, 1)$ is the depolarization parameter and d is the dimensionality of the underlying Hilbert space. In single-qubit scenarios, this can geometrically be interpreted as a uniform contraction of the Bloch sphere parametrized by p, pushing quantum states towards the completely mixed state. Analogously to classical randomized smoothing, we apply a depolarization channel to inputs before passing them through the classifier in order to artificially randomize the states and increase robustness against adversarial attacks. We then obtain a robustness guarantee by instantiating Theorem 8 in the following way. Let σ be a benign input state and suppose that the classifier $\mathcal{A}$ with score function $\mathbf{y}$ satisfies

$$\mathbf{y}_{k_A}(\mathcal{E}_p^{\mathrm{dep}}(\sigma)) \geqslant p_A > p_B \geqslant \max_{k \neq k_A} \mathbf{y}_k(\mathcal{E}_p^{\mathrm{dep}}(\sigma)). \tag{262}$$

Then $\mathcal{A}$ is robust at $\mathcal{E}_p^{\text{dep}}(\rho)$ for any adversarial input state ρ which satisfies the robustness condition (216), where β^* is the optimal type-II error probability for testing $\mathcal{E}_p^{\text{dep}}(\sigma)$ against $\mathcal{E}_p^{\text{dep}}(\rho)$. In particular, if σ and ρ are single-qubit pure states and in the case where we have $p_A + p_B = 1$, the robustness condition can be equivalently expressed in terms of the trace distance. We formalize this result in the following Corollary and provide a proof in Section F.1.2.

Corollary 7 (Depolarised single-qubit pure states). *Let $|\psi_\sigma\rangle, |\psi_\rho\rangle \in \mathbb{C}^2$ be single-qubit pure sates and let $\mathcal{E}_p^{\text{dep}}$ be a depolarising channel with noise parameter $p \in (0, 1)$. Then, if $p_A > 1/2$ and $p_B = 1 - p_A$, the robustness condition (216) for $\mathcal{E}_p^{\text{dep}}(\sigma)$ and $\mathcal{E}_p^{\text{dep}}(\rho)$ is equivalent to*

$$\frac{1}{2}\left\||\psi_\sigma\rangle\langle\psi_\sigma| - |\psi_\rho\rangle\langle\psi_\rho|\right\|_1 < r_Q(p) \tag{263}$$

where

$$r_Q(p) = \begin{cases} \sqrt{\frac{1}{2} - \frac{\sqrt{g(p,p_A)}}{1-p}}, & p_A < \frac{1+3(1-p)^2}{2+2(1-p)^2} \\ \sqrt{\frac{p\cdot(2-p)\cdot(1-2p_A)^2}{8(1-p)^2\cdot(1-p_A)}}, & p_A \geqslant \frac{1+3\cdot(1-p)^2}{2+2\cdot(1-p)^2} \end{cases} \tag{264}$$

with $g(p, p_A) = \frac{1}{2}\left(2p_A(1-p_A) - p(1-\frac{p}{2})\right)$.

The Hölder bound from Lemma 11 can also be adapted to the noisy setting. Specifically, since for two states σ and ρ, the trace distance obeys

$$T(\mathcal{E}_p^{\text{dep}}(\rho), \mathcal{E}_p^{\text{dep}}(\sigma)) = (1-p)\cdot T(\rho, \sigma), \tag{265}$$

Lemma 11 implies robustness given that the trace distance is less than $T(\rho, \sigma) < r_H(p)$ where

$$r_H(p) = \frac{2p_A - 1}{2(1-p)}. \tag{266}$$

It has been shown in [50] that naturally occurring noise in a quantum circuit can be harnessed to increase the robustness of quantum classification algorithms. Specifically, using techniques from quantum differential privacy, a robustness bound expressible in terms of the class probabilities p_A and the depolarization parameter p has been derived. Written in our notation and for single-qubit binary classification, the bound can be written as

$$r_{DP}(p) = \frac{p}{2(1-p)}\left(\sqrt{\frac{p_A}{1-p_A}} - 1\right) \tag{267}$$

and robustness is guaranteed for any adversarial state ρ with $T(\rho, \sigma) < r_{DP}(p)$. The three bounds are compared graphically in Figure 25 for different values of the noise parameter p, showing that the QHT bound gives rise to a tighter robustness condition for all values of p.

It is worth remarking that although the QHT robustness bounds can be, as shown here for the case of applying depolarization channel, enhanced by active input randomization, it already presents a valid, non-trivial condition with noiseless (without smoothing) quantum input (Theorems 8, 10, Corollary 6 and Lemma 11). This contrasts

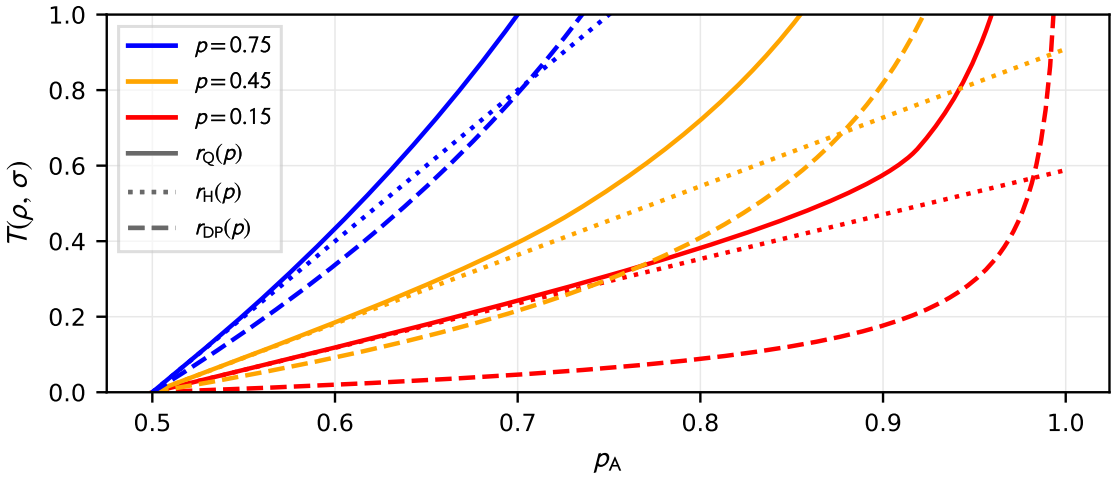


Figure 25: Comparison of Robustness bounds for single-qubit pure states derived from quantum hypothesis testing $r_Q(p)$, Hölder duality $r_H(p)$ and quantum differential privacy $r_{DP}(p)$ [50] with different levels of depolarization noise p.

with the deterministic classical scenario, where the addition of classical noise sources to the input state is necessary to generate a probability distribution corresponding to the input data, from which an adversarial robustness bound can be derived [39]. This distinction between the quantum and classical settings roots in the probabilistic nature of measurements on quantum states, which of course applies to both pure and mixed state inputs.

9.6 CONCLUSION

In this chapter of the thesis, we have investigated our fourth research question and sought to derive robustness guarantees for ML systems which are based on quantum machine learning. To that end, we have seen how a fundamental connection between adversarial robustness of quantum classifiers and QHT can be leveraged to provide a powerful framework for deriving tight conditions for robustness certification. The robustness condition is provably tight when expressed in the SDP formulation in terms of optimal error probabilities for binary classifications or, more generally, for multiclass classifications where the probability of the most likely class is greater than 1/2. The corresponding closed form expressions arising from the SDP formulation are proved to be tight for general states when expressed in terms of fidelity and Bures distance, whereas in terms of trace distance, tightness holds only for pure states. These bounds give rise to (1) a practical robustness protocol for assessing the resilience of a quantum classifier against adversarial and unknown noise sources; (2) a protocol to verify whether a classification given a noisy input has had the same outcome as a classification given the noiseless input state, without requiring access to the latter, and (3) conditions on noise parameters for amplitude and phase damping channels, under which the outcome of a classification is guaranteed to remain unaffected. Furthermore, we have shown how using a randomized input with depolarization channel enhances the QHT bound, consistent with previous results, in a manner akin to classical randomized smoothing.

A key difference between the quantum and classical formalism is that quantum states themselves have a naturally probabilistic interpretation, even though the classical data

that could be embedded in quantum states do not need to be probabilistic. Our quantum adversarial robustness bound can be shown independently of randomized input, even though it can be enhanced by it, like through a depolarization channel. In contrast, currently known classical probabilistic robustness guarantees rely on active input randomization what induces further challenges such as specific training protocols and degraded clean accuracy.

Our tight robustness guarantees and the connection to quantum hypothesis testing also provide a first step towards more rigorously identifying the limitations of quantum classifiers in their power of distinguishing quantum states. Our formalism hints at an intimate relationship between these fundamental limitations in the accuracy of distinguishing between different classes of states and robustness. This could shed light on the robustness and accuracy trade-offs observed in classification protocols [224] and is an important direction of future research. It is also of independent interest to explore possible connections between tasks that use quantum hypothesis testing, such as quantum illumination [250] and state discrimination [195], with accuracy and robustness in quantum classification.

10 ROBUSTNESS INTERVALS FOR QUANTUM EXPECTATION VALUES

In the previous chapter we have focused on establishing robustness guarantees for quantum classifiers. In other words, we derived conditions, under which the predicted class of a quantum state remains constant. Here we take a different view and characterize the error when estimating expectation values of quantum observables when the underlying quantum states are perturbed. In essence, this approach is very similar to the results presented in Chapter 8 where we have bounded the change of expectation values due to shifts in the underlying data distributions. Indeed, the core technique used to derive the results in Chapter 8 is a special case of the Gramian technique used in this chapter, and introduced in Section 5.2.

10.1 INTRODUCTION

10.1.1 *Overview*

Today's quantum computers are characterized by a low count of noisy qubits performing imperfect operations in a limited coherence time. In this era of quantum computing, the NISQ era [174], researchers and practitioners alike strive to heuristically harness limited quantum resources in order to solve classically difficult problems and also to benchmark and potentially develop new quantum subroutines. A typical pattern of these NISQ algorithms [12], exemplified by the seminal VQE [173] and quantum approximate optimization algorithm [62], consists of the preparation of ansatz states with a parameterized unitary circuit followed by useful classical output being extracted by means of quantum measurements, more generally as expectation values of quantum observables through repeated measurements.

The promising potential of these NISQ algorithms spans across a wide spectrum of applications, ranging from quantum chemistry, many-body physics, and machine learning to optimization and finance [12]. However, as a consequence of their heuristic nature and the prevalent imperfections in near-term implementation, NISQ algorithms in practice typically produce outputs deviating from the exact and ideal setting. This unfortunate hindrance practically arises from various sources such as circuit noise and decoherence [174], limited expressibility of Ansätze [160, 201], barren plateaus during optimization in variational hybrid quantum-classical algorithms [150, 167, 236], measurement noise and other experimental imperfections [102, 241]. To determine the usefulness of a given NISQ application, it is thus crucial to quantify the error on the final output in the presence of a multitude of the aforementioned sources of imperfection.

10.1.2 *Contributions*

In this work, we endeavour to systematically certify the reliability of quantum algorithms by deriving robustness bounds for expectation values of observables on ap-

proximations of a target state. To that end, based on analytical solutions to an SDP problem, we present lower and upper bounds to expectation values of quantum observables which depend only on the fidelity with the target state and post-processing of previously obtained measurement results. Furthermore, we take into account higher statistical moments of the observable by generalizing the Gramian method for pure states [243] to generic density operators, thus extending its application to bounding output errors resulting from noisy circuits. Although the focus of our investigation is on errors arising from circuit imperfection, the underlying techniques are also valid for other sources of errors such as algorithmic shortcomings. Finally, we apply these bounds to numerically obtain robustness intervals on simulated noisy and noiseless VQE for ground state energy estimation of electronic structure Hamiltonians of several molecules. The robustness certification protocol resulting from this work is integrated with the open source Tequila [114] library.

The remainder of this chapter is organised as follows. In Section 10.2, we present our main results, namely, the bounds based on the SDP formulation and the Gramian method. In Section 10.3, we present numerical simulations and explain the applicability of our bounds in the context of VQE. Section 10.3.2 highlights the implementation in Tequila and concluding remarks are given in Section 10.4.

10.2 ROBUSTNESS INTERVALS

The goal of this work is to provide techniques to compute intervals which are guaranteed to contain the expectation value of an observable A under an ideal, but unavailable, target state ρ. Any such interval is referred to as a robustness interval. More formally, instead of having access to the state ρ, we assume access to the approximate state σ, which is further assumed to have at least fidelity $1-\epsilon$ with the target state ρ. Given these assumptions, we define a robustness interval to be an interval $\mathcal{I} = [\underline{\chi}, \overline{\chi}] \subseteq \mathbb{R}$ for which it is guaranteed that

$$\underline{\chi}(\epsilon, \sigma, A) \leqslant \mathrm{Tr}[A\rho] \leqslant \overline{\chi}(\epsilon, \sigma, A) \tag{268}$$

and which is a function of the infidelity ϵ, the observable A, and the state σ.

NOTATION The Hilbert space corresponding to the quantum system of interest is denoted by $\mathcal{H} \equiv \mathbb{C}^d$ with dimension $d = 2^n$. We use the Dirac notation for quantum states, i.e. elements of $\mathcal{H}$ are written as kets $|\psi\rangle \in \mathcal{H}$ with the dual written as a bra $\langle\psi|$. The space of linear operators acting on elements of $\mathcal{H}$ is denoted by $\mathcal{L}(\mathcal{H})$ and elements thereof are written in capital letters $A \in \mathcal{L}(\mathcal{H})$. The set of density operators on $\mathcal{H}$ is written as $\mathcal{S}(\mathcal{H}) \subset \mathcal{L}(\mathcal{H})$ and lower case greek letters are used to denote its elements $\sigma \in \mathcal{S}(\mathcal{H})$ which are positive semidefinite and have trace equal to 1. For an element $A \in \mathcal{L}(\mathcal{H})$ we write $A \geqslant 0$ if it is positive semidefinite, A^T stands for its transpose, and $A^\dagger$ is the adjoint. We also use the Loewner partial order on the space of Hermitian operators, i.e. for two Hermitian operators $A, B \in \mathcal{L}(\mathcal{H})$, we write $A \geqslant B$ if and only if $A - B \geqslant 0$. Expectation values of observables, i.e. Hermitian operators $A \in \mathcal{L}(\mathcal{H})$, are written as $\langle A\rangle_\sigma = \mathrm{Tr}[A\sigma]$ for some $\sigma \in \mathcal{S}(\mathcal{H})$. The variance of an observable is given by $(\Delta A_\sigma)^2 = \langle A^2\rangle_\sigma - \langle A\rangle_\sigma^2$. We write $\|A\|_1 = \mathrm{Tr}[|A|]$ with $|A| = \sqrt{A^\dagger A}$ for the trace norm of an operator $A \in \mathcal{L}(\mathcal{H})$. The fidelity between quantum states $\sigma, \rho \in \mathcal{S}(\mathcal{H})$ is defined as $\mathcal{F}(\rho, \sigma) = \max_{\psi_\rho, \psi_\sigma} |\langle\psi_\rho|\psi_\sigma\rangle|^2$ where the maximum is taken over all

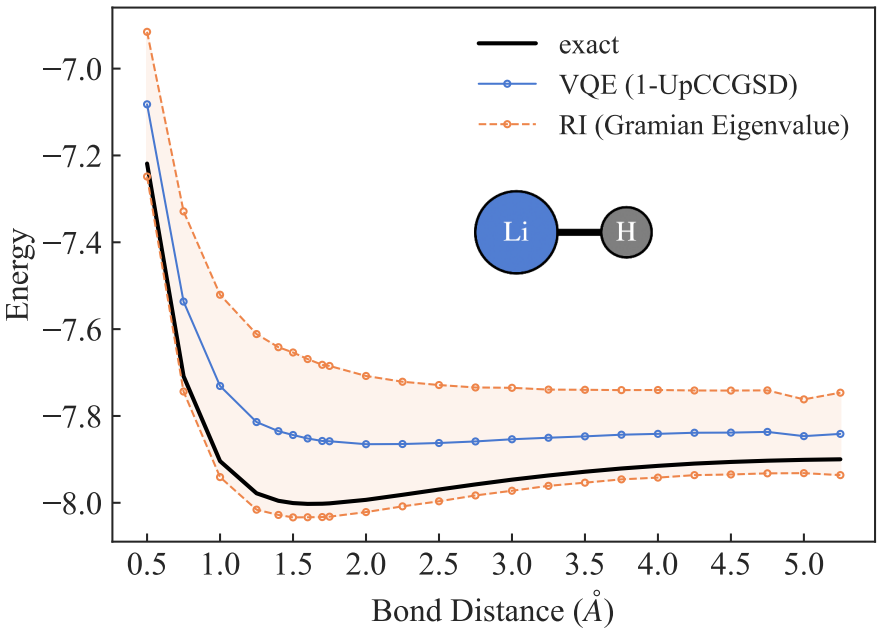


Figure 26: Bond dissociation curves and robustness interval (RI) for Lithium Hydride in a basis-set-free approach [117, 118]. The exact, theoretical energies are shown in black, the energy estimates provided by a noisy VQE with a UpCCGSD Ansatz [130] is shown in blue. The robustness interval is guaranteed to contain the true ground state energy and is based on the Gramian eigenvalue bounds for mixed states (Theorem 12).

purifications of ρ and σ. For pure states, the fidelity reduces to the squared overlap $\mathcal{F}(|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi|) = |\langle\psi|\phi\rangle|^2$. Finally, the real part of a complex number $z \in \mathbb{C}$ is written as $\Re(z)$ and the imaginary part as $\Im(z)$.

SUMMARY OF TECHNICAL RESULTS In this chapter, we apply the techniques introduced in Chapter 5 to the problem of providing an interval guarantee for quantum algorithms whose output relies on measurements of expectation values of quantum observables. The first technique is essentially based on the formulation of lower and upper bounds as SDP problems and makes use of the closed form solution of optimal type-II error probabilities from QHT presented in Lemma 3. The second technique, presented in Lemma 4 is based on the non-negativity of the determinant of Gram matrices for a suitable collection of vectors. This second technique was initially proposed by Weinhold [243] in the context of pure states. Using Uhlmann's Theorem [225], which relates the fidelity between two mixed states to the trace norm, we have extend these results to mixed states. This ultimately enables their applicability in the current NISQ era, where the assumption of a closed quantum system is violated and one needs to make use of the density operator formalism to accurately model these states and their evolutions.

In Table 10, we summarize our results, together with the conditions under which they apply and the quantities that are covered. Figure 26 shows the ground state energies of molecular Lithium Hydride in the basis-set-free approach of [117, 118], with energy estimates provided by VQE with an UpCCGSD ansatz. The lower and upper bounds on the true energy are obtained via the Gramian method from Theorem 12.

Table 10: Overview of bounds for the true expectation values and eigenvalues of a Hermitian operator A, with ρ the target state and σ the approximation. For the eigenvalue bound, $\rho = |\psi\rangle\langle\psi|$ is the density operator corresponding to the eigenstate $|\psi\rangle$ with eigenvalue $\lambda = \langle\psi|A|\psi\rangle$. We remark that the SDP lower and upper bounds are valid for fidelities with $\mathcal{F}(\rho, \sigma) \geqslant 1-\epsilon$ for $\epsilon \geqslant 0$ such that $\epsilon \leqslant \frac{1}{2}(1+\langle A\rangle_\sigma)$ and $\epsilon \leqslant \frac{1}{2}(1-\langle A\rangle_\sigma)$, respectively. The Gramian lower bound for expectation values is valid for $\epsilon \geqslant 0$ with $\sqrt{(1-\epsilon)/\epsilon} \geqslant \Delta A_\sigma / \langle A\rangle_\sigma$.

	SDP	Gramian	
	Expectation $\langle A\rangle_\rho$	Expectation $\langle A\rangle_\rho$	Eigenvalue λ
Lower Bound	$(1-2\epsilon)\langle A\rangle_\sigma - 2\sqrt{\epsilon(1-\epsilon)(1-\langle A\rangle_\sigma^2)}$	$(1-2\epsilon)\langle A\rangle_\sigma - 2\sqrt{\epsilon(1-\epsilon)}\Delta A_\sigma + \frac{\epsilon\langle A^2\rangle_\sigma}{\langle A\rangle_\sigma}$	$\langle A\rangle_\sigma - \Delta A_\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}$
Upper Bound	$(1-2\epsilon)\langle A\rangle_\sigma + 2\sqrt{\epsilon(1-\epsilon)(1-\langle A\rangle_\sigma^2)}$	—	$\langle A\rangle_\sigma + \Delta A_\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}$
Assumptions	$-\mathbb{1} \leqslant A \leqslant \mathbb{1}$	$A \geqslant 0$	$\sigma = \|\psi\rangle\langle\psi\| \wedge A|\psi\rangle = \lambda|\psi\rangle$

10.2.1 *Bounds via Semidefinite Programming*

Here we derive a robustness interval which is based on expressing lower and upper bounds as a semidefinite program which we express in terms of optimal type-II error probabilities for binary QHT. We remark that the reasoning presented here largely follows the proof of Lemma 2, although it is valid for observables $-\mathbb{1} \leqslant A \leqslant \mathbb{1}$ which cover the important Pauli operators.

10.2.1.1 *Robustness Interval*

Consider a bounded observable $-\mathbb{1} \leqslant A \leqslant \mathbb{1}$ and let σ be the approximate state, corresponding to the alternative hypothesis, and let ρ be the target state, corresponding to the null hypothesis. We can express lower and upper bounds to $\langle A\rangle_\rho$ as semidefinite programs which take into account measurements of σ. Namely, we have the upper bound

$$\langle A\rangle_\rho \leqslant \sup_{-\mathbb{1}\leqslant\Lambda\leqslant\mathbb{1}} \{\langle\Lambda\rangle_\rho : \langle\Lambda\rangle_\sigma = \langle A\rangle_\sigma\} \tag{269}$$

and the lower bound

$$\langle A\rangle_\rho \geqslant \inf_{-\mathbb{1}\leqslant\Lambda\leqslant\mathbb{1}} \{\langle\Lambda\rangle_\rho : \langle\Lambda\rangle_\sigma = \langle A\rangle_\sigma\}. \tag{270}$$

It is straight forward to see that these optimization problems are indeed valid lower and upper bounds to $\langle A\rangle_\rho$ by noting that the operator A is feasible. In addition, as shown in Section 10.2.1.2, the tightness of the bounds is an immediate consequence of the formulation of the robustness interval as an SDP. We can rewrite these SDPs and express them in terms of optimal type-II error probabilities, so that the upper bound reads

$$\begin{aligned}\langle A\rangle_\rho &\leqslant \sup_{-\mathbb{1}\leqslant\Lambda\leqslant\mathbb{1}} \{\langle\Lambda\rangle_\rho : \langle\Lambda\rangle_\sigma = \langle A\rangle_\sigma\} \\ &= 1 - 2\beta^*\left(\frac{1+\langle A\rangle_\rho}{2}; \sigma, \rho\right)\end{aligned} \tag{271}$$

and, similarly, for the lower bound

$$\begin{aligned}\langle A\rangle_\rho &\geqslant \inf_{-\mathbb{1}\leqslant\Lambda\leqslant\mathbb{1}}\{\langle\Lambda\rangle_\rho : \langle\Lambda\rangle_\sigma = \langle A\rangle_\sigma\}\\ &= 2\beta^*\left(\frac{1-\langle A\rangle_\rho}{2}; \sigma, \rho\right) - 1.\end{aligned} \tag{272}$$

These robustness bounds formalize the intuition that states which are hard to discriminate, i.e., which admit higher error probabilities, will have expectation values which are closer together. Furthermore, this connection also has the interesting interpretation that, if the approximate expectation $\langle A\rangle_\sigma$ is close to the extreme -1, a statistical hypothesis test is restricted to have type-I error probability close to 0. This makes it harder for the corresponding optimal type-II error probability β^* to be low and hence $\langle A\rangle_\rho$ will generally be closer to $\langle A\rangle_\sigma$. Finally, since Lemma 3 provides a closed form solution to the SDP β^*, which only depends on the fidelity between ρ and σ, we can establish a robustness interval of the form (268). This result is summarized in the following Theorem.

Theorem 11. *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H}_d)$ be density operators with $\mathcal{F}(\rho, \sigma) \geqslant 1-\epsilon$ for some $\epsilon \geqslant 0$. Let A be an observable with $-\mathbb{1} \leqslant A \leqslant \mathbb{1}$ and with expectation value $\langle A\rangle_\rho$ under ρ. For $\epsilon \leqslant \frac{1}{2}(1+\langle A\rangle_\sigma)$, the lower bound of $\langle A\rangle_\rho$ can be expressed as*

$$\langle A\rangle_\rho \geqslant (1-2\epsilon)\langle A\rangle_\sigma - 2\sqrt{\epsilon(1-\epsilon)(1-\langle A\rangle_\sigma^2)}. \tag{273}$$

Similarly, for $\epsilon \leqslant \frac{1}{2}(1-\langle A\rangle_\sigma)$, the upper bound of $\langle A\rangle_\rho$ becomes

$$\langle A\rangle_\rho \leqslant (1-2\epsilon)\langle A\rangle_\sigma + 2\sqrt{\epsilon(1-\epsilon)(1-\langle A\rangle_\sigma^2)}. \tag{274}$$

In practice, it is typically not feasible to measure the exact value of $\langle A\rangle_\sigma$ due to finite sampling errors, measurement noise, and other experimental imperfections. For this reason, one needs to rely on confidence intervals which contain the exact value of $\langle A\rangle_\sigma$ with high probability. This can be accounted for in the bounds from Theorem 11 by noting that they are monotonic in $\langle A\rangle_\sigma$, what allows us to replace the exact value by bounds which hold with high probability. Finally, it is worth noting that, if one has access to an estimate of the fidelity, i.e. some $\epsilon > 0$ with $\mathcal{F}(\rho, \sigma) \geqslant 1-\epsilon$, this interval can be calculated by post-processing previous measurement results, and hence does not cause any computational overhead. We now state the proof of this result which follows directly from applying the bound the optimal type-II error probability presented in Lemma 3 in Section 5.1.

Proof of Theorem 11. We first formulate the robustness bounds as SDP problems which take into account the first moment of A under σ and the assumption that $-\mathbb{1} \leqslant A \leqslant \mathbb{1}$. This is then connected to the SDP problem (41) from QHT for which we have established a closed form lower bound in Lemma 3. We start with the upper bound. Consider the optimization problem

$$\max\{\mathrm{Tr}\left[\Lambda\rho\right] \,|\, \mathrm{Tr}\left[\Lambda\sigma\right] = \mathrm{Tr}\left[A\sigma\right], -\mathbb{1} \leqslant \Lambda \leqslant \mathbb{1}\} \tag{275}$$

which is an upper bound to $\langle A\rangle_\rho$ since the operator A is feasible. We can rewrite this as

$$\begin{aligned}&\max\{\mathrm{Tr}\left[\Lambda\rho\right] | \mathrm{Tr}\left[\Lambda\sigma\right] = \mathrm{Tr}\left[A\sigma\right], -\mathbb{1} \leqslant \Lambda \leqslant \mathbb{1}\}\\ &\quad= -1 + 2\max\left\{\mathrm{Tr}\left[\tilde{\Lambda}\rho\right] | \mathrm{Tr}\left[\tilde{\Lambda}\sigma\right] = \frac{1}{2}(1+\mathrm{Tr}\left[A\sigma\right]), 0 \leqslant \tilde{\Lambda} \leqslant \mathbb{1}\right\}\\ &\quad= 1 - 2\beta^*\left(\frac{1}{2}(1+\mathrm{Tr}\left[A\sigma\right]); \sigma, \rho\right)\end{aligned} \tag{276}$$

where the second equality follows from the fact that replacing the equality with an inequality in the constraint of the SDP problem (41) leads to the same solution. It then follows from Lemma 3 that

$$\langle A\rangle_\rho \leqslant (1-2\epsilon)\langle A\rangle_\sigma + 2\sqrt{\epsilon(1-\epsilon)(1-\langle A\rangle_\sigma^2)} \tag{277}$$

for $\epsilon \geqslant 0$ with $1-\epsilon \geqslant \frac{1}{2}(1+\langle A\rangle_\sigma)$. To show the lower bound, consider the optimization problem

$$\min\{\mathrm{Tr}\left[\Lambda\rho\right] | \mathrm{Tr}\left[\Lambda\sigma\right] = \mathrm{Tr}\left[A\sigma\right], -\mathbb{1} \leqslant \Lambda \leqslant \mathbb{1}\} \tag{278}$$

which is a lower bound to $\langle A\rangle_\rho$ since the operator A is feasible. Analogous to the derivation of the upper bound, we rewrite this as

$$\begin{aligned}&\min\{\mathrm{Tr}\left[\Lambda\rho\right] | \mathrm{Tr}\left[\Lambda\sigma\right] = \mathrm{Tr}\left[A\sigma\right], -\mathbb{1} \leqslant \Lambda \leqslant \mathbb{1}\}\\ &\quad= 2\min\left\{\mathrm{Tr}\left[(\mathbb{1}-\tilde{\Lambda})\rho\right] | \mathrm{Tr}\left[\tilde{\Lambda}\sigma\right] = \frac{1}{2}(1-\mathrm{Tr}\left[A\sigma\right]), 0 \leqslant \tilde{\Lambda} \leqslant \mathbb{1}\right\} - 1\\ &\quad= 2\beta^*\left(\frac{1}{2}(1-\mathrm{Tr}\left[A\sigma\right]); \sigma, \rho\right) - 1\end{aligned} \tag{279}$$

and again use Lemma 3 and obtain the lower bound

$$\langle A\rangle_\rho \geqslant (1-2\epsilon)\langle A\rangle_\sigma - 2\sqrt{\epsilon(1-\epsilon)(1-\langle A\rangle_\sigma^2)} \tag{280}$$

for $\epsilon \geqslant 0$ with $1-\epsilon \geqslant \frac{1}{2}(1-\langle A\rangle_\sigma)$. This concludes the proof. □

10.2.1.2 *Tightness*

Stemming from the formulation as an SDP, the bounds in Theorem 11 are tight for pure states in the sense that, for each bound, there exists an observable A with expectation $\langle A\rangle_\sigma$ under σ, and whose expectation under ρ saturates the bound. Here we give a formal, constructive proof of this statement for pure states and in the case where $\mathcal{F}(\rho, \sigma) = 1-\epsilon$. Let us first consider the upper bound and recall that in the proof of Theorem 11 we have shown

$$\begin{aligned}\langle A\rangle_\rho &\leqslant 1 - 2\beta^*\left(\frac{1}{2}(1+\mathrm{Tr}\left[A\sigma\right]); \sigma, \rho\right)\\ &\leqslant (1-2\epsilon)\langle A\rangle_\sigma + 2\sqrt{\epsilon(1-\epsilon)(1-\langle A\rangle_\sigma^2)}\end{aligned} \tag{281}$$

where the last inequality follows from Lemma 3. Additionally, for pure states, this inequality is indeed an *equality*. That is, we have shown that for pure states we have

$$\beta^*(\alpha_0; \sigma, \rho) = \alpha_0(2\epsilon-1) + (1-\epsilon) - 2\sqrt{\epsilon\alpha_0(1-\epsilon)(1-\alpha_0)} \tag{282}$$

for $1-\epsilon \geqslant \alpha_0$ for arbitrary $\alpha_0 \in [0, 1]$. Let M_0 be a Helstrom operator such that $\mathrm{Tr}\left[\rho(\mathbb{1} - M_0^\star)\right] = \beta^*(\alpha_0; \sigma, \rho)$ and $\alpha_0 = \mathrm{Tr}[M_0\sigma]$, and let $A^\star := 2M_0 - \mathbb{1}$. The Helstrom operator M_0 exists and is well-defined as we have shown in Section B.1.1. Note that

$$\begin{aligned}\langle A^\star\rangle_\rho &= \mathrm{Tr}\left[\rho(2M_0 - \mathbb{1})\right] \\ &= 1 - 2\mathrm{Tr}\left[\rho(\mathbb{1} - M_0)\right] \\ &= 1 - 2\beta^*\left(\mathrm{Tr}[M_0\sigma]\right); \sigma, \rho)\end{aligned} \tag{283}$$

which shows that the bound is saturated for the observable $A^\star$ and hence tight. Tightness of the lower bound can be shown analogously.

10.2.2 *Bounds via non-negativity of the Gramian*

Here we study the bounds based on the Gramian technique, presented in Lemma 4 in Section 5.2, and adapt these bounds to the case where the target state is an eigenstate of the observables leading to a significant tightening of the bounds for these cases. As highlighted in Section 5.2, the Gramian technique was pioneered by Weinhold [243] in the context of pure states and problems related to quantum chemistry. However, this restriction to pure states hinders the applicability of this method in practice and, in particular, in the current NISQ era, where one often has to deal with noisy states that are expressed as mixed states. This motivates us to extend the result to the mixed state case, the result of which we restate here for completeness.

Lemma 4 (restated). *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H}_d)$ be density operators with fidelity $\mathcal{F}(\rho, \sigma) \geqslant 1-\epsilon$ for some $\epsilon \geqslant 0$ and let $A \geqslant 0$ be an observable. Let $\underline{m}, \bar{v} \geqslant 0$ such that*

$$\langle A\rangle_\sigma \geqslant \underline{m} \quad \textit{and} \quad \Delta A_\sigma \leqslant \bar{v}. \tag{284}$$

For ϵ with $\epsilon \leqslant \frac{\overline{m}^2}{\overline{m}^2+\bar{v}^2}$, a lower bound of $\langle A\rangle_\rho$ can be expressed as

$$\langle A\rangle_\rho \geqslant (1-\epsilon)\overline{m} - 2\bar{v}\sqrt{\epsilon(1-\epsilon)} + \epsilon\frac{\bar{v}^2}{\overline{m}}. \tag{285}$$

In the case where the target state σ is an eigenstate of an observable A, the Gramian method allows to derive a further bound. While the assumptions here are stronger, this bound is particularly useful in applications such as the variational quantum eigensolver and when the observable of interest commutes with a Hamiltonian H for which the target state is an eigenstate. Formally, we have the following result:

Theorem 12 (Eigenvalues). *Let $\sigma \in \mathcal{S}(\mathcal{H}_d)$ be a density operator and let A be an arbitrary observable with eigenstate $|\psi\rangle$ and eigenvalue λ, $A|\psi\rangle = \lambda|\psi\rangle$. Suppose that $\epsilon \geqslant 0$ is such that $\mathcal{F}(\sigma, |\psi\rangle) = \langle\psi|\sigma|\psi\rangle \geqslant 1-\epsilon$. Then, lower and upper bounds for λ can be expressed as*

$$\langle A\rangle_\sigma - \Delta A_\sigma\sqrt{\frac{\epsilon}{1-\epsilon}} \;\leqslant\; \lambda \;\leqslant\; \langle A\rangle_\sigma + \Delta A_\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}. \tag{286}$$

Proof. Recall that in the proof of Lemma 4 we have shown that a slight modification of the Gramian inequalities from (52) also holds for mixed states. Specifically, we have shown that

$$\begin{aligned}\sqrt{\mathcal{F}(\rho,\sigma)}\langle A\rangle_\sigma - \Delta A_\sigma\sqrt{1-\mathcal{F}(\rho,\sigma)} \;&\leqslant\; \Re(\langle A\sqrt{\rho}, \sqrt{\sigma}U\rangle_{HS}) \\ &\leqslant\; \sqrt{\mathcal{F}(\rho,\sigma)}\langle A\rangle_\sigma + \Delta A_\sigma\sqrt{1-\mathcal{F}(\rho,\sigma)}.\end{aligned} \tag{287}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt inner product and U is a unitary such that $\mathcal{F}(\rho, \sigma) = |\langle \Phi | \Psi \rangle|^2$ with

$$|\Psi\rangle \equiv (\sqrt{\rho} \otimes \mathbb{1})|\Omega\rangle, \qquad |\Phi\rangle \equiv (\sqrt{\sigma} \otimes U^{\mathsf{T}})|\Omega\rangle. \tag{288}$$

Here, by assumption $\rho = |\psi\rangle\langle\psi|$ is pure with $|\psi\rangle$ an eigenstate of A with eigenvalue λ, $A|\psi\rangle = \lambda|\psi\rangle$. Note that in this case

$$\begin{aligned} \langle A\sqrt{\rho}, \sqrt{\sigma}U\rangle_{\mathrm{HS}} &= \lambda \, \langle \sqrt{\rho}, \sqrt{\sigma}U\rangle_{\mathrm{HS}} \\ &= \lambda \, \mathrm{Tr}\left[\sqrt{\rho}\sqrt{\sigma}U\right] \\ &= \lambda \, \langle \Omega | (\sqrt{\rho}\sqrt{\sigma}U \otimes \mathbb{1})|\Omega\rangle \\ &= \lambda \, \langle \Omega | (\sqrt{\rho}\sqrt{\sigma} \otimes U^{\mathsf{T}})|\Omega\rangle \\ &= \lambda \, \langle \Psi | \Phi \rangle \end{aligned} \tag{289}$$

where $|\Phi\rangle$ and $|\Psi\rangle$ are the purifications of σ and ρ given in the proof of Lemma 4 in (405). Without loss of generality, we assume that $\langle \Psi | \Phi \rangle$ is real and positive, since otherwise each state can be multiplied by a global phase. Dividing each side in (287) by $\langle \Psi | \Phi \rangle$ and noting that $\sqrt{\mathcal{F}(\rho, \sigma)} = \langle \Psi | \Phi \rangle$ yields

$$\langle A \rangle_\sigma - \Delta A_\sigma \sqrt{\frac{1 - \mathcal{F}(\rho, \sigma)}{\mathcal{F}(\rho, \sigma)}} \leqslant \lambda \leqslant \langle A \rangle_\sigma + \Delta A_\sigma \sqrt{\frac{1 - \mathcal{F}(\rho, \sigma)}{\mathcal{F}(\rho, \sigma)}}. \tag{290}$$

Since the RHS (LHS) of this inequality is monotonically increasing (decreasing) as $\mathcal{F}(\rho, \sigma)$ decreases, we can replace the exact fidelity by an upper bound and still get valid bounds. That is, for $\epsilon \geqslant 0$ with $\mathcal{F}(\rho, \sigma) \geqslant 1 - \epsilon$, we have

$$\langle A \rangle_\sigma - \Delta A_\sigma \sqrt{\frac{\epsilon}{1 - \epsilon}} \leqslant \lambda \leqslant \langle A \rangle_\sigma + \Delta A_\sigma \sqrt{\frac{\epsilon}{1 - \epsilon}}. \tag{291}$$

which is the desired result. □

10.2.3 *Comparison of the bounds*

We have seen three different methods to derive robustness intervals. Namely, the interval based on SDP given in Theorem 11, the expectation value lower bound based on the Gramian method from Lemma 4, and the robustness interval for eigenvalues from Theorem 12, which is also based on the Gramian method. As a first observation, we notice that the SDP bounds are dependent only on the first moment of the observable, while the bounds derived from the Gramian method take into account the second moment via the variance. In principle, this hints at a trade-off between accuracy and efficiency. That is, by taking into account higher moments, which comes at a higher computational cost, one can hope for an improvement in accuracy as more information is included. On the other hand, the SDP bounds can be calculated as a postprocessing step and thus do not require to measure additional statistics. However, as less information is included, this typically comes at the cost of lower accuracy.

On the practical side, one needs to consider that for the SDP bounds to be applicable, it is required that the observable lies between $-\mathbb{1}$ and $\mathbb{1}$. In practice, however, this is not always the case and the observable needs to be appropriately rescaled, e.g. by using its eigenvalues. As the exact eigenvalues might not be available, one needs to use lower

and upper bounds on these, which results in a loss of tightness. This is because the set of feasible points in the SDP problem from (270) and (269) becomes larger and hence loosens the bounds. A similar issue emerges in the lower bound for expectation values based on the Gramian method where the observable needs to be positive semidefinite. If this assumption is violated, one again needs to apply an appropriate transformation of the observable, leading to a potentially looser bound. Instead of scaling, one can decompose the observable into individual terms, each satisfying the constraints, and then bound each term separately and aggregate the bounds over the decomposition. In Section 10.3 we consider such a decomposition in the context of VQE. Specifically, we decompose the underlying Hamiltonian into groups of mutually commuting Pauli terms and bound the expectation of each group separately. In contrast, the eigenvalue bound based on the Gramian method does not suffer from these issues and it is applicable for general observables. It is worth remarking that this comes at the cost of less generality in the sense that the bound only applies to eigenvalues rather than general expectation values.

Assuming that the observable $A = P$ is a projection, satisfying $P^2 = P$, we can directly compare the bounds. Note that, in this case, the variance is fully determined by the first moment via $(\Delta P_\sigma)^2 = \langle P\rangle_\sigma - \langle P\rangle_\sigma^2$ and we expect that the Gramian Expectation bound should not be tighter than the SDP bound. First, we incorporate the knowledge that P is a projection in the SDP lower bound by applying it to the observable $2P - \mathbb{1}$ so that we have the bound

$$\langle P\rangle_\rho \geqslant (1-2\epsilon)\langle P\rangle_\sigma + \epsilon - 2\sqrt{\epsilon(1-\epsilon)(\langle P\rangle_\sigma - \langle P\rangle_\sigma^2)} \tag{292}$$

which is exactly the same as the lower bound derived via the Gramian method in Lemma 4 when applied to the projection P. We can also compare this bound to the Gramian eigenvalue bound from Theorem 12. Since the latter is less general, in the sense that it only holds for target states which are eigenstates, we expect this to be tighter than the expectation value bound. As can be seen from Figure 27, this is indeed the case.

Finally, we notice that all of the above bounds are faithful in the sense that, as the approximation error vanishes $\epsilon \to 0$, the bounds converge to the true expectation value $\langle A\rangle_\rho$. To compare the rate of convergence, consider the case of pure states with the target state given by $\rho = |\psi\rangle\langle\psi|$ and the approximation state $\sigma = |\phi\rangle\langle\phi|$ with $|\phi\rangle = \sqrt{1-\epsilon}|\psi\rangle + \sqrt{\epsilon}|\psi^\perp\rangle$ where $|\psi^\perp\rangle$ is orthogonal to $|\psi\rangle$ so that $\mathcal{F}(\rho, \sigma) = 1 - \epsilon$. With this, one can explicitly show that the error between each bound and the true expectation $\langle A\rangle_\rho$ scales with $\mathcal{O}(\sqrt{\epsilon})$ as $\epsilon \to 0$. For values of ϵ close to 1 on the other hand, we remark that both expectation value bounds tend towards the trivial bounds, namely 0 for the expectation value bound, and ± 1 for the SDP bounds. This ultimately stems from the underlying assumptions required for the bounds to hold. In contrast, the Gramian eigenvalue bound has no assumptions on the observable A and the bounds diverge as ϵ approaches 1.

10.2.4 *Fidelity estimation*

All bounds presented so far have in common that they depend on the fidelity with the target state ρ. However, in many practical settings, it is not possible to access the target state and thus difficult to obtain even a lower bound to the true fidelity. Here we seek

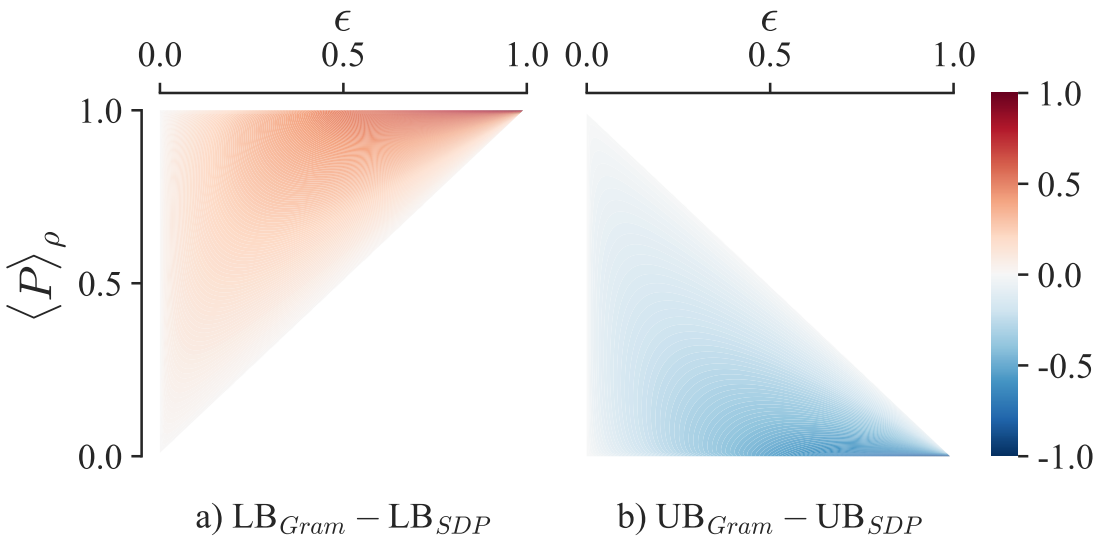


Figure 27: Difference between the SDP bounds from Theorem 11 and the Gramian eigenvalue bound from Theorem 12 as a function of the infidelity ϵ and the expectation value $\langle P\rangle_\rho$. The observable is assumed to be a projection P and the target state is an eigenstate of the observable. The difference is calculated by subtracting the SDP bound from the Gramian bound. a) shows the difference between lower bounds, b) shows the difference between the upper bounds. As can be seen from the figures, the Gramian eigenvalue bound is always more accurate than the expectation bound. Note that the Gramian expectation value lower bound (Lemma 4) equals the SDP lower bound under these assumptions.

to address this topic and present lower bounds on the true fidelity for the case where the target state is the ground state of a Hamiltonian H.

Let H be a Hamiltonian with eigenvalues $\lambda_0 \leqslant \lambda_1 \leqslant \ldots \leqslant \lambda_d$ and assume that λ_0 has geometric multiplicity 1 so that the corresponding ground state $|\psi_0\rangle$ is unique. Let σ be a possibly mixed state approximation of $|\psi_0\rangle$. If both λ_0 and λ_1 are known, one can make use of Eckart's criterion [57] to bound the fidelity via

$$\mathcal{F}(|\psi_0\rangle\langle\psi_0|, \sigma) = \langle\psi_0|\sigma|\psi_0\rangle \geqslant \frac{\lambda_1 - \langle H\rangle_\sigma}{\lambda_1 - \lambda_0}. \tag{293}$$

In scenarios where knowledge of the lowest lying eigenvalues λ_0 and λ_1 is available, one can thus directly lower-bound the fidelity and use (293) in the computation of the robustness intervals. In scenarios where one does not have full knowledge of these eigenvalues (or, in the least, corresponding bounds), Eckart's criterion cannot be directly applied. However, we can still use the inequality if less knowledge about the spectrum of H is available. If it is known that the energy estimate $\langle H\rangle_\sigma$ is closer to λ_0 than to λ_1 then, as an immediate consequence of Eckart's criterion, one finds that

$$\langle\psi_0|\sigma|\psi_0\rangle \geqslant \frac{1}{2}. \tag{294}$$

We remark that substituting (294) into the Gramian eigenvalue bounds from Theorem 12 yields the mixed state extension of the Weinstein bounds [145, 245] in the

non-degenerate case. If, in addition, a lower bound δ on the spectral gap is known such that $\lambda_1 - \lambda_0 \geqslant \delta$, then we have the bound derived in Ref. [151],

$$\langle\psi_0|\sigma|\psi_0\rangle \geqslant 1 - \frac{\Delta H_\sigma}{\delta}, \tag{295}$$

which is a nontrivial lower bound whenever the variance is small enough such that $\Delta H_\sigma \leqslant \delta$. With a similar technique, one obtains a further tightening of the bound:

$$\langle\psi_0|\sigma|\psi_0\rangle \geqslant \frac{1}{2}\left(1 + \sqrt{1 - \left(\frac{\Delta H_\sigma}{\delta/2}\right)^2}\right), \tag{296}$$

for variances with $\Delta H_\sigma \leqslant \delta/2$. We note that this bound has also been reported in [46, 244] and we provide an alternative proof in Section G.1. In practice, the bound δ on the spectral gap can also be estimated via classical methods, as for example truncated classical configuration interaction or density-matrix renormalization group techniques. In principle, also non-variational methods like truncated coupled-cluster (and the associated equation-of-motion or linear-response variants for the excited state energies) could be applied. In either case, the idea is to use these classical methods to compute the ground state and the first excited state energies to get an estimate of the spectral gap which can then be used for the fidelity estimation. The classical method which is the best to choose will generally depend on the system of interest and the available computational time. We refer the reader to [87] for a detailed treatment over some of those methods.

The above bounds hold for Hamiltonians H whose lowest eigenvalue is non-degenerate. In Section G.1 we consider the degenerate case and show that when the approximate state σ is pure, then there always exists a state $|\psi\rangle$ which is an element of the eigenspace associated with the lowest (possibly degenerate) eigenvalue, and for which the above fidelity lower bounds hold. However, if σ is a mixed state, this cannot be said in full generality, as is shown in the appendix with a counterexample. In summary, when the approximate state σ is allowed to be mixed, then the fidelity bounds are applicable only when the underlying Hamiltonian has a non-degenerate ground state. If, on the other hand, σ is pure, then the bounds also hold in the degenerate case. Finally, we remark that these fidelity bounds all require varying degrees of knowledge about the ground state and Hamiltonian in question. They thus can only partially address the topic of fidelity estimation in scenarios where such knowledge is not available.

At this point we would like to point out an interesting connection to Variational Quantum Time Evolution (VarQTE). In general, VarQTE is a technique to find the ground state of a Hamiltonian H [149, 158, 270] by projecting the time evolution of the initial state to the evolution of the ansatz parameters. VarQTE typically comes with an approximation error, stemming from a limited expressibility of the Ansatz state or from noise. In [282], this approximation error is quantified in terms of an upper bound on the Bures distance between the evolved state and the true ground state. Since there is a one-to-one correspondence between Bures distance and fidelity, these error bounds can be converted to a lower bound on the latter. This in turn can then be used to calculate the eigenvalue and expectation bounds presented in this work.

10.3 APPLICATIONS

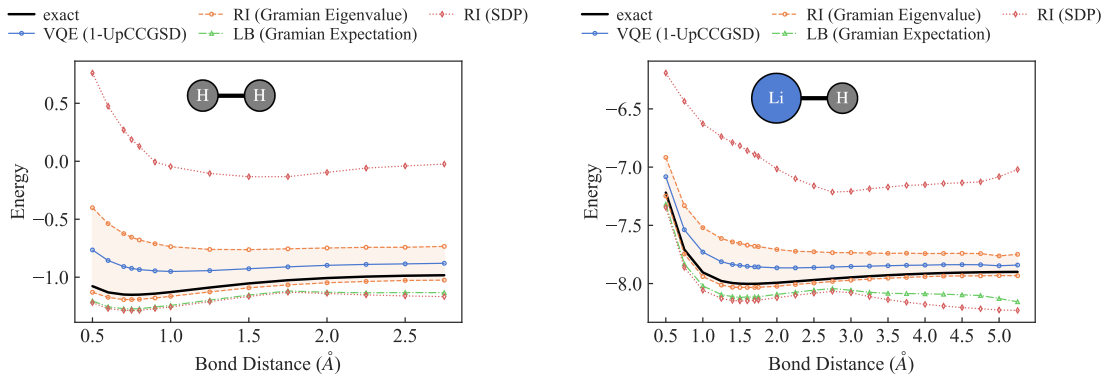
In this section, we put into practice the theoretical results presented in the previous sections and calculate the robustness intervals for ground state energies of electronic structure Hamiltonians when the approximation of the ground state is provided by VQE. We remark that, while VQE serves as an example application, our results are not limited to ground state energies but can be used in a more general context where the goal is to calculate error bounds for expectation values. Consider a Hamiltonian H with Pauli decomposition

$$H = \sum_{j=1}^{n} \omega_j P_j \tag{297}$$

and let σ be an approximation to the true ground state $|\psi_0\rangle$. Given $\epsilon \geqslant 0$ such that $\langle\psi_0|\sigma|\psi_0\rangle \geqslant 1-\epsilon$, and an estimate of the variance $\langle H^2\rangle_\sigma - \langle H\rangle_\sigma^2$, it is straightforward to evaluate the Gramian eigenvalue bounds from Theorem 12. In contrast, for the expectation value bounds derived via SDP and the Gramian method from Theorem 11 and Lemma 4), one needs to be more careful since the Hamiltonian H might violate the underlying assumptions. To evaluate the latter, we can account for this by adding a sufficiently large constant c such that $\tilde{H} := H + c\mathbb{1} \geqslant 0$ and calculate the bound for $\tilde{H}$, before reversing the translation in order to get the desired bound for H. Clearly, a valid choice for c is given by $-\lambda_0$ where λ_0 is the lowest eigenvalue of H. However, it is not always clear which constant c leads to the tightest lower bound. Similarly, to evaluate the SDP bounds, we need to apply Theorem 11 to operators which are bounded between $\pm\mathbb{1}$. If the full spectrum of H was known, one could normalize H using these eigenvalues. However, in the context of VQE, the spectrum is not a priori known as this is precisely the task that VQE is designed to solve, and we need a different approach for the expectation value bounds. The idea is to partition the terms in the Pauli decomposition from (297) into groups so that each term corresponding to a group can be normalized. To this end, we first partition H into groups of mutually qubit-wise commuting terms

$$H = \sum_{k=1}^{M} H_k, \quad H_k = \sum_j \omega_j^{(k)} P_j^{(k)}, \quad [P_i^{(k)}, P_j^{(k)}] = 0. \tag{298}$$

Given such a representation, the spectrum of each of the H_k can be calculated classically in order to scale $H_k \to \tilde{H}_k$ appropriately such that the assumptions for the bounds are satisfied. One can then compute the bounds for each of the terms in the summation and get the final bounds by aggregating the individual bounds. We further make use of the approach presented in [226, 269] where one applies a unitary transform U_k to each of the H_k terms so that single-qubit measurement protocols can be used. Specifically, instead of measuring H_k under the state σ, one measures $A_k = U_k H_k U_k^\dagger$ under the unitarily transformed $U_k \sigma U_k^\dagger$. One can then scale each A_k appropriately by classically computing its eigenvalues and apply the expectation value bounds (Theorem 11 and Lemma 4) to each term separately before aggregating. It is also worth noting the generality of Theorem 11. Although in the preceding demonstration, the matrix A is generally taken to be a Pauli observable for measuring the output of a quantum circuit, the condition $-\mathbb{1} \leqslant A \leqslant \mathbb{1}$ is satisfied much more generally (e. g., by Fermionic operators). The application of this theorem in settings without explicit Pauli decomposition would be a fruitful ground for future research.



(a) (b)

Figure 28: Comparison of the different lower bounds (LB) and robustness intervals (RI) presented in Section 10.2 for bond dissociation curves of $H_2(2, 4)$ and LiH(2, 4). The approximation states are provided by VQE with an UpCCGSD ansatz. Both the VQE optimization and the evaluation of the bounds were simulated with bit flip and depolarization noise with 1% error probability.

10.3.1 *Numerical simulations*

Here, we numerically investigate the different robustness bounds for the ground state energies for a set of electronic structure Hamiltonians, namely H_2, LiH and BeH_2 molecules where the qubit Hamiltonians are obtained within the basis-set-free approach of [118] using directly determined pair-natural orbitals on MP2 level [117].[1] All our experiments have been implemented in the TEQUILA [114] library using the qubit encodings from OPENFERMION [152], optimizers from SCIPY [228], MADNESS [80] as chemistry backend, QULACS [210] as the simulation backend for noiseless simulations and QISKIT [283] for simulations which include noise. We model noise as a combination of bitflip channels acting on single qubit gates with 1% error probability, and depolarizing noise acting on two qubit gates, also with an error probability of 1%.

For a given Hamiltonian H, we first approximate its ground state $|\psi_0\rangle$ via VQE. That is, for an Ansatz state σ_θ with parameters θ one minimizes the objective $\langle H\rangle_{\sigma_\theta}$ and obtains optimal parameters $\theta^* = \arg\min_\theta \langle H\rangle_{\rho_\theta}$. It follows from the Rayleigh-Ritz variational principle [178, 182] that the expectation $\langle H\rangle_{\sigma^*_\theta}$ is an upper bound to the true ground state energy λ_0. The such obtained ground state approximation σ_{θ^*} is then used to evaluate the bounds by computing the relevant statistics, i.e. expectation values and variances of observables under this state. We notice that the quality of this state in terms of a distance to the true ground state is not easily obtainable without having some prior knowledge over the system of interest (see also Section 10.2.4 in this regard). For this reason and in order to investigate and compare the bounds, here we assume knowledge of the true fidelity with the ground state. In practice, this is of course not realistic and, as discussed previously, one needs to approximate the true fidelity. Given the ground state approximation σ_{θ^*} and the fidelity $\mathcal{F}(\sigma_{\theta^*}, |\psi_0\rangle\langle\psi_0|)$, we then estimate the expectation values and variances under σ_{θ^*} in order to evaluate the bounds. In the noiseless scenario, these statistics can be calculated exactly, whereas in

1 The code for the simulations presented in this chapter is available at https://github.com/DS3Lab/robustness-intervals-quantum-measurements.

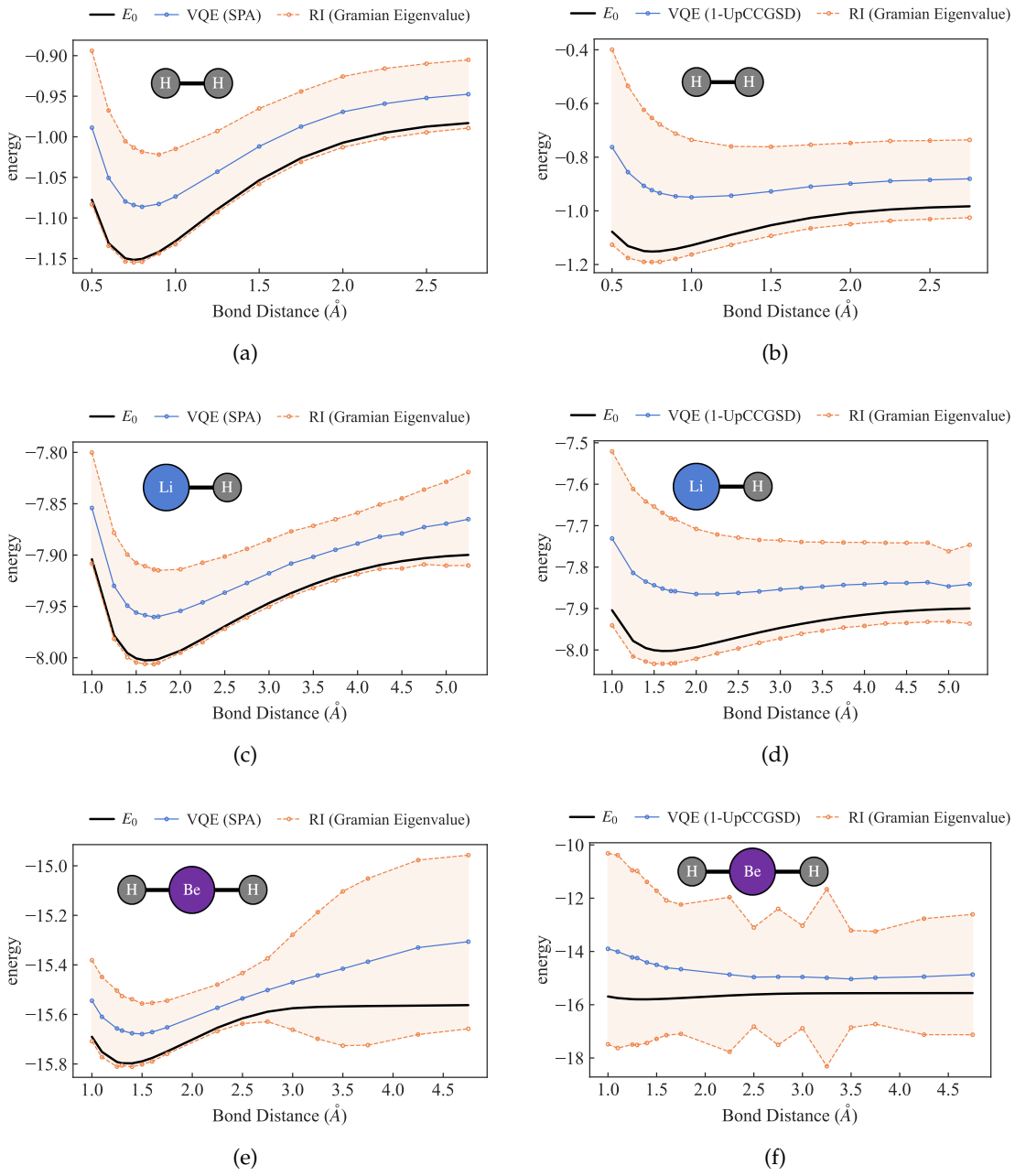
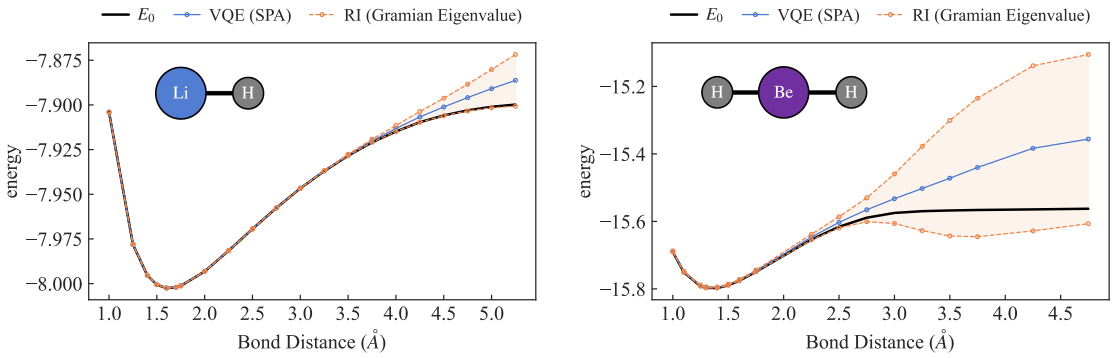


Figure 29: Bond dissociation curves and robustness intervals (RI) for eigenvalues based on the Gramian method (Theorem 12) for $H_2(2, 4)$, LiH(2, 4) and $BeH_2(4, 8)$. Both the VQE optimization and the evaluation of the bounds are done under a combination of bit flip and depolarization noise with 1% error probability.



(a) (b)

Figure 30: Bond dissociation curves, robustness interval (RI) for eigenvalues based on the Gramian method (Theorem 12) for LiH(2, 4)and BeH_2(4, 8). Here, an ideal scenario without noise is simulated and the approximation errors stem from the limited expressibility of the Ansatz state.

the noisy scenario they need to be estimated due to finite sampling errors. Thus, in the noisy case, we repeat the calculation of the bounds 20 times and report one-sided 99%-confidence intervals.

In Figure 28, we consider the noisy scenario and compare the different types of bounds for H_2(2, 4) and LiH(2, 4) with approximation states provided by VQE with an UpCCGSD Ansatz [130] and optimized fermionic gradients [115]. For both molecules, we notice that the Gramian eigenvalue bound is the tightest, while the expectation value bounds are less tight. However, this is not surprising, as the eigenvalue bound is more suited for this task, compared to the other bounds which hold more generally for expectation values. In Figure 29, we again consider the noisy scenario and compare the Gramian eigenvalue bounds for approximation states obtained via the SPA Ansatz [116] and via the UpCCGSD Ansatz for H_2(2, 4), LiH(2, 4) and BeH_2(4, 8). We first notice that the SPA Ansatz is generally less vulnerable to noise, which stems from the associated shallow circuits, compared to the UpCCGSD Ansatz. In particular, SPA and UpCCGSD have the same expressibility for H_2 but, since SPA uses more efficient compiling, its energy estimates and lower bounds are more accurate compared to UpCCGSD. In Section G.2 we show robustness intervals for the LiH(2, 4) molecule with the error rate increased to 10%. For the SPA ansatz, even with this error rate, the ground state fidelities vary between 0.51 and 0.65 while the UpCCGSD states have low ground state fidelities in the range 0.1. In other words, UpCCGSD fails to converge to states which are close to the true ground state. It is interesting to note that for lowest ground state fidelities, the expectation value bounds reduce to trivial bounds and the eigenvalue bound starts to diverge. In Figure 30 we consider the noiseless scenario with an SPA ansatz for LiH and BeH_2. In contrast to the noisy scenario, here the bounds based on the UpCCGSD Ansatz are tigther, compared to the ones based on the SPA Ansatz for large bond distances. This is due to the fact that SPA generally has more difficulties in approximating ground states for far stretched bond distances and hence results in lower ground state fidelities. Finally, it is worth remarking that these bounds are obtained under the assumption of having complete knowledge of the true ground state fidelity, an assumption which is idealistic and typically violated in practice.

10.3.2 *Implementation*

All our robustness intervals are implemented in the open source TEQUILA [114] library. In the following example, we run VQE for the H_2 Hamiltonian in a minimal representation (4 qubits), before computing the lower and upper bounds based on the optimized circuit, using the function `robustness_interval`:

```
import tequila as tq
from tequila.apps.robustness import robustness_interval

geom = 'H .0 .0 .0\nH .0 .0 .75'
mol = tq.Molecule(geom, n_pno=1)

H = mol.make_hamiltonian()
U = mol.make_upccgsd_ansatz()
E = tq.ExpectationValue(H=H, U=U)
result = tq.minimize(E)

lower_bound, energy, upper_bound, _ = robustness_interval(U, H, fidelity,
    variables=result.variables)
```

Used in this way, the function calculates the robustness interval for all three methods and returns the tightest bounds. Alternatively, one can specify the type of bound via the keywords `kind` and `method` where the former stands for which kind of interval is desired, that is expectation or eigenvalue, and the latter stands for the method used to obtain the bound (Gramian or SDP). For example, calculating a robustness interval for eigenvalues using the Gramian method, can be implemented as

```
robustness_interval(..., kind="eigenvalue", method="gramian").
```

In general, any type of expectation value can be used. Note that our implementation is agnostic with respect to the molecular representation, so that replacing `n_pno=1` with `basis_set="sto-3g"` will lead to a 4 qubit Hamiltonian in a traditional basis set.

10.4 CONLUSION

The current experimental stage of quantum computation offers the possibility to explore the physical and chemical properties of small systems and novel quantum algorithms are being developed to extract the most from this first generation of quantum devices. However, this potential for computational advantage, compared to classical methods, comes at a price of noisy and imperfect simulations stemming from low qubit counts and thus the lack of quantum error correcting qubits. The VQE is the canonical example of these NISQ algorithms that allow us to obtain an approximation of Hamiltonian eigenstates by exploiting the variational principle of quantum mechanics. Besides the broad applications and promising results of this approach [5, 151, 173], its performance guarantees should be studied and understood.

In this last of the five core chapters of this thesis, we have made first progress in this direction and have derived robustness intervals for quantum measurements of expectation values. For a target state ρ, these intervals are guaranteed to contain the true expectation value $\langle A \rangle_\rho$ of an observable A when we only have access to an approximation σ. Based on resource constraints, accuracy requirements, and depending on the

task, we have seen three different types of robustness intervals. Firstly, based on the formulation of robustness bounds as SDPs, we have derived upper and lower bounds to $\langle A \rangle_\rho$ which take into account only the first moment of the observable A and can thus be obtained by post-processing measurements of $\langle A \rangle_\sigma$ together with the fidelity with the target state $\mathcal{F}(\rho, \sigma)$. Secondly, we have revisited the Gramian method [243] to take into account higher statistical moments of A and extended this technique to mixed states, thereby enabling their applicability in noisy scenarios which are prevalent in the NISQ era. This has led to a further lower bound to expectation values and, additionally, to lower and upper bounds on eigenvalues of observables. We have also implemented our bounds in the open source TEQUILA [114] library. Furthermore, we have validated our results with numerical simulations of noisy and noiseless scenarios with VQE as an example application to calculate robustness intervals for ground state energies of electronic structure Hamiltonians of H_2, LiH and BeH_2. For the molecules considered in these experiments, we have observed that the robustness intervals provide accurate estimates of the errors incurred by noise, in particular when the ground state approximation is close enough to the true ground state in terms of fidelity.

The main requirement of the bounds obtained is the knowledge of the fidelity between the target state and its approximation. Although such a quantity is not always experimentally accessible and hence poses a challenge in the practical applicability of these bounds, there exist algorithms, such as within the variational quantum imaginary time evolution [282] framework, which allow for a quantification of the required approximation error in terms of distances between the target and the approximate state. Nonetheless, our aim is to provide a formal framework to study the robustness of broadly used approaches as are the Variational Quantum Algorithms. There are still many questions around the applicability of these quantum algorithms and its robustness against noise. Within this work, we seek to unravel the uncertainties around these state-of-the-art quantum algorithms with the goal of improving its performance and applicability.

11 CONCLUSION

The topic of safety and reliability of machine learning has gained substantial interest over the recent years and its importance is only increasing as ML systems become more integrated into our lives. Among the plethora of sub-fields of safe and reliable ML, this thesis was concerned with probabilistic robustness guarantees and took a holistic view on ML systems, covering different stages of an ML pipeline, and considering both classical and quantum computing frameworks. We have addressed several open questions related to robustness guarantees for data poisoning attacks, semantic transformations, and presented bounds to certify the out-of-domain generalization. We have further developed robustness guarantees for quantum machine learning and presented bounds on quantum expectation values with applications to NISQ algorithms.

11.1 SUMMARY

The core subject around which this thesis has revolved concerns the question of how we can provide robustness guarantees for different stages of an ML system. We divided this question up into five smaller parts, which we have treated in the different chapters of this thesis and have made progress towards building certifiably robust ML systems. Following the natural flow of an ML pipeline, we first set out to derive provable robustness against backdoor attacks in Chapter 6, and treated the first question:

Question 1: How can we develop certifiably robust ML models against backdoor attacks?

To address this question, we have built on the Neyman-Pearson approach of certifying robustness via randomised smoothing [39, 264], and developed a robust training algorithm that provides a robustness certificate against backdoor attacks. Specifically, given that our model was trained on a poisoned training set, the certificate guarantees that the prediction would have been the same, if the model was trained on the clean training set (i. e., with the backdoor pattern removed). This in turn implies that the adversary does not gain any advantage by knowing the backdoor pattern. As we have seen, the main limitation of the approach is the computational cost, arising from the method being based on an ensemble of models. Nevertheless, we believe that this result is an important first step towards provable robustness against backdoor attacks.

In the second question, which was the focus of Chapter 7, we were concerned with certifying robustness against a specific type of input perturbations, namely those arising from semantic transformations:

Question 2.1: How can we guarantee the robustness of ML models against input corruptions arising from semantic transformations that incur large ℓ_p-norm perturbations?

Similar to the approach we used in the treatment of the previous question, we have built on the Neyman Pearson approach to probabilistic robustness guarantees. To that end, we have developed a framework, TSS, that performs smoothing of a classifier over parameters of semantic transformations (e. g., rotation angle). Rather than a guarantee

on the absolute error between inputs, here we certify the transformation parameters since such transformations incur large ℓ_p norm distances which cannot be handled by classical additive smoothing approaches. Using this framework, we were able to certify a large number of different semantic transformations and set new state of the art on certified robustness for this type of perturbations.

In the third question, treated in Chapter 8, we moved away from instance-level robustness and focused our attention on how population-level metrics change due to changes in the data distributions:

Question 2.2 How can we certify the out-of-domain generalization abilities of ML models while only allowing blackbox access to these models?

To address this question, we have adapted a technique to bound quantum expectation values which is based on the non-negativity of Gram matrices and which was pioneered by Weinhold [243]. The such derived bounds take into account both variance and the expectation of the source distribution and, since they only require blackbox access to model predictions, can be efficiently estimated from data, in contrast to existing bounds. In our experimental validation, we have shown that this method can certify the out-of-domain generalization of diverse models which are as large as a full-fledged EfficientNet-B7 and BERT.

In Part iv of this thesis, we considered ML systems which make use of quantum computing components, and focused our attention on deriving robustness guarantees for quantum classifiers in Chapter 9, and more general quantum algorithms which rely on the readout of expectation values (Chapter 10). We started this investigation with the following question:

Question 3.1: How can we enable tight robustness guarantees for quantum classification models, taking into account the unique nature of QML algorithms?

Stemming from the observation that quantum circuits, and thus quantum classifiers, are naturally probabilistic, we presented a quantum analogue of the robustness guarantees based on the Neyman-Pearson Lemma. In addition, due to optimality of the Neyman-Pearson tests (i. e., Helstrom operators), this result was shown to be tight. Finally, as we have seen, due to the probabilistic nature of quantum classifiers, this result has revealed an intrinsic robustness guarantee which, in contrast to classical models, does not require smoothing over inputs, although it can be enhanced by it.

While the previous question addresses quantum algorithms whose output relies on the most likely measurement outcome, the final step in this thesis was to take a broader view and derive robustness intervals for quantum expectation values, when the prepared state is only an approximation to the ideal state. Inspired by the early days of computing [230, 251], we asked the following question:

Question 3.2: How can we characterize accurate error bounds on the output of quantum algorithms arising from imperfect representations of an ideal quantum state?

To answer this question, we have adapted two techniques to derive bounds on the worst-case error in quantum expectation values arising from approximations of an ideal quantum state. The first type of bounds relies on quantum hypothesis testing and is essentially an application of the Neyman-Pearson robustness guarantees presented in Chapter 9. In the second approach, we have extended the Gramian technique, whose

classical counterpart we have used to certify out-of-domain generalization in Chapter 8, to the mixed state formalism. This extension ultimately enables the applicability of this technique in the NISQ er of quantum computing. Finally, we have validated these bounds in the context of the Variational Quantum Eigensolver. While these bounds allow for an accurate characterization of the error and do not require knowledge of specific noise or state preparation models, the main limitation is that they do require knowledge, or, in the least, an approximation of the quantum state fidelity between approximate and ideal quantum state.

11.2 RESEARCH OUTLOOK

In this thesis, we have provided answers to five central open questions related to provably robust ML systems and, throughout every chapter, made extensive use of probabilistic approaches to robustness certification. While these results equip us with a better understanding and novel robustness certification techniques, we have only scratched the surface of designing provably robust ML systems. Many other types of vulnerabilities remain open, and optimizing these certificates in light of computational efficiency and tightness will be important cornerstones for their adoption in practice. In addition, with the current pace of integration of ML systems into everyday applications, we believe that new challenges related to reliable ML will emerge and require further careful analysis.

Many of the results presented in the five main chapters of this thesis share common limitations which, to a large extent, arise from current limitations of probabilistic robustness guarantees, and provide fruitful grounds for future research. Specifically, randomized smoothing based certification approaches usually suffer from the computational overhead caused by the requirement of estimating expectation values via Monte-Carlo methods. This limitation is perhaps most pronounced in our proposed approach to certifying robustness against backdoor attacks. Indeed the formulation of smoothing over the *training* data leads to the requirement of training an ensemble of models in order to estimate the associated expectation value, which is of course computationally expensive. Nevertheless, as we have shown in the context of KNN-classifiers, there exist problem instances where this limitation can be alleviated, giving hope that in the future, more algorithms which improve the efficiency of probabilistic guarantees will be discovered. We believe that such improvements in computational efficiency of probabilistic robustness guarantees will become even more important as models grow in size such as in the context of large language models where inference cost becomes a bottleneck for many applications. In addition, since we have developed *probabilistic* robustness guarantees, even if they hold with high probability (e. g., 99.9%), this still leaves the possibility of errors when a large number of predictions need to be certified. Therefore, especially in highly security-critical scenarios, developing deterministic guarantees for ML systems is paramount.

From the operational point of view, a further central limitation of the results presented in this thesis, is posed by the challenge of estimating statistical distances and distances between quantum states. In the classical case, this limitation is most pronounced in the context of the results on certifying out-of-domain generalization pre-

sented in Chapter 8. While the robustness guarantees presented scale to large models and datasets, the operational significance is currently hindered by a lack of techniques to accurately estimate the distribution drift in terms of the Hellinger distance, especially for high dimensional data. Indeed, such a result would enable to continuously estimate a worst-case drop in model performance and provide a signal when models need to be adapted to new data distributions. In the context of our robustness guarantees for quantum machine learning, a similar challenge exists. While our bounds are intrinsic guarantees of quantum classifiers and are independent of the underlying model architecture, they still depend on estimates of a notion of similarity between quantum states. Estimating these measures, like the fidelity to an ideal, but unavailable quantum state, without having access to it, is a challenging task in practice and developing such methods is needed.

Finally, while in this thesis we have provided results for *individual* components of an ML system, the *interplay* between guarantees for different components and vulnerabilities of such intelligent systems needs to be thoroughly understood in order to build ML systems with provable *end-to-end* robustness. Promising endeavours in this direction have already been made [268], giving fruitful grounds for further research in this direction.

Part V

APPENDIX

A

PROOFS FOR BOUNDS ON CLASSICAL EXPECTATION VALUES

In this chapter, we provide proofs for the results presented in Chapter 4.

A.1 LIKELIHOOD RATIO TESTS

Recall the definition of a likelihood ratio test for testing the null hypothesis $X \sim P$ against the alternative $X \sim Q$, given in Section 4.1,

$$\phi_{NP}(x) = \begin{cases} 1 & \Lambda(x) > t \\ q & \Lambda(x) = t \\ 0 & \Lambda(x) < t \end{cases}, \quad \text{with} \quad \Lambda(x) = \frac{f_Q(x)}{f_P(x)} \tag{299}$$

where $f_P = \frac{dP}{d\mu}$ and $f_Q = \frac{dQ}{d\mu}$ are the probability density functions with respect to a reference measure μ on $\mathcal{X}$. For a fixed significance level $\alpha_0 \in [0, 1]$, the value of t is chosen according to

$$t = \inf\{s \geqslant 0 \,|\, \mathbb{P}_{X\sim P}[\Lambda(X) \leqslant s] \geqslant 1 - \alpha_0\} \tag{300}$$

and q is then set to

$$q = \begin{cases} 0 & \text{if } \mathbb{P}_{X\sim P}[\Lambda(X) = t] = 0, \\ \frac{\alpha_0 - \mathbb{P}_{X\sim P}[\Lambda(X) > t]}{\mathbb{P}_{X\sim P}[\Lambda(X) = t]} & \text{otherwise.} \end{cases} \tag{301}$$

This choice ensures that

$$\alpha(\phi_{NP}; P) = q\mathbb{P}_{X\sim P}[\Lambda(X) = t] + \mathbb{P}_{X\sim P}[\Lambda(X) > t] = \alpha_0. \tag{302}$$

The following Lemma ensures that $q \in [0, 1]$ and that the resulting likelihood ratio test is well defined.

Lemma 12. *Let P and Q be two probability measures with densities f_P and f_Q with respect to a measure μ and denote by Λ the likelihood ratio $\Lambda(x) = f_Q(x)/f_P(x)$. For $p \in [0, 1]$ let $t_p := \inf\{t \geqslant 0 \colon \mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t] \geqslant p\}$. Then it holds that*

$$\mathbb{P}_{X\sim P}[\Lambda(X) < t_p] \leqslant p \leqslant \mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t_p]. \tag{303}$$

Proof. We first show the RHS of inequality (303). This follows directly from the definition of t_p if we show that the function $t \mapsto \mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t]$ is right-continuous. Let $t \geqslant 0$ and let $\{t_n\}_n$ be a sequence in $\mathbb{R}_{\geqslant 0}$ such that $t_n \downarrow t$. Define the sets $A_n := \{x \colon \Lambda(x) \leqslant t_n\}$ and note that $\mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t_n] = \mathbb{P}_{X\sim P}[X \in A_n]$. Clearly, if $x \in \{x \colon \Lambda(x) \leqslant t\}$ then $\forall n \colon \Lambda(x) \leqslant t \leqslant t_n$ and thus $x \in \cap_n A_n$. If on the other hand $x \in \cap_n A_n$ then $\forall n \colon \Lambda(x) \leqslant t_n \to t$ as $n \to \infty$. Hence, we have that $\cap_n A_n = \{x \colon \Lambda(x) \leqslant t\}$ and thus $\lim_{n\to\infty} \mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t_n] = \mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t]$ since $\lim_{n\to\infty} \mathbb{P}_{X\sim P}[X \in A_n] = \mathbb{P}_{X\sim P}[X \in \cap_n A_n]$ for $A_{n+1} \subseteq A_n$. We conclude that $t \mapsto \mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t]$ is right-continuous and

in particular $\mathbb{P}_{X\sim P}[\Lambda(X) \leqslant t_p] \geqslant p$. We now show the LHS of inequality (303). For that purpose, consider the set $B_n := \{x\colon \Lambda(x) < t_p - \frac{1}{n}\}$ and let $B := \{x\colon \Lambda(x) < t_p\}$. Clearly, if $x \in \cup_n B_n$, then $\exists n$ such that $\Lambda(x) < t_p - \frac{1}{n} < t_p$ and hence $x \in B$. If on the other hand $x \in B$, then we can choose n large enough such that $\Lambda(x) < t_p - \frac{1}{n}$ and thus $x \in \cup_n B_n$. It follows that $B = \cup_n B_n$. Furthermore, by the definition of t_p and since for any $n \in \mathbb{N}$ we have that $\mathbb{P}_{X\sim P}[X \in B_n] = \mathbb{P}_{X\sim P}\left[\Lambda(X) < t_p - \frac{1}{n}\right] < p$ it follows that $\mathbb{P}_{X\sim P}[\Lambda(X) < t_p] = \lim_{n\to\infty} \mathbb{P}_{X\sim P}[X_0 \in B_n] \leqslant p$ since $B_n \subseteq B_{n+1}$. This concludes the proof. □

The next Lemma is essentially the Neyman-Pearson Lemma [162] and establishes optimality of the likelihood ratio test.

Lemma 13. *Let P and Q be two probability measures with densities f_P and f_Q with respect to a measure μ and denote by Λ the likelihood ratio $\Lambda(x) = f_Q(x)/f_P(x)$ and denote by ϕ_{NP} be the likelihood ratio test defined in (299). Then, for any deterministic function $\phi\colon \mathcal{X} \to [0,1]$ the following holds:*

i) $\alpha(\phi; P) \geqslant 1 - \alpha(\phi_{NP;P}) \Rightarrow 1 - \beta(\phi; Q) \geqslant \beta(\phi_{NP}; Q)$

ii) $\alpha(\phi; P) \leqslant \alpha(\phi_{NP}; P) \Rightarrow \beta(\phi; Q) \geqslant \beta(\phi_{NP}; Q)$

Proof. We first show (i). We have

$$\begin{aligned} 1 - \beta(\phi_{NP}; Q) - \beta(\phi; Q) &= \int_{\Lambda>t} \phi \, dQ + \int_{\Lambda\leqslant t} (\phi - 1)\, dQ + q \int_{\Lambda=t} dQ \\ &= \int_{\Lambda>t} \phi\Lambda \, dP + \int_{\Lambda\leqslant t} \underbrace{(\phi - 1)}_{\leqslant 0} \Lambda \, dP + q \int_{\Lambda=t} \Lambda \, dP \\ &\geqslant t \cdot \left[\int_{\Lambda>t} \phi \, dP + \int_{\Lambda\leqslant t} (\phi - 1)\, dP + q \int_{\Lambda=t} dP \right] \\ &= t \cdot [\alpha(\phi; P) - (1 - \alpha(\phi_{NP}; P))] \geqslant 0 \end{aligned} \tag{304}$$

with the last inequality following from the assumption and $t \geqslant 0$. Thus, (i) follows; (ii) can be proved analogously. □

B

PROOFS FOR BOUNDS ON QUANTUM EXPECTATION VALUES

In this chapter, we provide proofs for the results presented in Chapter 5.

B.1 PROOF OF LEMMA 3

Lemma 3 constitutes one of the core theoretical results required to derive the robustness bounds for quantum classifiers (Chapter 9) and for the robustness intervals presented in Chapter 10. We start with a quick review of the central definitions related to QHT and an outline of the proof. We then proceed to constructing Helstrom operators and proving their optimality. Finally, we conclude with the proof of Lemma 3.

PRELIMINARIES Throughout this section, the null hypothesis is described by the density operator $\sigma \in \mathcal{S}(\mathcal{H})$, while the alternative hypothesis is denoted by $\rho \in \mathcal{S}(\mathcal{H})$. A quantum hypothesis test is a positive semi-definite operator $0 \leqslant M \leqslant \mathbb{1}_d$ and the type-I and type-II error probabilities associated with M are denoted by α and β and are defined by

$$\begin{aligned} \alpha(M;\sigma) &:= \operatorname{Tr}[\sigma M] && \text{(type-I error)} \\ \beta(M;\rho) &:= \operatorname{Tr}[\rho(\mathbb{1}-M)] && \text{(type-II error)} \end{aligned}$$

Throughout this section we will omit the explicit dependence on σ and ρ whenever it is clear from context. For two Hermitian operators A and B, we write $A \geqslant B$ ($A \leqslant B$) if $A - B$ is positive (negative) semi-definite and $A > B$ ($A < B$) if $A - B$ is positive (negative) definite. For a Hermitian operator A with spectral decomposition $A = \sum_i \lambda_i P_i$ with eigenvalues $\{\lambda_i\}_i$ and orthogonal projections onto the associated eigenspaces $\{P_i\}_i$, we write

$$\{A > 0\} := \sum_{i\colon \lambda_i > 0} P_i, \quad \{A < 0\} := \sum_{i\colon \lambda_i < 0} P_i \tag{305}$$

for the projections onto the eigenspaces associated with positive and negative eigenvalues respectively. Finally, for $t \geqslant 0$ define the operators

$$\begin{aligned} P_{t,+} &:= \{\rho - t\sigma > 0\}, \\ P_{t,-} &:= \{\rho - t\sigma < 0\}, \\ P_{t,0} &:= \mathbb{1} - P_{t,+} - P_{t,-}. \end{aligned} \tag{306}$$

Helstrom operators are the quantum counterpart of likelihood ratio tests from classical hypothesis testing and are defined as

$$M_t := P_{t,+} + X_t, \quad 0 \leqslant X_t \leqslant P_{t,0}. \tag{307}$$

Finally, in the Neyman-Pearson approach to QHT, an optimal hypothesis test is an operator that solves the SDP problem

$$\begin{aligned} \beta^*(\alpha_0;\sigma,\rho) := \text{minimize} \quad & \beta(M;\rho) \\ \text{s.t.} \quad & \alpha(M;\sigma) \leqslant \alpha_0, \\ & 0 \leqslant M \leqslant \mathbb{1}_d. \end{aligned} \tag{308}$$

for some predefined threshold $\alpha_0 \in [0, 1]$.

PROOF OUTLINE The first step in the proof is to make an explicit construction of Helstrom operators (307), and show that these are optimal hypothesis tests for the SDP (308). In a subsequent we will use this construction to derive a lower bound to (308), which in the case of pure states becomes an equality.

B.1.1 *Construction of Helstrom Operators and Optimality*

The first step is to show that if a sequence of bounded Hermitian operators A_n converges in operator norm to a bounded Hermitian operator A from above (below), then the projections $\{A_n < 0\}$ and $\{A_n > 0\}$ converge to $\{A < 0\}$ and $\{A > 0\}$, respectively, in operator norm. This subsequently allows us to show that the function $t \mapsto \alpha(P_{t,+})$ is non-increasing and continuous from the right, and that $t \mapsto \alpha(P_{t,+} + P_{t,0})$ is non-increasing and continuous from the left. As a consequence, for $\alpha_0 \in [0, 1]$, the quantity

$$\tau(\alpha_0) := \inf\{t \geqslant 0 \colon\ \alpha(P_{t,+}) \leqslant \alpha_0\} \tag{309}$$

is well defined. This implies the chain of inequalities

$$\alpha\left(P_{\tau(\alpha_0),+}\right) \leqslant \alpha_0 \leqslant \alpha\left(P_{\tau(\alpha_0),+} + P_{\tau(\alpha_0),0}\right). \tag{310}$$

Based on these, we can construct a Helstrom operator $M_{\tau(\alpha_0)}$ according to

$$M_{\tau(\alpha_0)} := P_{\tau(\alpha_0),+} + q_0 P_{\tau(\alpha_0),0}, \tag{311}$$

where

$$q_0 := \begin{cases} \frac{\alpha_0 - \alpha(P_{\tau(\alpha_0),+})}{\alpha(P_{\tau(\alpha_0),0})} & \text{if } \alpha\left(P_{\tau(\alpha_0),0}\right) \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{312}$$

which attains the preassigned type-I error probability α_0. Finally, we will show that these Helstrom operators are optimal for the SDP problem (308), so that

$$\beta^*(\alpha_0; \sigma, \rho) = \beta(M_{\tau(\alpha_0)}; \rho). \tag{313}$$

Lemma 14. *Denote by $\mathcal{B}(\mathcal{H})$ the space of bounded linear operators acting on the finite dimensional Hilbert space $\mathcal{H}$, $d := \dim(\mathcal{H}) < \infty$. Let $A \in \mathcal{B}(\mathcal{H})$ and $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{B}(\mathcal{H})$ be Hermitian operators and suppose that $\|A_n - A\|_{op} \xrightarrow{n\to\infty} 0$. Then, it holds that*

$$\text{(i) } A - A_n \leqslant 0 \Rightarrow \|\{A_n < 0\} - \{A < 0\}\|_{op} \xrightarrow{n\to\infty} 0, \tag{314}$$

$$\text{(ii) } A - A_n \geqslant 0 \Rightarrow \|\{A_n > 0\} - \{A > 0\}\|_{op} \xrightarrow{n\to\infty} 0. \tag{315}$$

Proof. We first show that convergence in operator norm implies that the eigenvalues of A_n converge towards the eigenvalues A. For a linear operator M let $\lambda_k(M)$ denote its k-th largest eigenvalue, $\lambda_1(M) \geqslant \ldots \geqslant \lambda_q(M)$, where $q \leqslant d$ is the number of distinct eigenvalues of M. By the minimax principle (e.g. [13], chapter 3), we can compute λ_k for any Hermitian operator M according to

$$\lambda_k(M) = \max_{\substack{V\subseteq\mathcal{H} \\ \dim(V)=k}} \min_{\substack{\psi\in V \\ \|\psi\|=1}} \langle\psi|M|\psi\rangle. \tag{316}$$

Now let $\varepsilon > 0$ and let $n \in \mathbb{N}$ large enough such that $\|A_n - A\|_{op} < \varepsilon$. Let $|\psi\rangle \in \mathcal{H}$ be a normalized state and note that by the Cauchy-Schwartz inequality we have

$$|\langle\psi|(A_n - A)|\psi\rangle| \leqslant \|(A_n - A)\psi\| \, \|\psi\| \leqslant \|A_n - A\|_{op} \, \|\psi\|^2 < \varepsilon \tag{317}$$

and thus

$$\langle\psi|A|\psi\rangle - \varepsilon < \langle\psi|A_n|\psi\rangle < \langle\psi|A|\psi\rangle + \varepsilon. \tag{318}$$

Hence, for any fixed $k \geqslant 1$ and any subspace $V \subset \mathcal{H}$ with $\dim(V) = k$, we have

$$\min_{\substack{\psi\in V \\ \|\psi\|=1}} \langle\psi|A|\psi\rangle - \varepsilon < \min_{\substack{\psi\in V \\ \|\psi\|=1}} \langle\psi|A_n|\psi\rangle < \min_{\substack{\psi\in V \\ \|\psi\|=1}} \langle\psi|A|\psi\rangle + \varepsilon \tag{319}$$

and thus, from (316), we see that

$$\lambda_k(A) - \varepsilon < \lambda_k(A_n) < \lambda_k(A) - \varepsilon \Rightarrow |\lambda_k(A) - \lambda_k(A_n)| < \varepsilon \tag{320}$$

and hence

$$\lambda_k(A_n) \xrightarrow{n\to\infty} \lambda_k(A), \qquad k = 1, \ldots, q. \tag{321}$$

Alternatively, this can be seen from Weyl's Perturbation Theorem (e.g. [13], ch. 3): namely, since A and A_n are Hermitian, it follows immediately from

$$|\lambda_k(A) - \lambda_k(A_n)| \leqslant \max_k |\lambda_k(A) - \lambda_k(A_n)| \leqslant \|A - A_n\|_{op} \tag{322}$$

that eigenvalues converge provided that $\|A_n - A\|_{op} \to 0$. We will now make use of function theory and the resolvent formalism to show the convergence of the positive and negative eigenprojections. Let $M \in \mathcal{B}(\mathcal{H})$ be Hermitian, let $\sigma(M)$ denote the spectrum of M and, for $\lambda \in \mathbb{C} \setminus \sigma(M)$, let

$$R_\lambda(M) := (M - \lambda\mathbb{1})^{-1} = -\sum_{k=0}^{\infty} \lambda^{-(k+1)} M^k \tag{323}$$

be the resolvent of the operator M. The sum is the Neumann series and converges for $\lambda \in \mathbb{C} \setminus \sigma(M)$. Since M is Hermitian, we can write its spectral decomposition in terms of contour integrals over the resolvent

$$M = \sum_{k=1}^{q} \lambda_k(M) P_k \quad \text{with} \quad P_k = \frac{1}{2\pi i} \oint_{(\gamma_k, -)} R_\lambda(M) \, d\lambda, \quad \text{and} \quad \sum_{k=1}^{q} P_k = \mathbb{1} \tag{324}$$

where P_k is the orthogonal projection onto the k-th eigenspace and the integration is to be understood element-wise. The symbol $(\gamma_k, -)$ indicates that the contour encircles $\lambda_k(M)$ once negatively, but does not encircle any other eigenvalue of M. We refer the reader to [181] for a detailed derivation.

We now show part (i) of the Lemma. For ease of notation, let λ_k denote the k-th eigenvalue of A and $\lambda_{k,n}$ the k-th eigenvalue of A_n. Since A_n and A are Hermitian

operators, we can write the eigenprojections $\{A_n < 0\}$ and $\{A < 0\}$ in terms of the resolvent as

$$\begin{aligned}\{A < 0\} &= \frac{1}{2\pi i} \sum_{k\colon \lambda_k < 0} \oint_{(\gamma_k, -)} R_\lambda(A)\, d\lambda, \\ \{A_n < 0\} &= \frac{1}{2\pi i} \sum_{k\colon \lambda_{k,n} < 0} \oint_{(\gamma_{k,n}, -)} R_\lambda(A_n)\, d\lambda\end{aligned} \tag{325}$$

where the symbols $(\gamma_k, -)$ and $(\gamma_{k,n}, -)$ indicate that the contours encircle only λ_k and $\lambda_{k,n}$ once negatively and no other eigenvalues of A and A_n respectively. Since by assumption $A_n \geqslant A$ and A_n, A are Hermitian, it follows from Weyl's Monotonicity Theorem that $\lambda_{k,n} \geqslant \lambda_k$. Let λ_K be the largest negative eigenvalue of A, that is $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \lambda_{K-1} \geqslant 0 > \lambda_K \geqslant \ldots \geqslant \lambda_q$. Note that if A is positive semidefinite, then so is A_n and the statement follows trivially from $\{A_n < 0\} = \{A < 0\} = 0$. Thus, without loss of generality, we can assume that at least one eigenvalue of A is negative. Since $\lambda_{k,n} \geqslant \lambda_k$, and in particular $\lambda_{K-1,n} \geqslant \lambda_{K-1} \geqslant 0$, there exists $N_0 \in \mathbb{N}$ large enough such that $\lambda_{K-1,n} \geqslant 0 > \lambda_{K,n}$ for all $n \geqslant N_0$. Let r_0 be the smallest distance between two eigenvalues of A

$$r_0 := \min_k |\lambda_k - \lambda_{k+1}| \tag{326}$$

and let $0 < \varepsilon < \frac{r_0}{2}$. Choose $N_1 \geqslant N_0$ large enough such that

$$\max_{k \geqslant K} |\lambda_{k,n} - \lambda_k| < \varepsilon/2. \tag{327}$$

Let $0 < \delta < \frac{r_0}{2} - \varepsilon$ and for $k \geqslant K$ let $B^k_{\delta+\varepsilon} := B_{\delta+\varepsilon}(\lambda_k)$ be the open ball of radius $\delta + \varepsilon$ centered at λ_k. Note that $\partial B^k_{\delta+\varepsilon}$ encircles both $\lambda_{k,n}$ and λ_k. Then, for $k \geqslant K$ and $n \geqslant N_1$, the mappings

$$\lambda \mapsto R_\lambda(A), \quad \lambda \in B^k_{\delta+\varepsilon} \setminus \{\lambda_k\}, \tag{328}$$

$$\lambda \mapsto R_\lambda(A_n), \quad \lambda \in B^k_{\delta+\varepsilon} \setminus \{\lambda_{k,n}\} \tag{329}$$

are holomorphic functions of λ and each has an isolated (simple) singularity at λ_k and $\lambda_{k,n}$ respectively. Let $\gamma_{k,n}$ be a contour around $\lambda_{k,n}$ encircling no other eigenvalue of A_n. Note that the contours $\gamma_{k,n}$ and $\partial B^k_{\delta+\varepsilon}$ are homotopic (in $B^k_{\delta+\varepsilon} \setminus \{\lambda_{k,n}\}$). Thus, for all $k \geqslant K$ and $n \geqslant N_1$, Cauchy's integral Theorem yields

$$\oint_{\gamma_{k,n}} R_\lambda(A_n)\, d\lambda = \oint_{\partial B^k_{\delta+\varepsilon}} R_\lambda(A_n)\, d\lambda. \tag{330}$$

With this, we see that for $n \geqslant N_1$

$$\{A_n < 0\} - \{A < 0\} = \frac{1}{2\pi i} \sum_{k=K}^{q} \left(\oint_{\partial B^k_{\delta+\varepsilon}} R_\lambda(A)\, d\lambda - \oint_{\gamma_{k,n}} R_\lambda(A_n)\, d\lambda \right) \tag{331}$$

$$= \frac{1}{2\pi i} \sum_{k=K}^{q} \left(\oint_{\partial B^k_{\delta+\varepsilon}} (R_\lambda(A) - R_\lambda(A_n))\, d\lambda \right) \tag{332}$$

and thus, by the triangle inequality

$$\|\{A_n < 0\} - \{A < 0\}\|_{op} \leqslant \frac{1}{2\pi} \sum_{k=K}^{q} \left\| \oint_{\partial B^k_{\delta+\varepsilon}} (R_\lambda(A) - R_\lambda(A_n))\, d\lambda \right\|_{op} \tag{333}$$

$$\leqslant \frac{1}{2\pi} \sum_{k=K}^{q} \sup_{\lambda \in \partial B^k_{\delta+\varepsilon}} \|R_\lambda(A) - R_\lambda(A_n)\|_{op} \cdot 2\pi \cdot (\delta + \varepsilon) \tag{334}$$

$$\leqslant (q - (K-1)) \cdot (\delta + \varepsilon) \cdot \max_{k \geqslant K} \left(\sup_{\lambda \in \partial B^k_{\delta+\varepsilon}} \|R_\lambda(A) - R_\lambda(A_n)\|_{op} \right). \tag{335}$$

Furthermore, for any k and $\lambda \in \partial B^k_{\delta+\varepsilon}$, the second resolvent identity yields

$$\begin{aligned} \|R_\lambda(A) - R_\lambda(A_n)\|_{op} &= \|R_\lambda(A)(A - A_n)R_\lambda(A)\|_{op} \\ &\leqslant \|A - A_n\|_{op} \cdot \|R_\lambda(A)\|_{op} \|R_\lambda(A_n)\|_{op}. \end{aligned} \tag{336}$$

We now show that the supremum over $\lambda \in \partial B^k_{\delta+\varepsilon}$ in the right hand side of (336) is bounded. Since both A and A_n are Hermitian, it follows that their resolvent is normal and bounded for $\lambda \in \mathbb{C} \setminus \sigma(A_n)$ and $\lambda \in \mathbb{C} \setminus \sigma(A)$ respectively. The operator norm is thus given by the spectral radius,

$$\|R_\lambda(A)\|_{op} = \max_k |\lambda_k(R_\lambda(A))| \tag{337}$$

and

$$\|R_\lambda(A_n)\|_{op} = \max_k |\lambda_k(R_\lambda(A_n))|. \tag{338}$$

Note that the eigenvalues of $R_\lambda(A)$ are given by $(\lambda_k(A) - \lambda)^{-1}$. To see this, let $\lambda \in \mathbb{C} \setminus \sigma(A)$ and consider

$$\det(R_\lambda(A) - \mu\mathbb{1}) = \det\left((A - \lambda\mathbb{1})^{-1}(\mathbb{1} - (A - \lambda\mathbb{1})\mu\mathbb{1})\right) \tag{339}$$

$$\propto \det(\mathbb{1} - (A - \lambda\mathbb{1})\mu\mathbb{1}) \tag{340}$$

$$= (-\mu)^m \det\left(A - (\mu^{-1} + \lambda)\mathbb{1}\right). \tag{341}$$

Since $\det(R_\lambda(A)) \neq 0$ it follows that $\mu = 0$ can not be an eigenvalue. Thus, eigenvalues of $R_\lambda(A)$ satisfy

$$\frac{1}{\mu} + \lambda = \lambda_k(A) \Rightarrow \mu = \frac{1}{\lambda_k(A) - \lambda}. \tag{342}$$

The same reasoning yields an expression for eigenvalues of $R_\lambda(A_n)$. Thus

$$\|R_\lambda(A)\|_{op} = \max_k \frac{1}{|\lambda_k(A) - \lambda|} \tag{343}$$

and

$$\|R_\lambda(A_n)\|_{op} = \max_k \frac{1}{|\lambda_k(A_n) - \lambda|}. \tag{344}$$

Note that, by the definition of δ, for $\lambda \in \partial B^k_{\delta+\varepsilon}$, the eigenvalue of A which is nearest to λ is given by $\lambda_k(A)$. Since this is exactly the center of the ball $B_{\delta+\varepsilon}$, it follows that

$$\sup_{\lambda \in \partial B^k_{\delta+\varepsilon}} \|R_\lambda(A)\|_{op} = \frac{1}{\delta + \varepsilon} < \infty. \tag{345}$$

Similarly, for $\lambda \in \partial B^k_{\delta+\varepsilon}$, the eigenvalue of A_n which is nearest to λ is given $\lambda_k(A_n)$ since n was chosen large enough such that $|\lambda_k(A_n) - \lambda_k(A)| < \varepsilon$ and $\varepsilon < r_0/2$. Since $\delta < \frac{r_0}{2} - \varepsilon$, it follows that the smallest distance from $\partial B^k_{\delta+\varepsilon}$ to $\lambda_k(A_n)$ is exactly δ and thus

$$\sup_{\lambda \in \partial B^k_{\delta+\varepsilon}} \|R_\lambda(A_n)\|_{op} = \frac{1}{\delta} < \infty. \tag{346}$$

Hence, we find that the RHS in (336) is bounded by $\frac{\|A_n - A\|_{op}}{(\delta+\varepsilon)\delta}$ for $\lambda \in \partial B^k_{\delta+\varepsilon}$. Finally, this yields

$$\begin{aligned}
\|\{A_n < 0\} - \{A < 0\}\|_{op} &\leqslant && (347)\\
&\leqslant (\delta+\varepsilon)(q-(K-1)) \max_{k \geqslant K} \left(\sup_{\lambda \in \partial B^k_{\delta+\varepsilon}} \|R_\lambda(A) - R_\lambda(A_n)\|_{op} \right) && (348)\\
&\leqslant (\delta+\varepsilon)(q-(K-1)) \|A_n - A\|_{op} \frac{1}{\delta+\varepsilon} \frac{1}{\delta} && (349)\\
&= \frac{q-(K-1)}{\delta} \|A_n - A\|_{op} \xrightarrow{n\to\infty} 0. && (350)
\end{aligned}$$

In an analogous way we can show that

$$\|\{A_n > 0\} - \{A > 0\}\|_{op} \leqslant \frac{R}{\delta} \|A_n - A\|_{op} \xrightarrow{n\to\infty} 0 \tag{351}$$

where R denotes the index of the smallest positive eigenvalue of A. This concludes the proof. □

Lemma 15. *The functions* $t \mapsto \alpha(P_{t,+})$ *and* $t \mapsto \alpha(P_{t,+} + P_{t,0})$ *are non-increasing.*

Proof. Let $0 \leqslant s < t$. We need to show that

$$\mathrm{Tr}\left[\sigma P_{s,+}\right] \geqslant \mathrm{Tr}\left[\sigma P_{t,+}\right] \quad \text{and} \quad \mathrm{Tr}\left[\sigma(P_{s,+} + P_{s,0})\right] \geqslant \mathrm{Tr}\left[\sigma(P_{t,+} + P_{t,0})\right]. \tag{352}$$

Note that for any Hermitian operator A on and for any Hermitian operator $0 \leqslant T \leqslant \mathbb{1}$ we have

$$\mathrm{Tr}\left[A\{A \geqslant 0\}\right] = \mathrm{Tr}\left[A\{A > 0\}\right] \geqslant \mathrm{Tr}\left[A \cdot T\right] \tag{353}$$

We first show $\mathrm{Tr}\left[\sigma P_{s,+}\right] \geqslant \mathrm{Tr}\left[\sigma P_{t,+}\right]$. Write $T_s := \rho - s\sigma$ and $T_t := \rho - t\sigma$ and note that the eigenprojections are Hermitian and satisfy

$$0 \leqslant \{T_s > 0\} \leqslant \mathbb{1}, \quad \text{and} \quad 0 \leqslant \{T_t > 0\} \leqslant \mathbb{1} \tag{354}$$

It follows that

$$\mathrm{Tr}\left[T_t\{T_t > 0\}\right] \geqslant \mathrm{Tr}\left[T_t\{T_s > 0\}\right] \tag{355}$$

and similarly

$$\mathrm{Tr}\left[T_s\{T_s > 0\}\right] \geqslant \mathrm{Tr}\left[T_s\{T_t > 0\}\right]. \tag{356}$$

Combining (355) and (356) yields

$$t \cdot \left(\mathrm{Tr}\left[\sigma\{T_s > 0\}\right] - \mathrm{Tr}\left[\sigma\{T_t > 0\}\right]\right) \geqslant s \cdot \left(\mathrm{Tr}\left[\sigma\{T_s > 0\}\right] - \mathrm{Tr}\left[\sigma\{T_t > 0\}\right]\right). \tag{357}$$

Since $s < t$ it follows that $\mathrm{Tr}\left[\sigma\{T_s > 0\}\right] \geqslant \mathrm{Tr}\left[\sigma\{T_t > 0\}\right]$ and thus $\alpha(P_{s,+}) \geqslant \alpha(P_{t,+})$. The other statement follows analogously by replacing $\{T_t > 0\}$ and $\{T_s > 0\}$ by $\{T_t \geqslant 0\}$ and $\{T_s \geqslant 0\}$. This concludes the proof. □

Lemma 16. *The function* $t \mapsto \alpha(P_{t,+})$ *is continuous from the right.*

Proof. Let $t \geqslant 0$ and let $\{t_n\}_{n\in\mathbb{N}} \subseteq [0,\infty)$ be a sequence such that $t_n \downarrow t$ (i.e. t_n converges to t from above). We show that $\lim_{n\to\infty} \alpha(P_{t_n,+}) = \alpha(P_{t,+})$. Let us define the operators

$$A_n := \rho - t_n\sigma, \tag{358}$$

and

$$A := \rho - t\sigma \tag{359}$$

and note that

$$\|A_n - A\|_{op} = |t_n - t| \cdot \|\sigma\|_{op} \xrightarrow{n\to\infty} 0. \tag{360}$$

Since, in addition, both A_n and A are Hermitian and $A - A_n = (t_n - t)\sigma \geqslant 0$, it follows from the second part of Lemma 14 that

$$\|\{A_n > 0\} - \{A > 0\}\|_{op} \xrightarrow{n\to\infty} 0 \tag{361}$$

and thus

$$\alpha(P_{t_n,+}) = \mathrm{Tr}\left[\sigma\{A_n > 0\}\right] \xrightarrow{n\to\infty} \mathrm{Tr}\left[\sigma\{A > 0\}\right] = \alpha(P_{t,+}) \tag{362}$$

since operator norm convergence implies convergence in the weak operator topology. This concludes the proof. □

Lemma 17. *The function* $t \mapsto \alpha(P_{t,+} + P_{t,0})$ *is continuous from the left.*

Proof. Let $\{t_n\}_{n\in\mathbb{N}} \subset [0,\infty)$ be a sequence of non-negative real numbers such that $t_n \uparrow t$ (i.e. t_n converges to t from below). Let A_n and A be the Hermitian operators defined by

$$A_n := \rho - t_n\sigma, \quad A := \rho - t\sigma \tag{363}$$

and note that $A - A_n = (t_n - t)\sigma \leqslant 0$ and $\|A_n - A\|_{op} \to 0$ as $n \to \infty$. It follows from the first part of Lemma 14 that

$$\|\{A_n < 0\} - \{A < 0\}\|_{op} \xrightarrow{n\to\infty} 0 \tag{364}$$

and thus, since operator norm convergence implies convergence in the weak operator topology, we have

$$\alpha(P_{t_n,+} + P_{t_n,0}) = \mathrm{Tr}\left[\sigma(\mathbb{1} - \{A_n < 0\})\right] \xrightarrow{n\to\infty} \mathrm{Tr}\left[\sigma(\mathbb{1} - \{A < 0\})\right] = \alpha(P_{t,+} + P_{t,0}). \tag{365}$$

This concludes the proof. □

Lemma 18. *For* $\alpha_0 \in [0,1]$ *we have the chain of inequalities*

$$\alpha\left(P_{\tau(\alpha_0),+}\right) \leqslant \alpha_0 \leqslant \alpha\left(P_{\tau(\alpha_0),+} + P_{\tau(\alpha_0),0}\right). \tag{366}$$

where $\tau(\alpha_0) := \inf\{t \geqslant 0 \colon\ \alpha(P_{t,+}) \leqslant \alpha_0\}$.

Proof. Recall that $\tau(\alpha_0) := \inf\{t \geqslant 0\colon\ \alpha(P_{t,+}) \leqslant \alpha_0\}$. Since, by Lemmas 15 and 16 the function $t \mapsto \alpha(P_{t,+})$ is non-decreasing and right-continuous, the left hand side of the inequality follows directly from the definition of $\tau(\alpha_0)$.

We now show the right hand side of the inequality. Note that if $t := \tau(\alpha_0) = 0$, then $P_{t,-} = \{\rho < 0\} = 0$ and thus $P_{t,+} + P_{t,0} = \mathbb{1}$ and $\alpha(P_{t,+} + P_{t,0}) = \mathrm{Tr}\,[\sigma] = 1 \geqslant \alpha_0$. Suppose that $\tau(\alpha_0) > 0$ and let $\{t_n\}_{n\in\mathbb{N}} \subset [0,\infty)$ be a sequence of non-negative real numbers such that $t_n \uparrow \tau(\alpha_0)$. Note that, since $t_n < \tau(\alpha_0)$ for all n and $t \mapsto \alpha(P_{t,+})$ is non-increasing and right-continuous by Lemmas 15 and 16, we have

$$\mathrm{Tr}\,[\sigma P_{t_n,+}] = \alpha(P_{t_n,+}) > \alpha_0 \tag{367}$$

by definition of $\tau(\alpha_0)$. Furthermore, since the projection $P_{t_n,0}$ is positive semidefinite, it follows that

$$\alpha(P_{t_n,+} + P_{t_n,0}) = \mathrm{Tr}\,[\sigma(P_{t_n,+} + P_{t_n,0})] \geqslant \mathrm{Tr}\,[\sigma P_{t_n,+}] = \alpha(P_{t_n,+}) > \alpha_0. \tag{368}$$

The Lemma now follows since by Lemma 17 the function $t \mapsto \alpha(P_{t,+} + P_{t,0})$ is continuous from the left and hence

$$\alpha(P_{\tau(\alpha_0),+} + P_{\tau(\alpha_0),0}) = \lim_{n\to\infty} \alpha(P_{t_n,+} + P_{t_n,0}) \geqslant \alpha_0 \tag{369}$$

which concludes the proof. □

B.1.1.1 *Optimality*

With these results, we can now see that the construction of Helstrom operators presented in the beginning of this section in (307) is well defined. Given $\alpha_0 \in [0, 1]$, it follows from Lemma 18 that firstly $\alpha_0 - \alpha\left(P_{\tau(\alpha_0),+}\right) \geqslant 0$, and secondly, since

$$\alpha_0 - \alpha\left(P_{\tau(\alpha_0),+}\right) \leqslant \alpha\left(P_{\tau(\alpha_0),+} + P_{\tau(\alpha_0),0}\right) - \alpha\left(P_{\tau(\alpha_0),+}\right) \tag{370}$$

we have

$$\frac{\alpha_0 - \alpha\left(P_{\tau(\alpha_0),+}\right)}{\alpha\left(P_{\tau(\alpha_0),0}\right)} \in [0, 1] \tag{371}$$

if $\alpha\left(P_{\tau(\alpha_0),0}\right) \neq 0$. It follows that, for any $\alpha_0 \in [0, 1]$, the Helstrom operator with type-I error probability α_0 is given by

$$M_{\tau(\alpha_0)} := P_{\tau(\alpha_0),+} + q_0 P_{\tau(\alpha_0),0}, \tag{372}$$

where

$$q_0 := \begin{cases} \frac{\alpha_0 - \alpha(P_{\tau(\alpha_0),+})}{\alpha(P_{\tau(\alpha_0),0})} & \text{if } \alpha\left(P_{\tau(\alpha_0),0}\right) \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{373}$$

is well defined. To see that $M_{\tau(\alpha_0)}$ indeed has type-I error probability α_0, note that if $q_0 \neq 0$ then

$$\alpha(M_{\tau(\alpha_0)}) = \mathrm{Tr}\left[\sigma M_{\tau(\alpha_0)}\right] = \mathrm{Tr}\left[\sigma P_{\tau(\alpha_0),+}\right] + q_0 \mathrm{Tr}\left[\sigma P_{\tau(\alpha_0),0}\right] = \alpha_0. \tag{374}$$

If on the other hand $q_0 = 0$, then, by Lemma 18, we have $\alpha_0 = \alpha(P_{\tau(\alpha_0),+}) = \alpha(M_{\tau(\alpha_0)})$. The following lemma shows optimality of the Helstrom operators, i.e. equation (313):

Lemma 19 (Optimality). *Let* $t \geqslant 0$ *and let* $M_t := P_{t,+} + X_t$ *for* $0 \leqslant X_t \leqslant P_{t,0}$ *be a Helstrom operator. Then, for any quantum hypothesis test* $0 \leqslant M \leqslant \mathbb{1}$ *for testing the null* σ *against the alternative* ρ*, the following implications hold*

(i) $\alpha(M) \leqslant \alpha(M_t) \;\Rightarrow\; \beta(M) \geqslant \beta(M_t)$

(ii) $\alpha(M) \geqslant 1 - \alpha(M_t) \;\Rightarrow\; 1 - \beta(M) \geqslant \beta(M_t)$

Proof. Let $\sum_i \lambda_i P_i$ be the spectral decomposition of the operator $\rho - t\sigma$ with orthogonal projections $\{P_i\}_i$ and associated eigenvalues $\{\lambda_i\}_i$. Recall that

$$\begin{aligned} P_{t,+} &:= \sum_{i\colon \lambda_i > 0} P_i, \\ P_{t,-} &:= \sum_{i\colon \lambda_i < 0} P_i, \\ P_{t,0} &:= \mathbb{1} - P_{t,+} - P_{t,-}. \end{aligned} \tag{375}$$

We notice that for any $0 \leqslant X_t \leqslant P_{t,0}$ we have

$$\begin{aligned} \operatorname{Tr}[(\rho - t\sigma)X_t] &= \operatorname{Tr}[(\rho - t\sigma)P_{t,+}X_t] + \operatorname{Tr}[(\rho - t\sigma)P_{t,-}X_t] \\ &\leqslant \operatorname{Tr}[(\rho - t\sigma)P_{t,+}P_{t,0}] \\ &= 0 \end{aligned} \tag{376}$$

and

$$\begin{aligned} \operatorname{Tr}[(\rho - t\sigma)X_t] &= \operatorname{Tr}[(\rho - t\sigma)P_{t,+}X_t] + \operatorname{Tr}[(\rho - t\sigma)P_{t,-}X_t] \\ &\geqslant \operatorname{Tr}[(\rho - t\sigma)P_{t,-}P_{t,0}] \\ &= 0 \end{aligned} \tag{377}$$

and thus $\operatorname{Tr}[(\rho - t\sigma)P_{t,0}] = \operatorname{Tr}[(\rho - t\sigma)X_t] = 0$. For the sequel, let $\bar{M}_t := \mathbb{1} - M_t$ and $\bar{M} := \mathbb{1} - M$.

We first show part (i) of the statement. Multiplying with the identity yields

$$\begin{aligned} M - M_t &= (\bar{M}_t + M_t)M - M_t(\bar{M} + M) \\ &= \bar{M}_t M - M_t \bar{M} \end{aligned} \tag{378}$$

and adding zero yields

$$\begin{aligned} \rho(M - M_t) &= (\rho - t\sigma)(M - M_t) + t\sigma(M - M_t) \\ &= (\rho - t\sigma)(\bar{M}_t M - M_t\bar{M}) + t\sigma(\bar{M}_t M - M_t\bar{M}). \end{aligned} \tag{379}$$

We need to show that $\beta(M_t) - \beta(M) = \operatorname{Tr}[\rho(M - M_t)] \leqslant 0$. Notice that

$$\operatorname{Tr}\left[(\rho - t\sigma)\bar{M}_t M\right] = -\operatorname{Tr}[(\rho - t\sigma)_- M] \leqslant 0 \tag{380}$$

and similarly

$$\operatorname{Tr}\left[(\rho - t\sigma)M_t\bar{M}\right] = \operatorname{Tr}\left[(\rho - t\sigma)_+\bar{M}\right] \geqslant 0 \tag{381}$$

where the inequalities follow from $\mathbb{1} \geqslant M \geqslant 0$. Finally, we see that

$$\begin{aligned} \operatorname{Tr}[\rho(M - M_t)] &= \operatorname{Tr}\left[(\rho - t\sigma)(\bar{M}_t M - M_t\bar{M})\right] + t \cdot \operatorname{Tr}\left[\sigma(\bar{M}_t M - M_t\bar{M})\right] \\ &\leqslant t \cdot \operatorname{Tr}\left[\sigma(\bar{M}_t M - M_t\bar{M})\right] \\ &= t \cdot \operatorname{Tr}[\sigma(M - M_t)] \\ &= t \cdot (\alpha(M) - \alpha(M_t)) \\ &\leqslant 0 \end{aligned} \tag{382}$$

where the last inequality follows from the assumption and $t \geqslant 0$.

Part (ii) now follows directly from part (i) by noting that $0 \leqslant M' := \mathbb{1} - M \leqslant \mathbb{1}$ and

$$\alpha(M) \geqslant 1 - \alpha(M_t) \Rightarrow \alpha(M') \leqslant \alpha(M_t) \overset{(i)}{\Longrightarrow} \beta(M_t) \leqslant \beta(M') = 1 - \beta(M). \tag{383}$$

This concludes the proof. □

B.1.2 *Main Proof*

Lemma 3 (restated). *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H}_d)$ be arbitrary quantum states, $\alpha_0 \in [0, 1]$ and $\epsilon \in [0, 1 - \alpha_0]$. Suppose that $\mathcal{F}(\rho, \sigma) \geqslant 1 - \epsilon$. Then*

$$\beta^*(\alpha_0; \rho, \sigma) \geqslant \alpha_0(2\epsilon - 1) + (1 - \epsilon) - 2\sqrt{\alpha_0 \epsilon (1 - \alpha_0)(1 - \epsilon)} \tag{384}$$

with equality if the states σ and ρ are pure and $\mathcal{F}(\rho, \sigma) = 1 - \epsilon$.

Proof. Let M_0 be a Helstrom operator such that $\alpha(M_0; \sigma) = \alpha_0$, i.e.

$$M_0 := P_{t_0,+} + q_0 \cdot P_{t_0,0}, \qquad \text{with} \tag{385}$$

where $t_0 := \inf\{t \geqslant 0 | \operatorname{Tr}[P_{t,+}\rho] \leqslant \alpha_0\}$ and

$$q_0 = \begin{cases} \frac{\alpha_0 - \alpha(P_{t_0,+}; \rho)}{\alpha(P_{t_0,0}; \rho)}, & \alpha(P_{t_0,0}; \rho) = 0 \\ 0 & \text{o.w.} \end{cases} \tag{386}$$

In addition, as an immediate consequence of part (i) of Lemma 19, for any $t \geqslant 0$, we have optimality, i. e.,

$$\beta^*(\alpha_0; \rho, \sigma) = \beta(M_0; \sigma). \tag{387}$$

We will now derive an explicit formula for the quantity $\beta(\Lambda_{t_0}; \sigma)$ for pure states $\rho = |\phi\rangle\langle\phi|$ and $\sigma = |\psi\rangle\langle\psi|$, and subsequently extend it to the general case for density matrices with arbitrary rank. For the sequel, let $\gamma = \langle\psi|\phi\rangle$. Since M_0 is a linear combination of the projections onto the eigenspaces of $\sigma - t\rho$, consider the eigenvalue problem

$$(\sigma - t\rho)|\eta\rangle = \eta|\eta\rangle. \tag{388}$$

Since here σ and ρ are projections of rank one, the operator $\sigma - t\rho$ has rank at most two and there exist at most two eigenstates $|\eta_0\rangle$, $|\eta_1\rangle$ corresponding to non-vanishing eigenvalues . In addition, they are linear combinations of $|\psi\rangle$, $|\phi\rangle$ so that we can write

$$|\eta_k\rangle = z_{k,\psi}|\psi\rangle + z_{k,\phi}|\phi\rangle, \quad k = 0, 1 \tag{389}$$

with constants $z_{k,\psi}$, $z_{k,\phi}$. Thus, solving the eigenvalue problem in (388) amounts to solving the problem

$$\begin{pmatrix} 1 & \gamma \\ -t\bar{\gamma} & -t \end{pmatrix} \cdot \begin{pmatrix} z_{k,\psi} \\ z_{k,\phi} \end{pmatrix} = \eta_k \begin{pmatrix} z_{k,\psi} \\ z_{k,\phi} \end{pmatrix}, \tag{390}$$

for which we find eigenvalues

$$\eta_0 = \frac{1}{2}(1-t) - R_t, \quad \eta_1 = \frac{1}{2}(1-t) + R_t \tag{391}$$

with

$$R_t = \sqrt{\frac{1}{4}(1-t)^2 + t(1-|\gamma|^2)}. \tag{392}$$

The corresponding eigenvectors $|\eta_0\rangle$, $|\eta_1\rangle$ are determined by their coefficients $z_{k,\psi}$, $z_{k,\phi}$ for $k = 0, 1$, for which we find

$$z_{k,\psi} = -\gamma A_k, \quad z_{k,\phi} = (1-\eta_k)A_k, \quad |A_k|^{-2} = 2R_t \left|\eta_k - 1 + |\gamma|^2\right| \tag{393}$$

where the coefficient A_k arises from requiring the eigenvectors $|\eta_k\rangle$ to be normalized (see [89], Section 8). Note that $\forall t \geqslant 0$ we have $\eta_0 \leqslant 0$ and $\eta_1 > 0$ so that $P_{t,+} = |\eta_1\rangle\langle\eta_1|$. Recall that we defined t_0 to be the positive number $t_0 := \inf\{t \geqslant 0\colon \operatorname{Tr}[P_{t,+}\rho] \leqslant \alpha_0\}$. It follows that

$$\operatorname{Tr}[P_{t,+}\rho] = |\langle\phi|\eta_1\rangle|^2 = \frac{1}{2}\left(1 - \frac{1+t-2|\gamma|^2}{\sqrt{(1+t)^2 - 4t|\gamma|^2}}\right) \tag{394}$$

and notice that the right hand side is continuous in t over $[0, \infty)$ whenever $|\gamma| < 1$. Since $t \mapsto \operatorname{Tr}[P_{t,+}\rho]$ is non-increasing, it's maximum is attained at $t = 0$ so that $\operatorname{Tr}[P_{t,+}\rho] \leqslant |\gamma|^2$ and hence $t_0 = 0$ if $\alpha_0 > |\gamma|^2$. In this case, we obtain $\beta^*(\alpha_0; \rho, \sigma) = 0$. If, on the other hand, $\alpha_0 \leqslant |\gamma|^2$, then we solve the equation $\operatorname{Tr}[P_{t,+}\rho] = \alpha_0$ and obtain the expression for t_0

$$t_0 = 2|\gamma|^2 - 1 - (2\alpha_0 - 1)\sqrt{\frac{|\gamma|^2(1-|\gamma|^2)}{\alpha_0(1-\alpha_0)}}. \tag{395}$$

For $t = t_0$ we have $\eta_0 < 0$ and $\eta_1 > 0$ so that $\Lambda_{t_0} = |\eta_1\rangle\langle\eta_1|$ and $|\eta_1\rangle = -\gamma A_1|\psi\rangle + (1-\eta_1)A_1|\phi\rangle$. Hence

$$\begin{aligned} \beta(M_0; \sigma) &= 1 - |\langle\eta_1|\psi\rangle|^2 \\ &= 1 - |A_1|^2|\gamma|^2\eta_1^2. \end{aligned} \tag{396}$$

Plugging t_0 into the expressions above yields

$$\begin{aligned} \beta^*(\alpha_0; \phi, \psi) &= \beta(M_0; \sigma) \\ &= \alpha_0 \cdot (1 - 2|\gamma|^2) + |\gamma|^2 - 2\sqrt{(1-\alpha_0)(1-|\gamma|^2)|\gamma|^2\alpha_0}. \end{aligned} \tag{397}$$

Since the right hand side of (397) is monotonically decreasing in $|\gamma|^2$ and $|\gamma|^2 \geqslant 1 - \epsilon$, the claim follows for pure states. To see that the above expression is also a valid lower bound for mixed states, let $|\Psi\rangle$ and $|\Phi\rangle$ be arbitrary purifications of σ and ρ respectively, both with purifying system $\mathcal{H}_E$. It is well known that β^* is monotonically increasing under the action of any completely positive and trace preserving map $\mathcal{E}$, i.e. $\beta^*(\alpha_0; \sigma, \rho) \leqslant \beta^*(\alpha_0; \mathcal{E}[\sigma], \mathcal{E}[\rho])$ for any $\alpha_0 \in [0, 1]$. Since the partial trace $\operatorname{Tr}_E[\cdot]$ is itself a CPTP map, we have the inequality

$$\begin{aligned} \beta^*(\alpha_0; \rho, \sigma) &= \beta^*(\alpha_0; \operatorname{Tr}_E[|\Phi\rangle\langle\Phi|], \operatorname{Tr}_E[|\Psi\rangle\langle\Psi|]) \\ &\geqslant \beta^*(\alpha_0; \Phi, \Psi) \\ &= \alpha_0 \cdot (1 - 2|\langle\Psi|\Phi\rangle|^2) + |\langle\Psi|\Phi\rangle|^2 - 2\sqrt{(1-\alpha_0)(1-|\langle\Psi|\Phi\rangle|^2)|\langle\Psi|\Phi\rangle|^2\alpha_0} \end{aligned} \tag{398}$$

where $\mathrm{Tr}_E[\cdot]$ denotes the partial trace over the purifying system. It follows from Uhlmann's Theorem that we can choose Ψ, Φ such that $|\langle\Psi|\Phi\rangle|^2 = \mathcal{F}(\rho, \sigma)$. The claim now follows from the observation that the RHS of (398) is monotonically decreasing in $|\langle\Psi|\Phi\rangle|^2$. This completes the proof. □

B.2 PROOF OF LEMMA 4

Lemma 4 (restated). *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H}_d)$ be density operators with fidelity $\mathcal{F}(\rho, \sigma) \geqslant 1-\epsilon$ for some $\epsilon \geqslant 0$ and let $A \geqslant 0$ be an observable. Let $\underline{m}, \bar{v} \geqslant 0$ such that*

$$\langle A\rangle_\sigma \geqslant \underline{m} \quad \textit{and} \quad \Delta A_\sigma \leqslant \bar{v}. \tag{399}$$

For ϵ with $\epsilon \leqslant \frac{\underline{m}^2}{\underline{m}^2+\bar{v}^2}$, a lower bound of $\langle A\rangle_\rho$ can be expressed as

$$\langle A\rangle_\rho \geqslant (1-\epsilon)\underline{m} - 2\bar{v}\sqrt{\epsilon(1-\epsilon)} + \epsilon\frac{\bar{v}^2}{\underline{m}}. \tag{400}$$

Proof. Let us recall that for any two pure states $|\psi\rangle$ and $|\phi\rangle$, the Gramian inequalities, derived in Section 5.2, read

$$\begin{aligned} |\langle\psi|\phi\rangle|\langle A\rangle_\phi - \Delta A_\phi\sqrt{1-|\langle\psi|\phi\rangle|^2} &\leqslant \Re(\langle\psi|A|\phi\rangle) \\ &\leqslant |\langle\psi|\phi\rangle|\langle A\rangle_\phi + \Delta A_\phi\sqrt{1-|\langle\psi|\phi\rangle|^2}. \end{aligned} \tag{401}$$

The first step is to show that these inequalities also hold for mixed states. Uhlmann's Theorem [225] states that for any two mixed states ρ and σ, we have

$$\mathcal{F}(\rho, \sigma) = \left\|\sqrt{\rho}\sqrt{\sigma}\right\|_1^2. \tag{402}$$

The trace norm in its variational form is given by $\|S\|_1 = \max_U |\mathrm{Tr}[US]|$ for arbitrary $S \in \mathcal{L}(\mathcal{H}_d)$ and where the maximization is taken over all unitaries. It follows that there exists U such that

$$\mathcal{F}(\rho, \sigma) = \left|\mathrm{Tr}\left[U\sqrt{\sigma}\sqrt{\rho}\right]\right|^2 \tag{403}$$

Let $|\Omega\rangle = \sum_{k=1}^d |k\rangle\otimes|k\rangle$ be the unnormalized maximally entangled state on $\mathcal{H}_d\otimes\mathcal{H}_d$ and note that the trace in the RHS of (403) can be rewritten as

$$\begin{aligned} \mathrm{Tr}\left[U\sqrt{\sigma}\sqrt{\rho}\right] &= \sum_k \langle k|U\sqrt{\sigma}\sqrt{\rho}|k\rangle \\ &= \sum_{k,l}\langle k|U\sqrt{\sigma}\sqrt{\rho}|l\rangle\langle k|l\rangle \\ &= \sum_{k,l}\langle k|\otimes\langle k|(U\sqrt{\sigma}\otimes\mathbb{1})(\sqrt{\rho}\otimes\mathbb{1})|l\rangle\otimes|l\rangle \\ &= \langle\Omega|(\sqrt{\sigma}\otimes(U^T)^\dagger)(\sqrt{\rho}\otimes\mathbb{1})|\Omega\rangle \end{aligned} \tag{404}$$

where the last equality follows from the definition of $|\Omega\rangle$ and the fact that $(\mathbb{1}\otimes U)|\Omega\rangle = (U^T\otimes\mathbb{1})|\Omega\rangle$. Define the (pure) states

$$|\Psi\rangle \equiv (\sqrt{\rho}\otimes\mathbb{1})|\Omega\rangle, \qquad |\Phi\rangle \equiv (\sqrt{\sigma}\otimes U^T)|\Omega\rangle \tag{405}$$

and note that these states are purifications of σ and ρ respectively and $\mathcal{F}(\rho, \sigma) = |\langle\Psi|\Phi\rangle|^2$. Furthermore, we have

$$\begin{aligned}\langle A \otimes \mathbb{1}\rangle_\Psi &= \langle\Omega|(\sqrt{\rho} \otimes \mathbb{1})(A \otimes \mathbb{1})(\sqrt{\rho} \otimes \mathbb{1})|\Omega\rangle \\ &= \mathrm{Tr}\left[\sqrt{\rho}A\sqrt{\rho}\right] \\ &= \langle A\rangle_\rho\end{aligned} \tag{406}$$

and similarly

$$\begin{aligned}\langle A \otimes \mathbb{1}\rangle_\Phi &= \langle\Omega|(\sqrt{\sigma} \otimes (U^\mathsf{T})^\dagger)(A \otimes \mathbb{1})(\sqrt{\sigma} \otimes U^\mathsf{T})|\Omega\rangle \\ &= \langle\Omega|(\sqrt{\sigma}A\sqrt{\sigma} \otimes (UU^\dagger)^\mathsf{T})|\Omega\rangle \\ &= \mathrm{Tr}\left[\sqrt{\sigma}A\sqrt{\sigma}\right] \\ &= \langle A\rangle_\sigma\end{aligned} \tag{407}$$

Replacing A with A^2, we find

$$\begin{aligned}(\Delta(A \otimes \mathbb{1})_\Phi)^2 &= \langle A^2 \otimes \mathbb{1}\rangle_\Phi - \langle A \otimes \mathbb{1}\rangle_\Phi^2 \\ &= \langle A^2\rangle_\sigma - \langle A\rangle_\sigma^2 \\ &= (\Delta A_\sigma)^2.\end{aligned} \tag{408}$$

Finally, we have

$$\begin{aligned}\langle\Psi|(A \otimes \mathbb{1})|\Phi\rangle &= \langle\Omega|(\sqrt{\rho}A\sqrt{\sigma}U \otimes \mathbb{1})|\Omega\rangle \\ &= \mathrm{Tr}\left[\sqrt{\rho}A\sqrt{\sigma}U\right] \\ &= \langle A\sqrt{\rho}, \sqrt{\sigma}U\rangle_{\mathrm{HS}}\end{aligned} \tag{409}$$

where $\langle\cdot, \cdot\rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt inner product. Without loss of generality, we assume that $\langle\Phi|\Psi\rangle$ is real and non-negative since otherwise we can multiply each purificiation by a global phase. Applying the Gramian inequalities to the observable $A \otimes \mathbb{1}$ and the purifications $|\Psi\rangle$, $|\Phi\rangle$, we find

$$\begin{aligned}\sqrt{\mathcal{F}(\rho, \sigma)}\langle A\rangle_\sigma - \Delta A_\sigma\sqrt{1 - \mathcal{F}(\rho, \sigma)} &= \langle\Phi|\Psi\rangle\langle A \otimes \mathbb{1}\rangle_\Phi - \Delta(A \otimes \mathbb{1})_\Phi\sqrt{1 - |\langle\Phi|\Psi\rangle|^2} \\ &\leqslant \Re(\langle\Psi|(A \otimes \mathbb{1})|\Phi\rangle) \\ &= \Re(\langle A\sqrt{\rho}, \sqrt{\sigma}U\rangle_{\mathrm{HS}}) \\ &\leqslant \langle\Phi|\Psi\rangle\langle A \otimes \mathbb{1}\rangle_\Phi + \Delta(A \otimes \mathbb{1})_\Phi\sqrt{1 - |\langle\Phi|\Psi\rangle|^2} \\ &= \sqrt{\mathcal{F}(\rho, \sigma)}\langle A\rangle_\sigma + \Delta A_\sigma\sqrt{1 - \mathcal{F}(\rho, \sigma)}.\end{aligned} \tag{410}$$

Thus, we have shown that inequalities similar to (401) also hold for mixed states. To finish the proof, note that by assumption $A \geqslant 0$ and hence A has a square root $A = A^{1/2}A^{1/2}$. The Cauchy-Schwarz inequality yields

$$\begin{aligned}\Re(\langle A\sqrt{\rho}, \sqrt{\sigma}U\rangle_{\mathrm{HS}}) &\leqslant \left|\langle A\sqrt{\rho}, \sqrt{\sigma}U\rangle_{\mathrm{HS}}\right| \\ &= \left|\langle A^{1/2}\sqrt{\rho}, A^{1/2}\sqrt{\sigma}U\rangle_{\mathrm{HS}}\right| \\ &\leqslant \left|\langle A^{1/2}\sqrt{\rho}, A^{1/2}\sqrt{\rho}\rangle_{\mathrm{HS}}\right|^{1/2} \times \left|\langle A^{1/2}\sqrt{\sigma}U, A^{1/2}\sqrt{\sigma}U\rangle_{\mathrm{HS}}\right|^{1/2} \\ &= |\mathrm{Tr}\left[A\rho\right]|^{1/2} \times |\mathrm{Tr}\left[A\sigma\right]|^{1/2}.\end{aligned} \tag{411}$$

Dividing the lower bound in (410) by $|\mathrm{Tr}\,[A\sigma]|^{1/2}$ leads to

$$|\mathrm{Tr}\,[A\rho]|^{1/2} \geqslant \sqrt{\mathcal{F}(\rho,\,\sigma)}\sqrt{\langle A\rangle_\sigma} - \frac{\Delta A_\sigma}{\sqrt{\langle A\rangle_\sigma}}\sqrt{1-\mathcal{F}(\rho,\,\sigma)} \tag{412}$$

Under the condition that $\sqrt{\mathcal{F}(\rho,\,\sigma)/(1-\mathcal{F}(\rho,\,\sigma))} \geqslant \Delta A_\sigma/\langle A\rangle_\sigma$, we can square both sides of the inequality

$$\begin{aligned}\langle A\rangle_\rho &\geqslant \left(\sqrt{\mathcal{F}(\rho,\,\sigma)}\sqrt{\langle A\rangle_\sigma} - \frac{\Delta A_\sigma}{\sqrt{\langle A\rangle_\sigma}}\sqrt{1-\mathcal{F}(\rho,\,\sigma)}\right)^2 \\ &= \mathcal{F}(\rho,\,\sigma)\langle A\rangle_\sigma - 2\Delta A_\sigma\sqrt{\mathcal{F}(\rho,\,\sigma)(1-\mathcal{F}(\rho,\,\sigma))} + \frac{1-\mathcal{F}(\rho,\,\sigma)}{\langle A\rangle_\sigma}(\Delta A_\sigma)^2.\end{aligned} \tag{413}$$

Note that the RHS is monotonically decreasing as $\mathcal{F}(\rho,\,\sigma)$ decreases. Hence, we can replace the true fidelity by a lower bound to it. In particular, for $\epsilon \geqslant 0$ with $\mathcal{F}(\rho,\,\sigma) \geqslant 1-\epsilon$ and $\sqrt{(1-\epsilon)/\epsilon} \geqslant \Delta A_\sigma/\langle A\rangle_\sigma$ we get

$$\langle A\rangle_\rho \geqslant (1-\epsilon)\langle A\rangle_\sigma - 2\sqrt{\epsilon(1-\epsilon)}\Delta A_\sigma + \epsilon\frac{(\Delta A_\sigma)^2}{\langle A\rangle_\sigma}. \tag{414}$$

Finally since this bound is also monotonically decreasing as $\langle A\rangle_\sigma$ decreases and ΔA_σ increases, we can replace these quantities by $\underline{m}$ and $\bar{v}$ and obtain

$$\langle A\rangle_\rho \geqslant (1-\epsilon)\underline{m} - 2\bar{v}\sqrt{\epsilon(1-\epsilon)} + \epsilon\frac{\bar{v}^2}{\underline{m}} \tag{415}$$

whenever

$$\epsilon \leqslant \frac{\underline{m}^2}{\underline{m}^2+\bar{v}^2} \tag{416}$$

which is the desired result. □

C

ADDITIONAL RESULTS IN PROVABLE ROBUSTNESS AGAINST BACKDOOR ATTACKS

C.1 EVALUATION AGAINST ADDITIONAL ATTACKS

ALL-TO-ALL ATTACKS Here, we consider all-to-all attacks which aim to trick the model such that it changes its prediction from the i-th class to the $((i+1)\%C)$-th class, where C is the number of classes. Different from the previous goal, here the model has to recognize both the image and the trigger to make a malicious prediction. Thus, intuitively, defenses like NC, which assume that the backdoored model only reacts to the backdoor trigger, are expected to perform worse.

The results of the all-to-all attack are shown in Table 11. We observe that our approach achieves a similar performance for empirical and certified robustness. The performance on MNIST and ImageNette is slightly better compared to the standard attack, while on CIFAR-10 it decreases slightly. We can observe that the performance of Mixup is also consistent with that on the standard attack. This is expected as Mixup also performs defense by processing the input and does not rely on model analysis. By comparison, the other baseline approaches based on model analysis do not achieve good performance here. We attribute this to the observation that in all-to-all attacks, the trained model needs to focus on both the original image and the trigger pattern, what makes it more difficult to detect the backdoors via model analysis compared to standard attacks where the model only focuses on the trigger pattern.

LARGER PERTURBATIONS Here we consider a larger perturbation consisting of a 4×4 trigger pattern with poison rate 20% and perturbation scale $\|\delta_i\| = 4.0$ on MNIST and $\|\delta_i\| = 4\sqrt{3}$ on CIFAR-10 and ImageNette (the $\sqrt{3}$ here comes from the fact that we add perturbation on all 3 channels). The results are shown in Table 12. We can see that such strong perturbation is too large to be within our certification radius, which is a limitation of our work. Therefore, the certified robust accuracy is 0. Nevertheless,

Table 11: Evaluation on DNNs with different datasets with an all-to-all attack goal. We use $\sigma = 0.5$ for MNIST and $\sigma = 0.2$ for CIFAR-10 and ImageNette. Vanilla denotes DNNs without RAB training and RAB-cert presents certified accuracy of RAB. The highest empirical robust accuracies are **bolded**.

	Backdoor Pattern	Acc. on Benign Instances		Empirical Robust Acc.								Certified Robust Acc.
		Vanilla	RAB	Vanilla	RAB	AC [32]	Spectral [223]	Sphere [207]	NC [234]	SCAn [215]	Mixup [18]	RAB-cert
MNIST	One-pixel	91.5%	90.2%	0%	**51.2%**	17.3%	3.0%	2.8%	28.4%	4.9%	37.1%	24.4%
	Four-pixel	91.6%	91.3%	0%	**60.3%**	16.1%	2.7%	1.8%	30.0%	1.8%	38.7%	39.9%
	Blending	91.5%	91.2%	0%	**59.7%**	15.4%	3.0%	1.8%	30.1%	4.7%	34.6%	39.1%
CIFAR-10	One-pixel	58.4%	52.2%	0%	24.9%	**26.7%**	5.7%	18.2%	13.2%	10.1%	19.7%	10.5%
	Four-pixel	57.5%	52.1%	0%	**25.1%**	11.2%	17.8%	18.3%	17.0%	13.3%	18.7%	11.6%
	Blending	58.3%	52.1%	0%	**24.8%**	10.0%	17.7%	15.9%	12.5%	10.7%	17.0%	10.9%
ImageNette	One-pixel	92.5%	93.0%	0%	43.1%	32.8%	19.6%	41.2%	23.5%	23.5%	**49.2%**	7.8%
	Four-pixel	93.6%	93.0%	0%	37.5%	18.8%	18.8%	43.8%	26.3%	21.7%	**58.3%**	18.7%
	Blending	95.0%	92.9%	0%	44.9%	46.9%	22.9%	34.7%	21.0%	14.3%	**49.0%**	16.3%

Table 12: Evaluation on DNNs with different datasets with a large attack perturbation. We use $\sigma = 0.5$ for MNIST and $\sigma = 0.2$ for CIFAR-10 and ImageNette. Vanilla denotes DNNs without RAB training and RAB-cert presents certified accuracy of RAB. The highest empirical robust accuracies are **bolded**.

	Backdoor Pattern	Acc. on Benign Instances		Empirical Robust Acc.								Certified Robust Acc.
		Vanilla	RAB	Vanilla	RAB	AC [32]	Spectral [223]	Sphere [207]	NC [234]	SCAn [215]	Mixup [18]	RAB-cert
MNIST	Large	86.8%	86.5%	0%	42.3%	65.5%	8.1%	0.6%	**70.9%**	11.9%	20.4%	0%
CIFAR-10	Large	52.1%	44.8%	0%	**27.2%**	20.88%	16.34%	11.96%	25.5%	8.6%	2.4%	0%
ImageNette	Large	84.7%	81.6%	0%	46.4%	62.6%	36.3%	1.5%	**74.9%**	55.5%	59.5%	0%

Table 13: Evaluation on Kernel KNN with different datasets. We use $\sigma = 0.5$ for MNIST and $\sigma = 0.2$ for CIFAR-10 and ImageNette. Vanilla denotes DNNs without RAB training and RAB-cert presents certified accuracy of RAB. The highest empirical robust accuracies are **bolded**.

	Backdoor Pattern	Acc. on Benign Instances		Empirical Robust Acc.								Certified Robust Acc.
		Vanilla	RAB	Vanilla	RAB	AC [32]	Spectral [223]	Sphere [207]	NC [234]	SCAn [215]	Mixup [18]	RAB-cert
MNIST	One-pixel	88.5%	78.2%	0%	35.7%	45.4%	53.0%	48.2%	53.0%	55.8%	**59.5%**	18.0%
	Four-pixel	88.5%	78.1%	0%	36.6%	50.6%	53.6%	48.3%	**69.9%**	55.6%	52.2%	18.8%
	Blending	88.4%	78.4%	0%	36.6%	44.8%	52.4%	47.4%	51.5%	**55.8%**	52.9%	18.8%
CIFAR-10	One-pixel	49.7%	46.5%	0%	21.6%	9.0%	24.9%	15.6%	16.5%	12.9%	**25.1%**	11.3%
	Four-pixel	49.5%	46.6%	0%	21.9%	15.9%	21.7%	**22.7%**	13.4%	15.0%	19.2%	11.7%
	Blending	49.8%	46.6%	0%	20.6%	17.0%	19.6%	15.1%	14.7%	16.8%	**21.8%**	10.5%
ImageNette	One-pixel	90.1%	88.6%	0%	35.3%	**56.8%**	22.2%	28.4%	40.9%	19.3%	31.3%	8.8%
	Four-pixel	90.7%	88.5%	0%	32.0%	**52.2%**	29.6%	41.5%	34.0%	30.8%	27.7%	7.6%
	Blending	91.5%	88.5%	0%	32.1%	**33.3%**	17.2%	2.5%	23.0%	13.8%	21.8%	7.6%

we can still achieve some non-trivial empirical robustness which is comparable with baseline approaches. This shows that our approach can be applied empirically to defend against strong backdoors with larger perturbation magnitude.

ADVERSARIAL ATTACKS ON RAB MODELS In [209], the authors show that if they smooth a backdoored model, the defended version will still be broken (i. e., with obvious adversarial pattern). We replicate the experiments on the RAB model by performing adversarial attacks against the RAB model. In order to perform the attack, we use the PGD attack where the gradient is calculated by aggregating the gradient on all the trained models. In Figure 31, We show the results on ImageNette with $\varepsilon = 60$ so that the pattern is the most clear. We observe that the adversarial examples look similar with those of unsmoothed model in [209]. Thus, the RAB pipeline is different from the smoothing process; rather, it is similar with a vanilla model.

Figure 31: Adversarial examples against the backdoored RAB model.

Table 14: Evaluation on Support Vector Machine (SVM) with different tabular datasets. We use $\sigma = 0.5$ for Spam and $\sigma = 0.2$ for `Adult` and `Mushroom`. Vanilla denotes DNNs without RAB training and RAB-cert presents certified accuracy of RAB. The highest empirical robust accuracies are **bolded**.

	Backdoor Pattern	Acc. on Benign Instances		Empirical Robust Acc.						Certified Robust Acc.
		Vanilla	RAB	Vanilla	RAB	AC [32]	Spectral [223]	Sphere [207]	SCAn [215]	RAB-cert
Spam	One-pixel	91.8%	88.4%	0%	**49.1%**	0%	18.3%	4.8%	12.9%	33.3%
	Four-pixel	91.2%	88.6%	0%	**48.2%**	0%	6.6%	7.4%	11.5%	32.1%
	Blending	92.0%	89.2%	0%	**44.7%**	0%	5.8%	5.8%	11.5%	29.8%
Adult	One-pixel	79.0%	77.2%	0%	**50.7%**	6.3%	15.3%	32.2%	8.4%	17.1%
	Four-pixel	77.4%	73.1%	0%	**53.0%**	5.4%	12.8%	14.4%	7.1%	21.5%
	Blending	78.8%	76.4%	0%	**55.9%**	8.0%	5.0%	11.6%	4.7%	26.1%
Mushroom	One-pixel	87.5%	82.0%	0%	**42.5%**	16.9%	0%	6.4%	17.3%	23.5%
	Four-pixel	86.6%	80.1%	0%	**42.2%**	14.2%	0%	2.8%	13.9%	22.5%
	Blending	87.4%	81.4%	0%	**43.5%**	13.1%	0%	11.1%	14.2%	24.0%

C.2 EVALUATION FOR ADDITIONAL MODELS

KERNEL-KNN We evaluate the defense on KNNs with a kernel function. The kernel function is learned with the convolution neural network trained on the supervised task and uses the hidden representation of the last layer before output as the kernel output. Note that in this case, our exact KNN certification algorithm cannot be applied since the output with Gaussian variable cannot be analyzed with the kernel function. Therefore, we use the same evaluation algorithm as in the case of DNNs to evaluate the certification performance. As shown in Table 13, our approach achieves worse performance than on DNNs, which is understandable since KNN models are known to usually underperform DNN models. On the other hand, we observe that many baselines actually have a better performance than DNN. We believe that the reason for this is that the baselines are based on the detection-and-removal algorithm. We found that the detection will only remove a subset of backdoored instances, so a trained DNN model remains backdoored; however, any removal of backdoored training data will help the performance of KNN since fewer backdoored instances will be viewed as a neighborhood, so the performance may improve. By comparison, RAB will not detect and remove instances and thus will not have a better performance on KNN.

SVM-BASED MODEL ON TABULAR DATA As our certification for DNN can be applied to any ML model, we now evaluate RAB on three tabular data - UCI Spambase dataset (Spambase) [51], and `Adult` and `Agaricus_lepiota` (Mushroom) in the Penn ML Benchmarks (PMLB) datasets[183]. These datasets are all binary classification tasks. Spambase contains 4,601 data points, with 57-dimensional input; Adult contains 48,842 data points with 14-dimensional input; Mushroom contains 8,145 data points with 22-dimensional input. We train a support vector machine (SVM) with RBF kernel using the default setting in scikit-learn toolkit [169]. As for the baselines where activation vectors are required, we use the output prediction vector as its representation, since there are no hidden activation layers in an SVM model.

The result of the SVM dataset is shown in Table 14. NC is not evaluated because it relies on anomaly detection among different classes, and therefore cannot be applied on these binary classification tasks; Mixup is not evaluated because it cannot be ap-

Table 15: Robustness of RAB on DNNs with and without test-time augmentation.

	Backdoor Pattern	With Aug		Without Aug	
		RAB	RAB-cert	RAB	RAB-cert
MNIST	One-pixel	41.2%	23.5%	27.0%	12.7%
	Four-pixel	40.7%	24.1%	27.4%	12.8%
	Blending	39.6%	23.1%	26.2%	12.1%
CIFAR-10	One-pixel	42.9%	24.5%	26.9%	15.2%
	Four-pixel	44.4%	25.7%	28.4%	16.4%
	Blending	42.8%	24.1%	27.8%	15.8%
ImageNette	One-pixel	38.6%	15.9%	22.7%	5.1%
	Four-pixel	38.4%	12.6%	22.6%	8.2%
	Blending	29.9%	9.2%	18.7%	4.1%

plied in the SVM training algorithm. We can see that our approach still achieves good robustness both empirically and certifiably. In contrast, the baseline approaches do not perform well as they are designed specifically for deep neural networks. In the SVM case, where they use the output as the representation vector, the detection performance is not favourable.

C.3 ABLATIONS

TEST-TIME AUGMENTATION Table 15 shows the comparison of empirical and certified robustness with and without test-time augmentation. We see that the test-time augmentation indeed helps with the model robustness both empirically and certifiably.

Table 16: The abstain rate of the certification on DNNs.

	Backdoor Pattern	Abstain Rate
MNIST	One-pixel	3.32%
	Four-pixel	3.21%
	Blending	3.02%
CIFAR-10	One-pixel	5.59%
	Four-pixel	6.00%
	Blending	5.29%
ImageNette	One-pixel	3.89%
	Four-pixel	4.08%
	Blending	1.90%

ABSTAIN RATE Table 16 shows the abstain rate of RAB against attacks. We see that in general, the abstain rate is relatively low and will not be a serious concern in the pipeline. Note that if the denial-of-service attack is indeed a concern, we can proceed in a similar way as in [39] to prove a certified radius in which we can certify our defense rather than abstaining the input.

Table 17: The mean and standard deviation of the RAB robustness on DNNs with 5 runs.

	Backdoor Pattern	RAB	RAB-cert
MNIST	One-pixel	40.79±0.72%	23.36±0.52%
	Four-pixel	40.27±0.87%	24.37±0.49%
	Blending	40.72±0.65%	23.58±0.88%
CIFAR-10	One-pixel	42.66±0.29%	24.35±0.31%
	Four-pixel	42.56±0.32%	25.25±0.37%
	Blending	42.89±0.21%	23.95±0.17%
ImageNette	One-pixel	38.64±0.80%	15.45±0.94%
	Four-pixel	37.23±0.69%	12.45±0.82%
	Blending	28.74±1.15%	9.20±1.40%

STABILITY To see the stability of RAB, we run our algorithm 5 times and report the mean and standard deviation in Table 17. We can see that the standard deviation is relatively small, indicating that our algorithm is stable.

D

ADDITIONAL RESULTS FROM TRANSFORMATION-SPECIFIC SMOOTHING FOR ROBUSTNESS CERTIFICATION

D.1 DERIVATIONS OF ROBUSTNESS BOUNDS

Here, we instantiate Theorem 5 with different smoothing distributions and solve the robustness condition (125) for the case where the distribution of ε_1 results from shifting the distribution of ε_0, i.e., $\varepsilon_1 = \alpha + \varepsilon_0$. For ease of notation, let $\zeta\colon \mathbb{R}_{\geqslant 0} \to [0, 1]$ be the function defined by

$$t \mapsto \zeta(t) := \mathbb{P}_0\left(\overline{S}_t\right) \tag{417}$$

where $\mathbb{P}_0$ is the distribution of ε_0 and $\overline{S}_t$ is a lower level set; recall that the definitions of lower level sets is

$$\begin{aligned} \underline{S}_t &:= \{z \in \mathcal{Z}\colon\ \Lambda(z) < t\}, \quad \overline{S}_t := \{z \in \mathcal{Z}\colon\ \Lambda(z) \leqslant t\}, \\ &\text{where} \quad \Lambda(z) := \frac{f_1(z)}{f_0(z)}. \end{aligned} \tag{418}$$

Note that the generalized inverse of ζ corresponds to τ_p, i.e.,

$$\zeta^{-1}(p) = \inf\{t \geqslant 0 |\ \zeta(t) \geqslant p\} = \tau_p \tag{419}$$

and the function ξ is correspondingly given by

$$\xi(p) = \sup\{\mathbb{P}_1(S) |\ \underline{S}(\zeta^{-1}(p)) \subseteq S \subseteq \overline{S}(\zeta^{-1}(p))\} \tag{420}$$

Finally, we make the following definition in order to reduce clutter and simplify the notation. This definition will be used throughout the proofs presented here.

Definition 10 ((p_A, p_B)-Confident Classifier)**.** *Let $x \in \mathcal{X}$, $y_A \in \mathcal{Y}$ and p_A, $p_B \in [0, 1]$ with $p_A > p_B$. We say that the ε-smoothed classifier q is (p_A, p_B)-confident at x if*

$$q(y_A | x; \varepsilon) \geqslant p_A \geqslant p_B \geqslant \max_{y \neq y_A} q(y | x; \varepsilon). \tag{421}$$

D.1.1 *Gaussian Smoothing*

Corollary 8. *Suppose $\mathcal{Z} = \mathbb{R}^m$, $\Sigma := \operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2)$, $\varepsilon_0 \sim \mathcal{N}(0, \Sigma)$ and $\varepsilon_1 := \alpha + \varepsilon_0$ for some $\alpha \in \mathbb{R}^m$. Suppose that the ε_0-smoothed classifier g is (p_A, p_B)-confident at $x \in \mathcal{X}$ for some $y_A \in \mathcal{Y}$. Then, it holds that $q(y_A | x; \varepsilon_1) > \max_{y \neq y_A} q(y | x; \varepsilon_1)$ if α satisfies*

$$\sqrt{\sum_{i=1}^{m} \left(\frac{\alpha_i}{\sigma_i}\right)^2} < \frac{1}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right). \tag{422}$$

Proof. By Theorem 5 we know that if ε_1 satisfies

$$\xi(p_A) + \xi(1 - p_B) > 1, \tag{423}$$

then it is guaranteed that

$$q(y_A|x;\varepsilon_1) > \max_{y\neq y_A} q(y|x;\varepsilon_1). \tag{424}$$

The proof is thus complete if we show that (423) reduces to (422). For that purpose, denote by f_0 and f_1 density functions of ε_0 and ε_1, respectively. Let $A := \Sigma^{-1}$ and note that the bilinear form $(z_1, z_2) \mapsto z_1^T A z_2 =: \langle z_1, z_2\rangle_A$ defines an inner product on $\mathbb{R}^m$. Let $z \in \mathbb{R}^m$ and consider

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} = \frac{\exp\left(-\frac{1}{2}\langle z-\alpha, z-\alpha\rangle_A\right)}{\exp\left(-\frac{1}{2}\langle z, z\rangle_A\right)} \tag{425}$$

$$= \exp\left(\langle z, \alpha\rangle_A - \frac{1}{2}\langle\alpha, \alpha\rangle_A\right) \tag{426}$$

and thus

$$\Lambda(z) \leqslant t \iff \langle z, \alpha\rangle_A \leqslant \log(t) + \frac{1}{2}\langle\alpha, \alpha\rangle. \tag{427}$$

Let $Z \sim \mathcal{N}(0, 1)$ and notice that

$$\frac{\langle\varepsilon_0, \alpha\rangle_A}{\sqrt{\langle\alpha, \alpha\rangle_A}} \overset{d}{=} Z \overset{d}{=} \frac{\langle\varepsilon_1, \alpha\rangle_A - \langle\alpha, \alpha\rangle_A}{\sqrt{\langle\alpha, \alpha\rangle_A}}. \tag{428}$$

Let $\partial_t := \overline{S}_t \setminus \underline{S}_t = \{z\colon \Lambda(z) = t\}$ and notice that $\mathbb{P}_0(\partial_t) = \mathbb{P}_1(\partial_t) = 0$ and $\mathbb{P}_0(\underline{S}_t) = \mathbb{P}_0(\overline{S}_t)$. Similarly, it holds that $\mathbb{P}_1(\underline{S}_t) = \mathbb{P}_1(\overline{S}_t)$. The function $p \mapsto \xi(p)$ is thus given by

$$\xi(p) = \mathbb{P}_1\left(\overline{S}_{\zeta^{-1}(p)}\right). \tag{429}$$

We compute ζ as

$$\begin{aligned}\zeta(t) &= \mathbb{P}(\Lambda(\varepsilon_0) \leqslant t)\\ &= \mathbb{P}\left(\langle\varepsilon_0, \alpha\rangle_A \leqslant \log(t) + \frac{1}{2}\langle\alpha, \alpha\rangle_A\right)\\ &= \Phi\left(\frac{\log(t) + \frac{1}{2}\langle\alpha, \alpha\rangle_A}{\sqrt{\langle\alpha, \alpha\rangle_A}}\right)\end{aligned} \tag{430}$$

and for $p \in [0, 1]$ its inverse is

$$\zeta^{-1}(p) = \exp\left(\Phi^{-1}(p)\sqrt{\langle\alpha, \alpha\rangle_A} - \frac{1}{2}\langle\alpha, \alpha\rangle_A\right). \tag{431}$$

Thus

$$\begin{aligned}\mathbb{P}\left(\Lambda(\varepsilon_1) \leqslant \zeta^{-1}(p)\right) &= \mathbb{P}\left(\frac{\langle\varepsilon_1, \alpha\rangle_A - \langle\alpha, \alpha\rangle_A}{\sqrt{\langle\alpha, \alpha\rangle_A}} \leqslant \frac{\log(\zeta^{-1}(p)) - \frac{1}{2}\langle\alpha, \alpha\rangle_A}{\sqrt{\langle\alpha, \alpha\rangle_A}}\right)\\ &= \Phi\left(\frac{\left(\Phi^{-1}(p)\sqrt{\langle\alpha, \alpha\rangle_A} - \frac{1}{2}\langle\alpha, \alpha\rangle_A\right) - \frac{1}{2}\langle\alpha, \alpha\rangle_A}{\sqrt{\langle\alpha, \alpha\rangle_A}}\right)\\ &= \Phi\left(\Phi^{-1}(p) - \sqrt{\langle\alpha, \alpha\rangle_A}\right).\end{aligned} \tag{432}$$

Finally, algebra shows that

$$\Phi\left(\Phi^{-1}(p_A) - \sqrt{\langle \alpha, \alpha \rangle_A}\right) + \Phi\left(\Phi^{-1}(1-p_B) - \sqrt{\langle \alpha, \alpha \rangle_A}\right) > 1 \tag{433}$$

is equivalent to

$$\sqrt{\sum_{i=1}^{m} \left(\frac{\alpha_i}{\sigma_i}\right)^2} < \frac{1}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right) \tag{434}$$

what concludes the proof. □

D.1.2 *Exponential Smoothing*

Corollary 9. *Suppose* $\mathcal{Z} = \mathbb{R}^m_{\geqslant 0}$, *fix some* $\lambda > 0$ *and let* $\varepsilon_{0,i} \overset{\text{iid}}{\sim} \text{Exp}(1/\lambda)$, $\varepsilon_0 := (\varepsilon_{0,1}, \ldots, \varepsilon_{0,m})^T$ *and* $\varepsilon_1 := \alpha + \varepsilon_0$ *for some* $\alpha \in \mathbb{R}^m_{\geqslant 0}$. *Suppose that the* ε_0*-smoothed classifier* g *is* (p_A, p_B)*-confident at* $x \in \mathcal{X}$ *for some* $y_A \in \mathcal{Y}$. *Then, it holds that* $q(y_A|x;\varepsilon_1) > \max_{y \neq y_A} q(y|x;\varepsilon_1)$ *if* α *satisfies*

$$\|\alpha\|_1 < -\frac{\log(1-p_A+p_B)}{\lambda}. \tag{435}$$

Proof. By Theorem 5 we know that if ε_1 satisfies

$$\xi(p_A) + \xi(1-p_B) > 1, \tag{436}$$

then it is guaranteed that

$$q(y_A|x;\varepsilon_1) > \max_{y \neq y_A} q(y|x;\varepsilon_1). \tag{437}$$

The proof is thus complete if we show that (436) reduces to (435). For that purpose, denote by f_0 and f_1 density functions of ε_0 and ε_1, respectively, and note that

$$f_1(z) = \begin{cases} \lambda \cdot \exp(-\lambda\|z-\alpha\|_1), & \min_i(z_i - \alpha_i) \geqslant 0, \\ 0, & \text{otherwise}, \end{cases} \tag{438}$$

$$f_0(z) = \begin{cases} \lambda \cdot \exp(-\lambda\|z\|_1), & \min_i(z_i) \geqslant 0, \\ 0, & \text{otherwise}, \end{cases} \tag{439}$$

and $\forall i$, $z_i - \alpha_i \leqslant z_i$ and hence $f_0(z) = 0 \Rightarrow f_1(z) = 0$. Thus

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} = \begin{cases} \exp\left(\lambda \cdot \|\alpha\|_1\right) & \min_i(z_i - \alpha_i) \geqslant 0, \\ 0, & \text{otherwise}. \end{cases} \tag{440}$$

Let $S_0 := \{z \in \mathbb{R}^m_{\geqslant 0} \colon \min_i(z_i - \alpha_i) < 0\}$ and note that due to independence

$$\mathbb{P}_0(S_0) = \mathbb{P}\left(\bigcup_{i=1}^m \{\varepsilon_{0,i} < \alpha_i\}\right) \tag{441}$$

$$= 1 - \mathbb{P}\left(\bigcap_{i=1}^m \{\varepsilon_{0,i} \geqslant \alpha_i\}\right) \tag{442}$$

$$= 1 - \prod_{i=1}^m \mathbb{P}(\varepsilon_{0,i} \geqslant \alpha_i) \tag{443}$$

$$= 1 - \prod_{i=1}^m (1 - (1 - \exp(-\lambda\,\alpha_i))) \tag{444}$$

$$= 1 - \exp(-\lambda \|\alpha\|_1)\,. \tag{445}$$

Let $t_\alpha := \exp(\lambda\|\alpha\|_1)$ and compute ζ as

$$\zeta(t) = \mathbb{P}(\Lambda(\varepsilon_0) \leqslant t) \tag{446}$$

$$= \mathbb{P}\left(\mathbb{1}\{\min_i(\varepsilon_{0,i} - \alpha_i) \geqslant 0\} \leqslant t \cdot \exp(-\lambda\|\alpha\|_1)\right) \tag{447}$$

$$= \begin{cases} 1 - \exp(-\lambda\|\alpha\|_1) & t < t_\alpha, \\ 1 & t \geqslant t_\alpha. \end{cases} \tag{448}$$

Recall that $\zeta^{-1}(p) := \inf\{t \geqslant 0 \colon \zeta(t) \geqslant p\}$ for $p \in [0, 1]$ and hence

$$\zeta^{-1}(p) = \begin{cases} 0 & p \leqslant 1 - \exp(-\lambda\|\alpha\|_1), \\ \exp(\lambda\|\alpha\|_1) & p > 1 - \exp(-\lambda\|\alpha\|_1). \end{cases} \tag{449}$$

In order to evaluate ξ, we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R}^m_{\geqslant 0} \colon \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R}^m_{\geqslant 0} \colon \Lambda(z) \leqslant t\}$ and consider

$$\begin{aligned} \underline{S}_{\zeta^{-1}(p)} = &\left(S_0^c \cap \left\{z \in \mathbb{R}^m_{\geqslant 0} \colon \exp(\lambda\|\alpha\|_1) < \zeta^{-1}(p)\right\}\right) \\ &\cup \left(S_0 \cap \left\{z \in \mathbb{R}^m_{\geqslant 0} \,|\, 0 < \zeta^{-1}(p)\right\}\right) \end{aligned} \tag{450}$$

$$= \begin{cases} \varnothing & p \leqslant 1 - \exp(-\lambda\|\alpha\|_1), \\ S_0 & p > 1 - \exp(-\lambda\|\alpha\|_1) \end{cases} \tag{451}$$

and

$$\begin{aligned} \overline{S}_{\zeta^{-1}(p)} = &\left(S_0^c \cap \left\{z \in \mathbb{R}^m_{\geqslant 0} \colon \exp(\lambda\|\alpha\|_1) \leqslant \zeta^{-1}(p)\right\}\right) \\ &\dot\cup \left(S_0 \cap \left\{z \in \mathbb{R}^m_{\geqslant 0} \colon 0 \leqslant \zeta^{-1}(p)\right\}\right) \end{aligned} \tag{452}$$

$$= \begin{cases} S_0 & p \leqslant 1 - \exp(-\lambda\|\alpha\|_1), \\ \mathbb{R}^m_+ & p > 1 - \exp(-\lambda\|\alpha\|_1). \end{cases} \tag{453}$$

Suppose that $p \leqslant 1 - \exp(-\lambda\|\alpha\|_1)$. Then we have that $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = S_0$ and hence

$$p \leqslant 1 - \exp(-\lambda\|\alpha\|_1) \quad \Rightarrow \quad \xi(p) = \sup\{\mathbb{P}_1(S) \colon S \subseteq S_0 \wedge \mathbb{P}_0(S) \leqslant p\} = 0. \tag{454}$$

Condition (436) can thus be satisfied only if $p_A > 1 - \exp(-\lambda\|\alpha\|_1)$ and $1 - p_B > 1 - \exp(-\lambda\|\alpha\|_1)$. In this case $\underline{S}_{\zeta^{-1}(p)} = S_0$ and $\overline{S}_{\zeta^{-1}(p)} = \mathbb{R}^m_{\geqslant 0}$. For $p \in [0, 1]$ let $\mathcal{S}_p = \{S \subseteq \mathbb{R}^m_{\geqslant 0} \colon S_0 \subseteq S \subseteq \mathbb{R}^m_{\geqslant 0},\ \mathbb{P}_0(S) \leqslant p\}$. Then

$$p > 1 - \exp(-\lambda\|\alpha\|_1) \quad \Rightarrow \quad \xi(p) = \sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S). \tag{455}$$

We can write any $S \in \mathcal{S}_p$ as the disjoint union $S = S_0 \dot\cup T$ for some $T \subseteq \mathbb{R}^m_{\geqslant 0}$ such that $\mathbb{P}_0(S_0 \dot\cup T) \leqslant p$. Note that $\mathbb{P}_1(S_0) = 0$ and since $S_0 \cap T = \varnothing$ any $z \in T$ satisfies $0 \leqslant \min_i (z_i - \alpha_i) \leqslant \min_i z_i$ and hence $\Lambda(z) = \exp(\lambda\|\alpha\|_1)$. Thus

$$\begin{aligned} \mathbb{P}_1(S) &= \mathbb{P}_1(T) \\ &= \int_T f_1(z)\,dz \\ &= \int_T \exp(\lambda\|\alpha\|_1) f_0(z)\,dz \\ &= \exp(\lambda\|\alpha\|_1) \cdot \mathbb{P}_0(T). \end{aligned} \tag{456}$$

Thus, The supremum of the left hand side over all $S \in \mathcal{S}_p$ equals the supremum of the right hand side over all $T \in \{T' \subseteq S_0^c \colon \mathbb{P}_0(T') \leqslant 1 - \mathbb{P}_0(S_0)\}$

$$\begin{aligned} \sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S) &= \exp(\lambda\|\alpha\|_1) \cdot \sup\{\mathbb{P}_1(T') \colon T' \subseteq S_0^c,\ \mathbb{P}_0(T') \leqslant p - \mathbb{P}_0(S_0)\} \\ &= \exp(\lambda\|\alpha\|_1) \cdot (p - \mathbb{P}_0(S_0)). \end{aligned} \tag{457}$$

Computing ξ at p_A thus yields

$$\begin{aligned} \xi(p_A) &= \sup_{S \in \mathcal{S}_{p_A}} \mathbb{P}_1(S) \\ &= \exp(\lambda\|\alpha\|_1) \cdot (p_A - \mathbb{P}_0(S_0)) \\ &= \exp(\lambda\|\alpha\|_1) \cdot (p_A - (1 - \exp(-\lambda\|\alpha\|_1))) \\ &= \exp(\lambda\|\alpha\|_1) \cdot (p_A + \exp(-\lambda\|\alpha\|_1) - 1) \end{aligned} \tag{458}$$

where the third equality follows from (445). Similarly, computing ξ at $1 - p_B$ yields

$$\xi(1 - p_B) = \sup_{S \in \mathcal{S}_{1-p_B}} \mathbb{P}_1(S) \tag{459}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot (1 - p_B - \mathbb{P}_0(S_0)) \tag{460}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot (1 - p_B - (1 - \exp(-\lambda\|\alpha\|_1))) \tag{461}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot (-p_B + \exp(-\lambda\|\alpha\|_1)). \tag{462}$$

Finally, condition (436) is satisfied whenever α satisfies

$$\begin{aligned} &\exp(\lambda\|\alpha\|_1)\cdot(p_A+\exp(-\lambda\|\alpha\|_1)-1) \\ &\qquad +\exp(\lambda\|\alpha\|_1)\cdot(-p_B+\exp(-\lambda\|\alpha\|_1))>1 \end{aligned} \tag{463}$$

$$\iff$$

$$\begin{aligned} &\exp(-\lambda\|\alpha\|_1)+p_B-\exp(-\lambda\|\alpha\|_1) \\ &\qquad <p_A+\exp(-\lambda\|\alpha\|_1)-1 \end{aligned} \tag{464}$$

$$\iff$$

$$1-p_A+p_B<\exp(-\lambda\|\alpha\|_1) \tag{465}$$

$$\iff$$

$$\|\alpha\|_1<-\frac{\log(1-p_A+p_B)}{\lambda} \tag{466}$$

what completes the proof. □

D.1.3 *Uniform Smoothing*

Corollary 10. *Suppose $\mathcal{Z}=\mathbb{R}^m$, and $\varepsilon_0\sim\mathcal{U}([a,b]^m)$ for some $a<b$. Set $\varepsilon_1:=\alpha+\varepsilon_0$ for $\alpha\in\mathbb{R}^m$. Suppose that the ε_0-smoothed classifier g is (p_A,p_B)-confident at $x\in\mathcal{X}$ for some $y_A\in\mathcal{Y}$. Then, it holds that $q(y_A|x;\varepsilon_1)>\max_{y\neq y_A}q(y|x;\varepsilon_1)$ if α satisfies*

$$1-\left(\frac{p_A-p_B}{2}\right)<\prod_{i=1}^{m}\left(1-\frac{|\alpha_i|}{b-a}\right)_+ \tag{467}$$

where $(x)_+:=\max\{x,0\}$.

Proof. By Theorem 5 we know that if ε_1 satisfies

$$\xi(p_A)+\xi(1-p_B)>1, \tag{468}$$

then it is guaranteed that

$$q(y_A|x;\varepsilon_1)>\max_{y\neq y_A}q(y|x;\varepsilon_1). \tag{469}$$

The proof is thus complete if we show that (468) reduces to (467). For that purpose, denote by f_0 and f_1 density functions of ε_0 and ε_1, respectively, and let $I_0=[a,b]^m$ and $I_1:=\prod_{i=1}^m[a+\alpha_i,b+\alpha_i]$ bet the support of ε_0 and ε_1. Consider

$$f_0(z)=\begin{cases}(b-a)^{-m} & z\in I_0,\\ 0 & \text{otherwise}\end{cases} \tag{470}$$

$$f_1(z)=\begin{cases}(b-a)^{-m} & z\in I_1,\\ 0 & \text{otherwise.}\end{cases} \tag{471}$$

Let $S_0:=I_0\setminus I_1$. Then, for any $z\in I_0\cup I_1$

$$\Lambda(z)=\frac{f_1(z)}{f_0(z)}=\begin{cases}0 & z\in S_0,\\ 1 & z\in I_0\cap I_1,\\ \infty & z\in I_1\setminus I_0.\end{cases} \tag{472}$$

Note that

$$\mathbb{P}_0(S_0) = 1 - \mathbb{P}_0(I_1) \tag{473}$$

$$= 1 - \prod_{i=1}^{m} \mathbb{P}(a + \alpha_i \leqslant \varepsilon_{0,i} \leqslant b + \alpha_i) \tag{474}$$

$$= 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{475}$$

where $(x)_+ = \max\{x, 0\}$. We then compute ζ for $t \geqslant 0$

$$\zeta(t) = \mathbb{P}(\Lambda(\varepsilon_0) \leqslant t) = \begin{cases} \mathbb{P}_0(S_0) & t < 1, \\ \mathbb{P}_0(I_0) & t \geqslant 1. \end{cases} \tag{476}$$

$$= \begin{cases} 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ & t < 1, \\ 1 & t \geqslant 1. \end{cases} \tag{477}$$

Recall that $\zeta^{-1}(p) := \inf\{t \geqslant 0 \colon \zeta(t) \geqslant p\}$ for $p \in [0, 1]$ and hence

$$\zeta^{1}(p) = \begin{cases} 0 & p \leqslant 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+, \\ 1 & p > 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+. \end{cases} \tag{478}$$

In order to evaluate ξ, we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R}^m_{\geqslant 0} \colon \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R}^m_{\geqslant 0} \colon \Lambda(z) \leqslant t\}$ and consider

$$\underline{S}_{\zeta^{-1}(p)} = \begin{cases} \varnothing & p \leqslant 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+, \\ S_0 & p > 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \end{cases} \tag{479}$$

and

$$\overline{S}_{\zeta^{-1}(p)} = \begin{cases} S_0 & p \leqslant 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+, \\ I_0 & p > 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \end{cases} \tag{480}$$

Suppose $p \leqslant 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+$. Then $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = S_0$ and hence

$$p \leqslant 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{481}$$

$$\Rightarrow \xi(p) = \sup\{\mathbb{P}_1(S) \colon S \subseteq S_0, \mathbb{P}_0(S) \leqslant p\} = 0.$$

Condition (468) can thus be satisfied only if $p_A > 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+$ and $1 - p_B > 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+$. In this case $\underline{S}_{\zeta^{-1}(p)} = S_0$ and $\overline{S}_{\zeta^{-1}(p)} = I_0$. For $p \in [0, 1]$ let $\mathcal{S}_p = \{S \subseteq \mathbb{R}^m \colon S_0 \subseteq S \subseteq I_0, \mathbb{P}_0(S) \leqslant p\}$. Then

$$p > 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \Rightarrow \xi(p) = \sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S). \tag{482}$$

We can write any $S \in \mathcal{S}_p$ as the disjoint union $S = S_0 \dot\cup T$ for some $T \subseteq I_0 \cap I_1$ such that $\mathbb{P}_0(S_0 \dot\cup T) \leqslant p$. Note that $\mathbb{P}_1(S_0) = 0$ and for any $z \in T$, we have $f_0(z) = f_1(z)$. Hence

$$\mathbb{P}_1(S) = \mathbb{P}_1(T) = \mathbb{P}_0(T) \tag{483}$$

$$\leqslant p - \mathbb{P}_0(S_0) = p - \left(1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right). \tag{484}$$

Thus, The supremum of the left hand side over all $S \in \mathcal{S}_p$ equals the supremum of the right hand side over all $T \in \{T' \subseteq I_0 \cap I_1 \colon \mathbb{P}_0(T') \leqslant 1 - \mathbb{P}_0(S_0)\}$

$$\sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S) = \sup\{\mathbb{P}_1(T') \colon T' \subseteq I_0 \cap I_1, \mathbb{P}_0(T') \leqslant p - \mathbb{P}_0(S_0)\} \tag{485}$$

$$= p - \left(1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right). \tag{486}$$

Hence, computing ξ at p_A and $1 - p_B$ yields

$$\xi(p_A) = p_A - \left(1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right), \tag{487}$$

$$\xi(1 - p_B) = 1 - p_B - \left(1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right). \tag{488}$$

Finally, condition (468) is satisfied whenever α satisfies

$$1 - \left(1 - p_B - \left(1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right)\right) < p_A - \left(1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right) \tag{489}$$

$$\iff$$

$$p_B + 1 - \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ < p_A - 1 + \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{490}$$

$$\iff$$

$$2 - p_A + p_B < 2 \cdot \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{491}$$

$$\iff$$

$$1 - \left(\frac{p_A - p_B}{2}\right) < \prod_{i=1}^{m} \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{492}$$

what concludes the proof. □

D.1.4 *Laplacian Smoothing*

Corollary 11. *Suppose* $\mathcal{Z} = \mathbb{R}$ *and* $\varepsilon_0 \sim \mathcal{L}(0, b)$ *follows a Laplace distribution with mean* 0 *and scale parameter* $b > 0$. *Let* $\varepsilon_1 := \alpha + \varepsilon_0$ *for* $\alpha \in \mathbb{R}$. *Suppose that the* ε_0*-smoothed*

classifier g *is* (p_A, p_B)*-confident at* $x \in \mathcal{X}$ *for some* $y_A \in \mathcal{Y}$*. Then, it holds that* $q(y_A|x;\varepsilon_1) > \max_{y\neq y_A} q(y|x;\varepsilon_1)$ *if* α *satisfies*

$$|\alpha| < \begin{cases} -b\cdot\log\left(4\,p_B\,(1-p_A)\right) & (p_A = \frac{1}{2} \wedge\ p_B < \frac{1}{2}) \\ & \vee\ (p_A > \frac{1}{2} \wedge p_B = \frac{1}{2}), \\ -b\cdot\log\left(1-p_A+p_B\right) & p_A > \frac{1}{2} \wedge p_B < \frac{1}{2}. \end{cases} \tag{493}$$

Proof. By Theorem 5 we know that if ε_1 satisfies

$$\xi(p_A) + \xi(1-p_B) > 1, \tag{494}$$

then it is guaranteed that

$$q(y_A|x;\varepsilon_1) > \max_{y\neq y_A} q(y|x;\varepsilon_1). \tag{495}$$

The proof is thus complete if we show that (494) reduces to (493). For that purpose denote by f_0 and f_1 density functions of ε_0 and ε_1, respectively, and consider

$$f_0(z) = \frac{1}{2b}\exp\left(-\frac{|z|}{b}\right), \quad f_1(z) = \frac{1}{2b}\exp\left(-\frac{|z-\alpha|}{b}\right). \tag{496}$$

Due to symmetry, assume without loss of generality that $\alpha \geqslant 0$. Then for $z \in \mathbb{R}$

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} = \exp\left(-\frac{|z-\alpha|-|z|}{b}\right) \tag{497}$$

$$= \begin{cases} \exp\left(-\frac{\alpha}{b}\right) & z < 0, \\ \exp\left(\frac{2z-\alpha}{b}\right) & 0 \leqslant z < \alpha, \\ \exp\left(\frac{\alpha}{b}\right) & z \geqslant \alpha. \end{cases} \tag{498}$$

Note that the CDFs for ε_0 and ε_1 are given by

$$F_0(z) = \begin{cases} \frac{1}{2}\exp\left(\frac{z}{b}\right) & z \leqslant 0, \\ 1-\frac{1}{2}\exp\left(-\frac{z}{b}\right) & z > 0, \end{cases} \tag{499}$$

$$F_1(z) = \begin{cases} \frac{1}{2}\exp\left(\frac{z-\alpha}{b}\right) & z \leqslant \alpha, \\ 1-\frac{1}{2}\exp\left(-\frac{z-\alpha}{b}\right) & z > \alpha. \end{cases} \tag{500}$$

Note that for $\exp\left(-\frac{\alpha}{b}\right) \leqslant t < \exp\left(\frac{\alpha}{b}\right)$ we have

$$\mathbb{P}_0\left(\exp\left(\frac{2\varepsilon_0-\alpha}{b}\right) \leqslant t \wedge 0 \leqslant \varepsilon_0 < \alpha\right)$$

$$= \mathbb{P}_0\left(\exp\left(-\frac{\alpha}{b}\right) \leqslant \exp\left(\frac{2\varepsilon_0-\alpha}{b}\right) \leqslant t\right) \tag{501}$$

$$= \mathbb{P}_0\left(0 \leqslant \varepsilon_0 \leqslant \frac{b\log(t)+\alpha}{2}\right) \tag{502}$$

$$= F_0\left(\frac{b\log(t)+\alpha}{2}\right) - F_0(0) \tag{503}$$

$$= \frac{1}{2} - \frac{1}{2}\exp\left(-\frac{1}{b}\left(\frac{b\log(t)+\alpha}{2}\right)\right) \tag{504}$$

$$= \frac{1}{2} - \frac{1}{2\sqrt{t}}\exp\left(-\frac{\alpha}{2b}\right). \tag{505}$$

Computing ζ yields

$$\begin{aligned}\zeta(t) &= \mathbb{P}\left(\Lambda(\varepsilon_0) \leqslant t\right) && (506)\\ &= \mathbb{P}\left(\exp\left(-\frac{\alpha}{b}\right) \leqslant t \wedge \varepsilon_0 < 0\right) \\ &\quad + \mathbb{P}\left(\exp\left(\frac{\alpha}{b}\right) \leqslant t \wedge \varepsilon_0 \geqslant \alpha\right) && (507)\\ &\quad + \mathbb{P}\left(\exp\left(\frac{2\varepsilon_0 - \alpha}{b}\right) \leqslant t \wedge 0 \leqslant \varepsilon_0 < \alpha\right) \\ &= \begin{cases} 0 & t < \exp\left(-\frac{\alpha}{b}\right), \\ 1 - \frac{1}{2\sqrt{t}} \exp\left(-\frac{\alpha}{2b}\right) & \exp\left(-\frac{\alpha}{b}\right) \leqslant t < \exp\left(\frac{\alpha}{b}\right), \\ 1 & t \geqslant \exp\left(\frac{\alpha}{b}\right). \end{cases} && (508)\end{aligned}$$

The inverse is then given by

$$\zeta^{-1}(p) = \begin{cases} 0 & p < \frac{1}{2}, \\ \frac{1}{4(1-p)^2} \exp\left(-\frac{\alpha}{b}\right) & \frac{1}{2} \leqslant p < 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}), \\ \exp\left(\frac{\alpha}{b}\right) & p \geqslant 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}). \end{cases} \tag{509}$$

In order to evaluate ξ, we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R} \colon \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R} \colon \Lambda(z) \leqslant t\}$ and consider

$$\underline{S}_{\zeta^{-1}(p)} = \begin{cases} \varnothing & p \leqslant \frac{1}{2}, \\ \left(-\infty, b \cdot \log\left(\frac{1}{2(1-p)}\right)\right) & \frac{1}{2} < p < 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right), \\ (-\infty, \alpha], & p \geqslant 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right) \end{cases} \tag{510}$$

and

$$\overline{S}_{\zeta^{-1}(p)} = \begin{cases} \varnothing & p < \frac{1}{2}, \\ \left(-\infty, b \cdot \log\left(\frac{1}{2(1-p)}\right)\right] & \frac{1}{2} \leqslant p < 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right), \\ \mathbb{R} & p \geqslant 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right). \end{cases} \tag{511}$$

Suppose $p < 1/2$. Then $\underline{S}_{\zeta^{-1}(p)} = \overline{S}_{\zeta^{-1}(p)} = \varnothing$ and hence $\xi(p) = 0$ and condition (494) cannot be satisfied. If $p = 1/2$, then $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = (-\infty, 0]$. Note that for $z \leqslant 0$ we have $f_1(z) = f_0(z)\exp(-\alpha/b)$ and hence for any $S \subseteq \overline{S}_{\zeta^{-1}(1/2)}$ we have $\mathbb{P}_1(S) = \exp(-\alpha/b) \cdot \mathbb{P}_0(S)$. We can thus compute ξ at $1/2$ as

$$\begin{aligned} &p = \frac{1}{2} \\ &\Rightarrow \xi(1/2) = \sup\left\{\mathbb{P}_1(S) \colon S \subseteq (-\infty, 0], \mathbb{P}_0(S) \leqslant \frac{1}{2}\right\} = \frac{1}{2}. \end{aligned} \tag{512}$$

Now suppose $1/2 < p < 1 - 1/2\exp(-\alpha/b)$. In this case, $\underline{S}_{\zeta^{-1}(p)} = (-\infty, b \cdot \log(1/2(1-p)))$ and $\overline{S}_{\zeta^{-1}(p)} = (-\infty, b \cdot \log(1/2(1-p))]$. Since the singleton $\{b \cdot \log(1/2(1-p))\}$ has no prob-

ability mass under both $\mathbb{P}_0$ and $\mathbb{P}_1$, the function ξ is straight forward to compute: if $\frac{1}{2} < p < 1 - \frac{1}{2}\exp(-\frac{\alpha}{b})$, then

$$\xi(p) = \mathbb{P}\left(\varepsilon_1 \leqslant b \cdot \log\left(\frac{1}{2(1-p)}\right)\right) \tag{513}$$

$$= \frac{1}{2}\exp\left(\frac{b \cdot \log\left(\frac{1}{2(1-p)}\right) - \alpha}{b}\right) \tag{514}$$

$$= \frac{1}{4(1-p)}\exp\left(-\frac{\alpha}{b}\right). \tag{515}$$

Finally, consider the case where $p \geqslant 1 - 1/2\exp(-\alpha/b)$. Then $\underline{S}_{\zeta^{-1}(p)} = (-\infty, \alpha]$ and $\overline{S}_{\zeta^{-1}(p)} = \mathbb{R}$. Any $(-\infty, \alpha] \subseteq S \subseteq \mathbb{R}$ can then be written as $S = (-\infty, \alpha] \dot\cup T$ for some $T \subseteq (\alpha, \infty)$. Hence

$$\mathbb{P}_1(S) = \mathbb{P}(\varepsilon_1 \leqslant \alpha) + \mathbb{P}_1(T) = \frac{1}{2} + \exp\left(\frac{\alpha}{b}\right)\mathbb{P}_0(T), \tag{516}$$

$$\mathbb{P}_0(S) = \mathbb{P}(\varepsilon_0 \leqslant \alpha) + \mathbb{P}_0(T) = 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}) + \mathbb{P}_0(T). \tag{517}$$

Thus, if $p \geqslant 1 - \frac{1}{2}\exp(-\frac{\alpha}{b})$, then

$$\xi(p) = \sup\{\mathbb{P}_1(S)\colon\ (-\infty, \alpha] \subseteq S \subseteq \mathbb{R},\ \mathbb{P}_0(S) \leqslant p\} \tag{518}$$

$$= \frac{1}{2} + \sup\left\{\mathbb{P}_1(T)\colon\ T \subseteq (\alpha, \infty),\ \mathbb{P}_0(T) \leqslant p - 1 + \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right)\right\} \tag{519}$$

$$= \frac{1}{2} + \exp\left(\frac{\alpha}{b}\right)\left(p - 1 + \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right)\right) \tag{520}$$

$$= 1 - \exp\left(\frac{\alpha}{b}\right)(1-p). \tag{521}$$

In order to evaluate condition (494), consider

$$1 - \xi(1 - p_B) = \begin{cases} 1 & p_B > \frac{1}{2} \\ \frac{1}{2} & p_B = \frac{1}{2} \\ 1 - \frac{1}{4p_B}\exp\left(-\frac{\alpha}{b}\right) & \frac{1}{2} > p_B > \exp\left(-\frac{\alpha}{b}\right) \\ \exp\left(\frac{\alpha}{b}\right)p_B & \exp\left(-\frac{\alpha}{b}\right) \geqslant p_B, \end{cases} \tag{522}$$

$$\xi(p_A) = \begin{cases} 0 & p_A < \frac{1}{2} \\ \frac{1}{2} & p_A = \frac{1}{2} \\ \frac{1}{4(1-p_A)}\exp\left(-\frac{\alpha}{b}\right) & \frac{1}{2} < p_A < 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}) \\ 1 - \exp\left(\frac{\alpha}{b}\right)(1-p_A) & p_A \geqslant 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}). \end{cases} \tag{523}$$

Note that the case $p_B > 1/2$ can be ruled out, since by assumption $p_A \geqslant p_B$. If $p_A = 1/2$, then we need $p_B < 1/2$. Thus, if $p_A = 1/2$, then condition (494) is satisfied if $p_B < 1/2$ and

$$\max\left\{1 - \frac{1}{4p_B}\exp\left(-\frac{\alpha}{b}\right), \exp\left(\frac{\alpha}{b}\right)\cdot p_B\right\} < \frac{1}{2} \tag{524}$$

$$\iff p_B \cdot \exp\left(\frac{\alpha}{b}\right) < \frac{1}{2} \tag{525}$$

$$\iff \alpha < -b \cdot \log\left(2p_B\right). \tag{526}$$

Now consider the case where $p_A > 1/2$. If $p_B = 1/2$, then condition (494) is satisfied if

$$\frac{1}{2} < \min\left\{\frac{1}{4(1-p_A)}\exp\left(-\frac{\alpha}{b}\right), 1 - \exp\left(\frac{\alpha}{b}\right)(1-p_A)\right\} \tag{527}$$

$$\iff \frac{1}{2} < 1 - \exp\left(\frac{\alpha}{b}\right)(1-p_A) \tag{528}$$

$$\iff \alpha < -b \cdot \log\left(2(1-p_A)\right). \tag{529}$$

If on the other hand, $p_A > 1/2$ and $p_B < 1/2$, condition (494) is satisfied if

$$\begin{aligned}&\max\left\{1 - \frac{1}{4p_B}\exp\left(-\frac{\alpha}{b}\right), \exp\left(\frac{\alpha}{b}\right)\cdot p_B\right\} < \\ &\qquad \min\left\{\frac{1}{4(1-p_A)}\exp\left(-\frac{\alpha}{b}\right), 1 - \exp\left(\frac{\alpha}{b}\right)(1-p_A)\right\}\end{aligned} \tag{530}$$

$$\iff$$

$$p_B \cdot \exp\left(\frac{\alpha}{b}\right) < 1 - \exp\left(\frac{\alpha}{b}\right)(1-p_A) \tag{531}$$

$$\iff$$

$$\alpha < -b \cdot \log\left(1 - p_A + p_B\right). \tag{532}$$

Finally, we get that condition (494) is satisfied, if

$$|\alpha| < \begin{cases} -b \cdot \log\left(4\, p_B\,(1-p_A)\right) & (p_A = \frac{1}{2} \wedge\ p_B < \frac{1}{2}) \\ & \vee\ (p_A > \frac{1}{2} \wedge p_B = \frac{1}{2}) \\ -b \cdot \log\left(1 - p_A + p_B\right) & p_A > \frac{1}{2} \wedge p_B < \frac{1}{2} \end{cases} \tag{533}$$

what concludes the proof. □

D.1.5 *Folded Gaussian Smoothing*

Corollary 12. *Suppose* $\mathcal{Z} = \mathbb{R}_{\geqslant 0}$, $\varepsilon_0 \sim |\mathcal{N}(0, \sigma)|$ *and* $\varepsilon_1 := \alpha + \varepsilon_0$ *for some* $\alpha > 0$. *Suppose that the* ε_0*-smoothed classifier* g *is* (p_A, p_B)*-confident at* $x \in \mathcal{X}$ *for some* $y_A \in \mathcal{Y}$. *Then, it holds that* $q(y_A|x;\varepsilon_1) > \max_{y \neq y_A} q(y|x;\varepsilon_1)$ *if* α *satisfies*

$$\alpha < \sigma \cdot \left(\Phi^{-1}\left(\frac{1 + \min\{p_A, 1-p_B\}}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right). \tag{534}$$

Proof. By Theorem 5 we know that if ε_1 satisfies

$$\xi(p_A) + \xi(1-p_B) > 1, \tag{535}$$

then it is guaranteed that

$$q(y_A|x;\varepsilon_1) > \max_{y\neq y_A} q(y|x;\varepsilon_1). \tag{536}$$

The proof is thus complete if we show that (535) reduces to (534). For that purpose denote by f_0 and f_1 density functions of ε_0 and ε_1, respectively, and consider

$$f_0(z) = \begin{cases} \frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right) & z \geqslant 0 \\ 0 & z < 0 \end{cases} \tag{537}$$

$$f_1(z) = \begin{cases} \frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(z-\alpha)^2}{2\sigma^2}\right) & z \geqslant \alpha \\ 0 & z < \alpha. \end{cases} \tag{538}$$

Then, for $z \geqslant 0$,

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} = \begin{cases} 0 & z < \alpha, \\ \exp\left(\frac{z\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2}\right) & z \geqslant \alpha. \end{cases} \tag{539}$$

Let $t_\alpha := \exp\left(\frac{\alpha^2}{2\sigma^2}\right)$ and suppose $t < t_\alpha$. Then

$$\zeta(t) = \mathbb{P}\left(\Lambda(\varepsilon_0) \leqslant t\right] = \mathbb{P}\left(\varepsilon_0 < \alpha\right) \tag{540}$$

$$= \int_0^\alpha \frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)\,dz \tag{541}$$

$$= 2\cdot\int_0^{\alpha/\sigma} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{s^2}{2}\right)\,ds = 2\cdot\Phi\left(\frac{\alpha}{\sigma}\right) - 1. \tag{542}$$

If $t \geqslant t_\alpha$, then

$$\zeta(t) = \mathbb{P}\left(\Lambda(\varepsilon_0) \leqslant t\right) \tag{543}$$

$$= \mathbb{P}\left(\frac{\varepsilon_0\,\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} \leqslant \log(t) \wedge \varepsilon_0 \geqslant \alpha\right) + \mathbb{P}\left(\varepsilon_0 < \alpha\right) \tag{544}$$

$$= \mathbb{P}\left(\varepsilon_0 \leqslant \frac{\sigma^2}{\alpha}\log(t) + \frac{1}{2}\alpha\right) \tag{545}$$

$$= 2\cdot\Phi\left(\frac{\sigma}{\alpha}\log(t) + \frac{\alpha}{2\sigma}\right) - 1 \tag{546}$$

and hence

$$\zeta(t) = \begin{cases} 2\cdot\Phi\left(\frac{\alpha}{\sigma}\right) - 1 & t < t_\alpha \\ 2\cdot\Phi\left(\frac{\sigma}{\alpha}\log\left(t\right) + \frac{\alpha}{2\sigma}\right) - 1 & t \geqslant t_\alpha. \end{cases} \tag{547}$$

Note that $\zeta(t_\alpha) = 2\cdot\Phi\left(\frac{\alpha}{\sigma}\right) - 1$ and let $p_\alpha := \zeta(t_\alpha)$. Recall that $\zeta^{-1}(p) := \inf\{t \geqslant 0\colon \zeta(t) \geqslant p\}$, which yields

$$\zeta^{-1}(p) = \begin{cases} 0 & p \leqslant p_\alpha \\ \exp\left(\frac{\alpha}{\sigma}\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha^2}{2\sigma^2}\right) & p > p_\alpha. \end{cases} \tag{548}$$

In order to evaluate ξ we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R}_{\geqslant 0} \colon \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R}_{\geqslant 0} \colon \Lambda(z) \leqslant t\}$. Let $S_0 := [0, \alpha)$ and note that if $p \leqslant p_\alpha$, we have $\zeta^{-1}(p) = 0$ and hence $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = S_0$. If, on the other hand $p > p_\alpha$, then

$$\begin{aligned}
\underline{S}_{\zeta^{-1}(p)} &= \left\{z \geqslant 0 \colon \Lambda(z) < \zeta^{-1}(p)\right\} && (549)\\
&= S_0 \cup \left\{z \geqslant \alpha \colon \frac{z\,\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} < \frac{\alpha}{\sigma}\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha^2}{2\sigma^2}\right\} && (550)\\
&= S_0 \cup \left\{z \geqslant \alpha \colon z < \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\} && (551)\\
&= S_0 \cup \left[\alpha, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right) && (552)
\end{aligned}$$

and

$$\begin{aligned}
\overline{S}_{\zeta^{-1}(p)} &= \left\{z \geqslant 0 \colon \Lambda(z) \leqslant \zeta^{-1}(p)\right\} && (553)\\
&= S_0 \cup \left\{z \geqslant \alpha \colon \frac{z\,\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} \leqslant \frac{\alpha}{\sigma}\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha^2}{2\sigma^2}\right\} && (554)\\
&= S_0 \cup \left\{z \geqslant \alpha \colon z \leqslant \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\} && (555)\\
&= S_0 \cup \left[\alpha, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right] && (556)\\
&= \underline{S}_{\zeta^{-1}(p)} \cup \left\{\sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\}. && (557)
\end{aligned}$$

In other words

$$\underline{S}_{\zeta^{-1}(p)} = \begin{cases} \varnothing & p \leqslant p_\alpha, \\ S_0 \cup \left[\alpha, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right) & p > p_\alpha, \end{cases} \tag{558}$$

$$\overline{S}_{\zeta^{-1}(p)} = \begin{cases} S_0 & p \leqslant p_\alpha, \\ S_0 \cup \left[\alpha, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right] & p > p_\alpha. \end{cases} \tag{559}$$

Let $\mathcal{S}_p := \{S \subseteq \mathbb{R}_{\geqslant 0} \colon \underline{S}_{\zeta^{-1}(p)} \subseteq S \subseteq \overline{S}_{\zeta^{-1}(p)}, \mathbb{P}_0(S) \leqslant p\}$ and recall that $\xi(p) = \sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S)$. Note that for $p \leqslant p_\alpha$, we have $\mathcal{S}_p = \{S \subseteq \mathbb{R}_{\geqslant 0} \colon S \subseteq S_0 \wedge \mathbb{P}_0(S) \leqslant p\}$ and for $S \subseteq S_0$, it holds that $\mathbb{P}_1(S) = 0$. Hence

$$p \leqslant p_\alpha \Rightarrow \xi(p) = \sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S) = 0. \tag{560}$$

If $p > p_\alpha$, then

$$\begin{aligned}
\mathcal{S}_p = \Big\{ & S \subseteq \mathbb{R}_{\geqslant 0} \colon S_0 \cup \left[\alpha, \sigma \cdot \Phi^{-1}\left(1+p/2\right)\right) \subseteq S \\
& \subseteq S_0 \cup \left[\alpha, \sigma \cdot \Phi^{-1}\left(1+p/2\right)\right], \wedge \mathbb{P}_0(S) \leqslant p \Big\}.
\end{aligned} \tag{561}$$

Since the singleton $\left\{\sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\}$ has no mass under both $\mathbb{P}_0$ and $\mathbb{P}_1$, we find that if $p > p_\alpha$, then

$$\begin{aligned}\xi(p) &= \mathbb{P}\left(0 \leqslant \varepsilon_1 \leqslant \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right) && (562)\\ &= \mathbb{P}\left(0 \leqslant \varepsilon_0 \leqslant \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right) - \alpha\right) && (563)\\ &= 2 \cdot \Phi\left(\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha}{\sigma}\right) - 1. && (564)\end{aligned}$$

Condition (535) can thus be satisfied only if $p_B < p_A$ and

$$\begin{aligned}2 \cdot \Phi\left(\frac{\alpha}{\sigma}\right) - 1 &< \min\{p_A, 1 - p_B\}\\ &\wedge\, \xi(p_A) + \xi(1 - p_b) > 1\end{aligned} \tag{565}$$

that is equivalent to

$$\begin{aligned}\alpha < \sigma \cdot \Phi^{-1}&\left(\frac{1 + \min\{p_A, 1 - p_B\}}{2}\right)\\ &\wedge\, \Phi\left(\Phi^{-1}\left(\frac{1 + (1 - p_B)}{2}\right) - \frac{\alpha}{\sigma}\right)\\ &\quad + \Phi\left(\Phi^{-1}\left(\frac{1 + p_A}{2}\right) - \frac{\alpha}{\sigma}\right) > \frac{3}{2}.\end{aligned} \tag{566}$$

Thus, the following is a sufficient condition for the two inequalities in (566) and hence (535) to hold

$$\alpha < \sigma \cdot \left(\Phi^{-1}\left(\frac{1 + \min\{p_A, 1 - p_B\}}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right) \tag{567}$$

what completes the proof. □

D.2 PROOFS FOR RESOLVABLE TRANSFORMATIONS

Here, we state the proofs and technical details concerning our results for resolvable transformations. Let us start by recalling the definition of resolvable transformations.

Definition 3 (restated). *A transformation $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ is called resolvable if for any $\alpha \in \mathcal{Z}$ there exists a resolving function $\gamma_\alpha\colon \mathcal{Z} \to \mathcal{Z}$ that is injective, continuously differentiable, has non-vanishing Jacobian and for which*

$$\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma_\alpha(\beta)) \quad x \in \mathcal{X},\ \beta \in \mathcal{Z}. \tag{568}$$

Furthermore, we say that ϕ is additive, if $\gamma_\alpha(\beta) = \alpha + \beta$.

D.2.1 *Proof of Corollary 3*

Corollary 3 (restated). *Suppose that the transformation ϕ in Theorem 5 is resolvable with resolving function γ_α. Let $\alpha \in \mathcal{Z}$ and set $\varepsilon_1 := \gamma_\alpha(\varepsilon_0)$ in the definition of the functions ζ and ξ. Then, if α satisfies condition (125), it is guaranteed that $g(\phi(x, \alpha); \varepsilon_0) = g(x; \varepsilon_0)$.*

Proof. Since ϕ is a resolvable transformation, by definition γ_α is injective, continuously differentiable and has non-vanishing Jacobian. By Jacobi's transformation formula (see e.g., [105]), it follows that the density of ε_1 vanishes outside the image of γ_α and is elsewhere given by

$$f_1(z) = f_0(\gamma_\alpha^{-1}(z))|\det(J_{\gamma_\alpha^{-1}(z)})| \quad \text{for any } z \in \mathrm{Im}(\gamma_\alpha) \tag{569}$$

where $J_{\gamma_\alpha^{-1}(z)}$ is the Jacobian of $\gamma_\alpha^{-1}(z)$. Since f_1 is paramterized by α, it follows from Theorem 5 that if α satisfies (125) it is guaranteed that $\arg\max_y q(y|x, \varepsilon_1) = \arg\max_y q(y|x, \varepsilon_0)$. The statement of the corollary then follows immediately from the observation that for any $y \in \mathcal{Y}$ we have

$$\begin{aligned} q(y|x; \varepsilon_1) &= \mathbb{E}(p(y|\phi(x, \varepsilon_1))) \\ &= \mathbb{E}(p(y|\phi(x, \gamma_\alpha(\varepsilon_0)))) \\ &= \mathbb{E}(p(y|\phi(\phi(x, \alpha), \varepsilon_0))) \\ &= q(y|\phi(x, \alpha); \varepsilon_0). \end{aligned} \tag{570}$$

□

D.2.2 *Proofs for Gaussian Blur*

Recall that the Gaussian blur transformation is given by a convolution with a Gaussian kernel

$$G_\alpha(k) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{k^2}{2\alpha}\right) \tag{571}$$

where $\alpha > 0$ is the squared kernel radius. Here we show that the transformation $x \mapsto \phi_B(x) := x * G$ is additive.

Lemma 5 (restated). *The Gaussian blur transformation is additive, i.e., for any $\alpha, \beta \geqslant 0$, we have $\phi_B(\phi_B(x, \alpha), \beta) = \phi_B(x, \alpha + \beta)$.*

Proof. Note that associativity of the convolution operator implies that

$$\begin{aligned} \phi_B(\phi_B(x, \alpha), \beta) &= (\phi_B(x, \alpha) * G_\beta) \\ &= ((x * G_\alpha) * G_\beta) \\ &= (x * (G_\alpha * G_\beta)). \end{aligned} \tag{572}$$

The claim thus follows, if we can show that $(G_\alpha * G_\beta) = G_{\alpha+\beta}$. Let $\mathcal{F}$ denote the Fourier transformation and $\mathcal{F}^{-1}$ the inverse Fourier transformation and note that by the convolution theorem $(G_\alpha * G_\beta) = \mathcal{F}^{-1}\{\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta)\}$. Therefore we have to show that $\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta) = \mathcal{F}(G_{\alpha+\beta})$. For that purpose, consider

$$\mathcal{F}(G_\alpha)(\omega) = \int_{-\infty}^{\infty} G_\alpha(y) \exp(-2\pi i \omega y)\, dy \tag{573}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{y^2}{2\alpha}\right) \exp\left(-2\pi i \omega y\right)\, dy \tag{574}$$

$$= \frac{1}{\sqrt{2\pi\alpha}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\alpha}\right) \left(\cos\left(2\pi\omega y\right) + i \sin\left(2\pi\omega y\right)\right)\, dy \tag{575}$$

$$\overset{(i)}{=} \frac{1}{\sqrt{2\pi\alpha}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\alpha}\right) \cos\left(2\pi\omega y\right)\, dy \tag{576}$$

$$\overset{(ii)}{=} \exp\left(-\omega^2\pi^2 2\alpha\right), \tag{577}$$

where (i) follows from the fact that the second term is an integral of an odd function over a symmetric range and (ii) follows from $\int_{-\infty}^{\infty} \exp\left(-a\,y^2\right) \cos\left(2\pi\omega y\right) \mathrm{d}y = \sqrt{\frac{\pi}{a}} \exp(\frac{-(\pi\omega)^2}{a})$ with $a = \frac{1}{2\alpha}$ (see p. 302, eq. 7.4.6 in [2]). This concludes our proof since

$$\begin{aligned}(\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta))(\omega) &= \exp\left(-\omega^2\pi^2 2\alpha\right) \cdot \exp\left(-\omega^2\pi^2 2\beta\right) && (578)\\ &= \exp\left(-\omega^2\pi^2 2(\alpha+\beta)\right) && (579)\\ &= \mathcal{F}(G_{\alpha+\beta})(\omega) && (580)\end{aligned}$$

and hence

$$\begin{aligned}(G_\alpha * G_\beta) &= \mathcal{F}^{-1}\{\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta)\} && (581)\\ &= \mathcal{F}^{-1}\{\mathcal{F}(G_{\alpha+\beta})\} && (582)\\ &= G_{\alpha+\beta}. && (583)\end{aligned}$$

□

Remark 2. *We notice that the preceding theorem naturally extends to higher dimensional Gaussian kernels of the form*

$$G_\alpha(k) = \frac{1}{(2\pi\alpha)^{\frac{m}{2}}} \exp\left(-\frac{\|k\|^2}{2\alpha}\right), \qquad k \in \mathbb{R}^m. \tag{584}$$

Consider

$$\begin{aligned}\mathcal{F}(G_\alpha)(\omega) &= \int_{\mathbb{R}^m} G_\alpha(y) \exp\left(-2\pi i\langle\omega, y\rangle\right) \mathrm{d}y && (585)\\ &= \frac{1}{(2\pi\alpha)^{\frac{m}{2}}} \int_{\mathbb{R}^m} \exp\left(-\frac{\|y\|_2^2}{2\alpha} - 2\pi i\langle\omega, y\rangle\right) \mathrm{d}y && (586)\\ &= \prod_{j=1}^{m} \left(\frac{1}{\sqrt{2\pi\alpha}} \int_{\mathbb{R}} \exp\left(-\frac{y_j^2}{2\alpha} - 2\pi i\omega_j y_j\right) \mathrm{d}y_j\right) && (587)\\ &= \exp\left(-\|\omega\|_2^2\,\pi^2 2\alpha\right) && (588)\end{aligned}$$

that leads to $(G_\alpha * G_\beta) = G_{\alpha+\beta}$*, and hence additivity.*

D.2.3 *Proofs for Brightness and Contrast*

Recall that the brightness and contrast transformation is defined as

$$\phi_{BC} \colon \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}, \quad (x, \alpha) \mapsto e^{\alpha_1}(x + \alpha_2). \tag{589}$$

Lemma 6 (restated). *Let* $x \in \mathcal{X}$, $k \in \mathbb{R}$, *and suppose that*

$$\varepsilon_0 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, \tau^2)) \quad \text{and} \quad \varepsilon_1 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, e^{-2k}\tau^2)). \tag{590}$$

Suppose that $q(y|x; \varepsilon_0) \geqslant p$ *for some* $p \in [0, 1]$ *and* $y \in \mathcal{Y}$*. Let* Φ *be the cumulative density function of the standard Gaussian. Then*

$$q(y|x; \varepsilon_1) \geqslant \begin{cases} 2\Phi\left(e^k \Phi^{-1}\left(\frac{1+p}{2}\right)\right) - 1 & k \leqslant 0 \\ 2\left(1 - \Phi\left(e^k \Phi^{-1}(1 - \frac{p}{2})\right)\right) & k > 0. \end{cases} \tag{591}$$

Proof. Note that $\varepsilon_0 \sim \mathcal{N}(0, \Sigma)$ and $\varepsilon_1 = A\,\varepsilon_0 \sim \mathcal{N}(0, A^2\,\Sigma)$ where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & e^{-k} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \tag{592}$$

and denote by f_0 and f_1 the probability density functions of ε_0 and ε_1, respectively, and denote by $\mathbb{P}_0$ and $\mathbb{P}_1$ the corresponding probability measures. Recall the Definition of Lower Level sets (Definition 8: for $t \geqslant 0$, (strict) lower level sets are defined as

$$\begin{aligned} \underline{S_t} &:= \{z \in \mathcal{Z}\colon\ \Lambda(z) < t\}, \quad \overline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) \leqslant t\}, \\ &\text{where} \quad \Lambda(z) := \frac{f_1(z)}{f_0(z)}. \end{aligned} \tag{593}$$

Furthermore, recall that the function ζ is given by

$$t \mapsto \zeta(t) := \mathbb{P}_0\left(\overline{S}_t\right) \tag{594}$$

where $\mathbb{P}_0$ is the distribution of ε_0 and note that the generalized inverse of ζ corresponds to τ_p, i.e.,

$$\zeta^{-1}(p) = \inf\{t \geqslant 0 | \zeta(t) \geqslant p\} = \tau_p \tag{595}$$

and the function ξ is correspondingly given by

$$\xi(p) = \sup\{\mathbb{P}_1(S) | \underline{S}(\zeta^{-1}(p)) \subseteq S \subseteq \overline{S}(\zeta^{-1}(p))\}. \tag{596}$$

By assumption we know that $\mathbb{E}(p(y|\phi(x, \varepsilon_0))) = q(y|x; \varepsilon_o) \geqslant p$. Note that by Lemma 12, for any $p \in [0, 1]$ we have that

$$\mathbb{P}_0(\underline{S}_{\zeta^{-1}(p)}) \leqslant p. \tag{597}$$

Let $S \subseteq \mathcal{Z}$ be such that $\underline{S}_{\zeta^{-1}(p)} \subseteq S \subseteq \overline{S}_{\zeta^{-1}(p)}$ and $\mathbb{P}_0(S) \leqslant p$. Then, from part (i) of Lemma 9, it follows that $\mathbb{E}(p(y|\phi(x, \varepsilon_1))) = q(y|x; \varepsilon_1) \geqslant \mathbb{P}_1(S)$. Note that

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} \tag{598}$$

$$= \frac{\left((2\pi)^2|A^2\,\Sigma|\right)^{-\frac{1}{2}} \exp(-\frac{1}{2}(z^T(A^2\Sigma)^{-1}z))}{((2\pi)^2|\Sigma|)^{-\frac{1}{2}} \exp(-\frac{1}{2}(z^T(\Sigma)^{-1}z))} \tag{599}$$

$$= \frac{1}{|A|} \exp\left(-\frac{1}{2} z^T\left((A^2\,\Sigma)^{-1} - \Sigma^{-1}\right) z\right) \tag{600}$$

$$= \exp\left(k - \frac{z_2^2}{2\tau^2}\left(e^{2k} - 1\right)\right). \tag{601}$$

Note that, if $k = 0$, then $\varepsilon_1 = \varepsilon_0$ and hence the statement holds in this case. Suppose that $k > 0$ and consider

$$\zeta(t) = \mathbb{P}_0\left(\overline{S}_t\right) = \mathbb{P}\left(\exp\left(k - \frac{\varepsilon_{0,2}^2}{2\tau^2}\left(e^{2k} - 1\right)\right) \leqslant t\right) \tag{602}$$

$$= 1 - \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau}\right)^2 \leqslant 2 \cdot \frac{k - \log(t)}{e^{2k} - 1}\right) \tag{603}$$

$$= 1 - F_{\chi^2}\left(2 \cdot \frac{k - \log(t)}{e^{2k} - 1}\right) \tag{604}$$

$$= \begin{cases} 0 & t = 0, \\ 1 - F_{\chi^2}\left(2 \cdot \frac{k - \log(t)}{e^{2k} - 1}\right) & 0 < t < e^k, \\ 1 & t \geqslant e^k, \end{cases} \tag{605}$$

where F_{χ^2} denotes the CDF of the χ^2-distribution with one degree of freedom. Note that for any $t \geqslant 0$ we have that $\mathbb{P}_0(\overline{S}_t) = \mathbb{P}_0(\underline{S}_t)$ and thus the inverse $\zeta^{-1}(p) = \inf\{t \geqslant 0 \colon \zeta(t) \geqslant p\}$ is given by

$$\zeta^{-1}(p) = \begin{cases} 0 & p = 0 \\ \exp\left(k - F_{\chi^2}^{-1}(1-p) \cdot \frac{e^{2k}-1}{2}\right) & 0 < p < 1 \\ e^k & p = 1. \end{cases} \tag{606}$$

Thus, for any $p \in [0, 1]$, we find that

$$\mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)}) = \mathbb{P}_0(\underline{S}_{\zeta^{-1}(p)}) = \zeta(\zeta^{-1}(p)) = p \tag{607}$$

and

$$\mathbb{E}_0(p(y|\,\phi(x, \varepsilon_0))) = q(y|\,x; \varepsilon_0) \geqslant p = \mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)}). \tag{608}$$

Part (i) of Lemma 9 implies that $q(y|\,x; \varepsilon_1) \geqslant \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$. Computing $\mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$ yields

$$q(y|\,x; \varepsilon_1) \geqslant \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)}) \tag{609}$$

$$= 1 - \mathbb{P}\left(\left(\frac{\varepsilon_{1,2}}{\tau^2}\right)^2 \leqslant (k - \log(\zeta^{-1}(p)))\frac{2}{e^{2k} - 1}\right) \tag{610}$$

$$= 1 - \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau^2}\right)^2 \leqslant (k - \log(\zeta^{-1}(p)))\frac{2e^{2k}}{e^{2k} - 1}\right) \tag{611}$$

$$= 1 - F_{\chi^2}\left((k - \log(\zeta^{-1}(p)))\frac{2e^{2k}}{e^{2k} - 1}\right) \tag{612}$$

$$= 1 - F_{\chi^2}\left(\left(k - \left(k - \frac{e^{2k} - 1}{2}F_{\chi^2}^{-1}(1-p)\right)\right)\frac{2e^{2k}}{e^{2k} - 1}\right) \tag{613}$$

$$= 1 - F_{\chi^2}\left(e^{2k}F_{\chi^2}^{-1}(1-p)\right). \tag{614}$$

If, on the other hand, $k < 0$, then

$$\begin{aligned}
\zeta(t) &= \mathbb{P}_0\left(\overline{S}_t\right) && (615)\\
&= \mathbb{P}\left(\exp\left(k + \frac{\varepsilon_{0,2}^2}{2\tau^2}\left|e^{2k} - 1\right|\right) \leqslant t\right) && (616)\\
&= \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau}\right)^2 \leqslant 2 \cdot \frac{\log(t) - k}{|e^{2k} - 1|}\right) && (617)\\
&= F_{\chi^2}\left(2 \cdot \frac{\log(t) - k}{|e^{2k} - 1|}\right) && (618)\\
&= \begin{cases} 0 & t \leqslant e^k, \\ F_{\chi^2}\left(2 \cdot \frac{\log(t)-k}{|e^{2k}-1|}\right) & t > e^k. \end{cases} && (619)
\end{aligned}$$

A similar computation as in the case where $k > 0$ leads to an expression for the inverse $\zeta^{-1}(p) = \inf\{t \geqslant 0 | \, \zeta(t) \geqslant p\}$

$$\zeta^{-1}(p) = \begin{cases} 0 & p = 0, \\ \exp\left(k + F_{\chi^2}^{-1}(p) \cdot \frac{|e^{2k}-1|}{2}\right) & p > 0. \end{cases} \tag{620}$$

Thus, for any $p \in [0, 1]$, we find that

$$\mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)}) = \mathbb{P}_0(\underline{S}_{\zeta^{-1}(p)}) = \zeta(\zeta^{-1}(p)) = p \tag{621}$$

and

$$\mathbb{E}(p(y | \, \phi(x, \varepsilon_0))) = q(y | \, x; \varepsilon_0) \geqslant p = \mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)}). \tag{622}$$

Part (i) of Lemma 9 implies that $g_c^{\varepsilon_1}(x) \geqslant \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$. Computing $\mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$ yields

$$\begin{aligned}
q(y | \, x; \varepsilon_1) &\geqslant \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)}) && (623)\\
&= \mathbb{P}\left(\left(\frac{\varepsilon_{1,2}}{\tau}\right)^2 \leqslant 2 \cdot \frac{\log(\zeta^{-1}(p)) - k}{|e^{2k} - 1|}\right) && (624)\\
&= \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau}\right)^2 \leqslant 2e^{2k} \cdot \frac{\log(\zeta^{-1}(p)) - k}{|e^{2k} - 1|}\right) && (625)\\
&= F_{\chi^2}\left(\left(\left(k + F_{\chi^2}^{-1}(p)\frac{|e^{2k} - 1|}{2}\right) - k\right)\frac{2\,e^{2k}}{|e^{2k} - 1|}\right) && (626)\\
&= F_{\chi^2}\left(e^{2k} F_{\chi^2}^{-1}(p)\right). && (627)
\end{aligned}$$

Finally, note the following relation between the $\chi^2(1)$ and the standard normal distribution. Let $Z \sim \mathcal{N}(0, 1)$ and denote by Φ the CDF of Z. Then, for any $z \geqslant 0$, $F_{\chi^2}(z) = \mathbb{P}(Z^2 \leqslant z) = \mathbb{P}(-\sqrt{z} \leqslant Z \leqslant \sqrt{z}) = \Phi(\sqrt{z}) - \Phi(-\sqrt{z}) = 2\Phi(\sqrt{z}) - 1$ and the inverse is thus given by $F_{\chi^2}^{-1}(p) = (\Phi^{-1}(\frac{1+p}{2}))^2$. It follows that

$$q(y|, \, x; \varepsilon_1) \geqslant \begin{cases} 2\Phi\left(e^k \Phi^{-1}\left(\frac{1+p}{2}\right)\right) - 1 & k \leqslant 0, \\ 2\left(1 - \Phi\left(e^k \Phi^{-1}(1 - \frac{p}{2})\right)\right) & k > 0, \end{cases} \tag{628}$$

what concludes the proof. □

The following Lemma establishes another useful property of the distribution of ε_1.

Lemma 20. *Let* $\varepsilon_0 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, \tau^2))$, $\alpha = (k, b)^T \in \mathbb{R}^2$ *and* $\varepsilon_1 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, e^{-2k}\tau^2))$. *Then, for all* $x \in \mathcal{X}$, *it holds that* $g(\phi_{BC}(x, \alpha); \varepsilon_0) = g(x; \alpha + \varepsilon_1)$.

Proof. Let $x \in \mathcal{X}$, and write $\varepsilon_i = (\varepsilon_{i,1}, \varepsilon_{i,2})^T$ for $i = 0, 1$. Note that

$$\phi_{BC}(\phi_{BC}(x, \alpha), \varepsilon_0) = e^{\varepsilon_{0,1}} (\phi_{BC}(x, \alpha) + \varepsilon_{0,2}) = e^{\varepsilon_{0,1}} \left(e^k (x + b) + \varepsilon_{0,2}\right) \tag{629}$$

$$= e^{\varepsilon_{0,1}+k} \left(x + \left(b + e^{-k}\varepsilon_{0,2}\right)\right) = \phi_{BC}(x, \alpha + \tilde{\varepsilon}_0) \tag{630}$$

where $\tilde{\varepsilon}_0 = (\varepsilon_{0,1}, e^{-k}\varepsilon_{0,2})^T$. Note that $\tilde{\varepsilon}_0$ follows a Gaussian distribution since

$$\tilde{\varepsilon}_0 = A \cdot \varepsilon_0, \quad A = \begin{pmatrix} 1 & 0 \\ 0 & e^{-k} \end{pmatrix} \tag{631}$$

and hence $\mathbb{E}(\tilde{\varepsilon}_0) = A \cdot \mathbb{E}(\varepsilon_0) = 0$ and

$$\operatorname{Cov}(\tilde{\varepsilon}_0) = \mathbb{E}\left(\varepsilon_0 A A^T \varepsilon_0^T\right) = A^2 \cdot \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & e^{-2k}\tau^2 \end{pmatrix}. \tag{632}$$

The choice $\varepsilon_1 \equiv \tilde{\varepsilon}_0 \sim \mathcal{N}(0, \operatorname{diag}(\sigma_1^2, e^{-2k}\sigma_2^2))$ shows that for any $y \in \mathcal{Y}$

$$q(y|\, \phi_{BC}(x, \alpha); \varepsilon_0) = \mathbb{E}(p(y|\, \phi(\phi(x, \alpha), \varepsilon_0)) \tag{633}$$

$$= \mathbb{E}(p(y|\, \phi(x, \alpha + \varepsilon_1)) \tag{634}$$

$$= q(y|\, x; \alpha + \varepsilon_1) \tag{635}$$

what concludes the proof. □

These observations, together with the Gaussian robustness bound from Corollary 8 allow us to prove Lemma 7.

Lemma 7 (restated). *Let* ε_0 *and* ε_1 *be as in Lemma 6 and suppose that*

$$q(y_A|\, x; \varepsilon_1) \geqslant \tilde{p}_A > \tilde{p}_B \geqslant \max_{y \neq y_A} q(y|\, x; \varepsilon_1). \tag{636}$$

Then it is guaranteed that $y_A = g(\phi_{BC}(x, \alpha); \varepsilon_0)$ *as long as* $\alpha = (k, b)^T$ *satisfies*

$$\sqrt{\left(\frac{k}{\sigma}\right)^2 + \left(\frac{b}{e^{-k}\tau}\right)^2} < \frac{1}{2}\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(\tilde{p}_B)\right). \tag{637}$$

Proof. Since $\varepsilon_1 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, e^{-2k}\tau^2))$, it follows from Corollary 8 that whenever $\alpha = (k, b)^T$ satisfies

$$\sqrt{\left(\frac{k}{\sigma}\right)^2 + \left(\frac{b}{e^{-k}\tau}\right)^2} < \frac{1}{2}\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(\tilde{p}_B)\right), \tag{638}$$

then it is guaranteed that $y_A = g(x; \varepsilon_1)$. The statement now directly follows from Lemma 20. □

D.2.4 *Composition of Gaussian Blur, Brightness, Contrast, and Translation*

D.2.4.1 *Certification Strategy*

The certification generally follows the same procedure as in certifying brightness and contrast. In the following, we first provide a formal definition of this transformation composition. Specifically, the transformation ϕ_{BTBC} is defined as:

$$\phi_{BTBC}(x, \alpha) := \phi_B(\phi_T(\phi_{BC}(x, \alpha_k, \alpha_b), \alpha_{Tx}, \alpha_{Ty}), \alpha_B), \tag{639}$$

where ϕ_B, ϕ_T and ϕ_{BC} are Gaussian blur, translation, and brightness and contrast transformations respectively as defined before; $\alpha := (\alpha_k, \alpha_b, \alpha_{Tx}, \alpha_{Ty}, \alpha_B)^T \in \mathbb{R}^4 \times \mathbb{R}_{\geqslant 0}$ is the transformation parameter.

Our certification relies on the following key results:

Corollary 13. *Let $x \in \mathcal{X}$, $k \in \mathbb{R}$ and let $\epsilon_0 := (\epsilon_0^a, \epsilon_0^b)^T$ be a random variable defined as*

$$\epsilon_0^a \sim \mathcal{N}(0, \mathrm{diag}(\sigma_k^2, \sigma_b^2, \sigma_T^2, \sigma_T^2)) \text{ and } \epsilon_0^b \sim \mathrm{Exp}(\lambda_B). \tag{640}$$

Similarly, let $\epsilon_1 := (\epsilon_1^a, \epsilon_1^b)$ be a random variable with

$$\epsilon_1^a \sim \mathcal{N}(0, \mathrm{diag}(\sigma_k^2, e^{-2k}\sigma_b^2, \sigma_T^2, \sigma_T^2)) \text{ and } \epsilon_1^b \sim \mathrm{Exp}(\lambda_B). \tag{641}$$

For either random variable (denoted as ϵ), recall that $q(y|x; \epsilon) := \mathbb{E}(p(y|\phi_{BTBC}(x, \epsilon)))$. Suppose that $q(y|x; \epsilon_0) \geqslant p$ for some $p \in [0, 1]$ and $y \in \mathcal{Y}$. Then $q(y|x; \epsilon_1)$ satisfies (131).

Lemma 21. *Let ϵ_0 and ϵ_1 be as in 13 and suppose that*

$$q(y_A|x; \varepsilon_1) \geqslant \tilde{p}_A > \tilde{p}_B \geqslant \max_{y \neq y_A} q(y|x; \varepsilon_1). \tag{642}$$

Then it is guaranteed that $y_A = g(\phi_{BTBC}(x, \alpha); \varepsilon_0)$ as long as $p'_A > p'_B$, where

$$p'_A = \begin{cases} 0, & \text{if } \tilde{p}_A \leqslant 1 - \exp(-\lambda_B \alpha_B), \\ \Phi\left(\Phi^{-1}\left(1 - (1 - \tilde{p}_A)\exp(\lambda_B \alpha_B)\right)\right. & \\ \quad \left. -\sqrt{\alpha_k^2/\sigma_k^2 + \alpha_b^2/(e^{-2\alpha_k}\sigma_b^2) + (\alpha_{Tx}^2 + \alpha_{Ty}^2)/\sigma_T^2}\right) & \text{otherwise} \end{cases} \tag{643}$$

and

$$p'_B = \begin{cases} 1, & \text{if } \tilde{p}_B \geqslant \exp(-\lambda_B \alpha_B), \\ 1 - \Phi\left(\Phi^{-1}\left(1 - \tilde{p}_B \exp(\lambda_B \alpha_B)\right)\right. & \\ \quad \left. -\sqrt{\alpha_k^2/\sigma_k^2 + \alpha_b^2/(e^{-2\alpha_k}\sigma_b^2) + (\alpha_{Tx}^2 + \alpha_{Ty}^2)/\sigma_T^2}\right). & \text{otherwise} \end{cases} \tag{644}$$

The ϵ_0 specified by (640) is the smoothing distribution. Similar as in brightness and contrast certification, we first obtain p_A, a lower bound of $q(y_A|x, \varepsilon_0)$ by Monte-Carlo sampling. For a given transformation parameter $\alpha := (\alpha_k, \alpha_b, \alpha_{Tx}, \alpha_{Ty}, \alpha_B)^T$, we then trigger 13 to get $\tilde{p}_A$, a lower bound of $q(y_A|x, \varepsilon_1)$ and set $\tilde{p}_B = 1 - \tilde{p}_A$. Finally, we use the explicit condition in 21 to obtain the certification. Indeed, with $\tilde{p}_B = 1 - \tilde{p}_A$, 21 can be simplified to the following corollary.

Corollary 14. *Let ϵ_0 and ϵ_1 be as in 13 and suppose that*

$$q(y_A|x;\, \varepsilon_1) \geqslant \tilde{p}_A. \tag{645}$$

Then it is guaranteed that $y_A = g(\phi_{BTBC}(x,\, \alpha);\, \varepsilon_0)$ as long as

$$\tilde{p}_A > 1 - \exp(-\lambda_B \alpha_B)\left(1 - \Phi\left(\sqrt{\frac{\alpha_k^2}{\sigma_k^2} + \frac{\alpha_b^2}{e^{-2\alpha_k}\sigma_b^2} + \frac{\alpha_{Tx}^2 + \alpha_{Ty}^2}{\sigma_T^2}}\right)\right). \tag{646}$$

To certify against a set of transformation parameters

$$\begin{aligned} \mathcal{S}_{adv} = \{&(\alpha_k,\, \alpha_b,\, \alpha_{Tx},\, \alpha_{Ty},\, \alpha_B)^T | \\ &\alpha_k \in [-k_0,\, k_0], \alpha_b \in [-b_0,\, b_0], \\ &\|(\alpha_{Tx}, \alpha_{Ty})\|_2 \leqslant T, \alpha_B \leqslant B_0\}, \end{aligned} \tag{647}$$

we relax the robust condition in (646) to

$$\begin{aligned} \tilde{p}_A > 1 - \exp(-\lambda_B \alpha_B)\Bigg(1 - \\ \Phi\left(\sqrt{\frac{\alpha_k^2}{\sigma_k^2} + \frac{\alpha_b^2}{\min\{e^{-2\alpha_k},\, 1\}\sigma_b^2} + \frac{\alpha_{Tx}^2 + \alpha_{Ty}^2}{\sigma_T^2}}\right)\Bigg). \end{aligned} \tag{648}$$

The LHS of 648 is monotonically decreasing w.r.t. $|\alpha_k|$ and the RHS is monotonically increasing w.r.t. $|\alpha_k|, |\alpha_b|, \|(\alpha_{Tx}, \alpha_{Ty}\|_2$, and $|\alpha_B|$, and the RHS is symmetric w.r.t. α_b and $\|(\alpha_{Tx}, \alpha_{Ty})\|_2$. As a result, we only need to check the condition for $(-k_0,\, b_0,\, T,\, 0,\, B_0)$ and $(k_0,\, b_0,\, T,\, 0,\, B_0)$ to certify the entire set $\mathcal{S}_{adv}$. Throughout the experiments, we use this strategy for certification.

D.2.4.2 *Proofs*

Recall that the composition of Gaussian Blur, with brightness, contrast and translation is defined as

$$\phi_{BTBC}(x, \alpha) := \phi_B(\phi_T(\phi_{BC}(x, \alpha_k, \alpha_b), \alpha_{Tx}, \alpha_{Ty}), \alpha_B), \tag{649}$$

where ϕ_B, ϕ_T and ϕ_{BC} are Gaussian blur, translation, and brightness and contrast transformations respectively as defined before and $\alpha := (\alpha_k, \alpha_b, \alpha_{Tx}, \alpha_{Ty}, \alpha_B)^T \in \mathbb{R}^4 \times \mathbb{R}_{\geqslant 0}$ is the transformation parameter. It is easy to see that this transformation composition satisfies the following properties:

- **(P1)** For arbitrary $\alpha^{(1)}, \alpha^{(2)} \in \mathbb{R}^4 \times \mathbb{R}_{\geqslant 0}$,

$$\phi_{BTBC}(\phi_{BTBC}(x, \alpha^{(1)}), \alpha^{(2)}) = \phi_{BTBC}(x, \alpha) \tag{650}$$

where

$$\begin{aligned} \alpha = \Big(&\alpha_k^{(1)} + \alpha_k^{(2)}, \alpha_b^{(1)} + \alpha_b^{(2)}/e^{\alpha_k^{(1)}}, \\ &\alpha_{Tx}^{(1)} + \alpha_{Tx}^{(2)}, \alpha_{Tx}^{(1)} + \alpha_{Tx}^{(2)}, \alpha_B^{(1)} + \alpha_B^{(2)}\Big). \end{aligned} \tag{651}$$

- **(P2)** For an arbitrary $\alpha \in \mathbb{R}^4 \times \mathbb{R}_{\geqslant 0}$, define the parameterized operators:

$$\begin{aligned} \phi_B^\alpha := \phi_B(\cdot; \alpha_B), \quad \phi_T^\alpha := \phi_T(\cdot; \alpha_{Tx}, \alpha_{Ty}), \\ \phi_{BC}^\alpha := \phi_{BC}(\cdot; \alpha_k, \alpha_b) \end{aligned} \tag{652}$$

and let $\phi_1^\alpha, \phi_2^\alpha, \phi_3^\alpha$ be an arbitrary permutation of the above three operators. Then, we have that

$$\phi_{BTBC}(x, \alpha) = \phi_1^\alpha \circ \phi_2^\alpha \circ \phi_3^\alpha(x). \tag{653}$$

The property **(P1)** states that ϕ_{BTBC} is almost additive where the exception happens only on the brightness dimension (α_b). The brightness dimension is subject to the same contrast effect implied. The property **(P2)** states that all the three transformations ϕ_B, ϕ_T, and ϕ_{BC} are commutative. The reason is that: (1) ϕ_{BC} is a per-pixel color shift and independent of ϕ_B and ϕ_T; (2) ϕ_B, Gaussian blur, relies on relative position of pixels and the translation with reflection padding, ϕ_T, does not change it.

Based on these two properties, we prove the key results as follows.

Proof of Corollary 13. According to the commutative property **(P2)**, we can view $q(y|x;\, \epsilon)$ as

$$q(y|x,\, \epsilon) = \mathbb{E}_\epsilon p(y|\phi_{BTBC}(x,\, \epsilon)) \tag{654}$$

$$= \mathbb{E}_{\epsilon_k, \epsilon_b} \underbrace{\mathbb{E}_{\epsilon_{Tx}, \epsilon_{Ty}, \epsilon_B} p(y|\phi_{BC}(\phi_T(\phi_B(x, \epsilon_B), \epsilon_{Tx}, \epsilon_{Ty}), \epsilon_k, \epsilon_b))}_{=: q'(y|x; \epsilon_k, \epsilon_b)}. \tag{655}$$

Notice that $q'(y|x;\, \epsilon_k,\, \epsilon_b)$ is a deterministic value in $[0,\, 1]$. Its value is dependent on the distribtuion of $\epsilon_{Tx}, \epsilon_{Ty}, \epsilon_B$ and the underlying base classifier. Luckily, the random variables ϵ_0 and ϵ_1 have the same distribution over the components $\epsilon_{Tx}, \epsilon_{Ty}$ and ϵ_B. Thus, they share the same q' and we write $q(y|x;\, \epsilon_0)$ and $q(y|x;\, \epsilon_1)$ as

$$q(y|x;\, \epsilon_0) = \mathop{\mathbb{E}}_{(\epsilon_k, \epsilon_b) \sim \mathcal{N}(0, \mathrm{diag}(\sigma_k^2, \sigma_b^2))} q'(y|x;\, \epsilon_k, \epsilon_b), \tag{656}$$

$$q(y|x;\, \epsilon_1) = \mathop{\mathbb{E}}_{(\epsilon_k, \epsilon_b) \sim \mathcal{N}(0, \mathrm{diag}(\sigma_k^2, e^{-2k}\sigma_b^2))} q'(y|x;\, \epsilon_k, \epsilon_b). \tag{657}$$

Now, we directly apply Lemma 6 and the desired lower bound for $q(y|x;\, \epsilon_1)$ follows. □

Proof. Proof of Lemma 21 We notice that for any $y \in \mathcal{Y}$,

$$q(y|\phi_{BTBC}(x,\, \alpha);\, \epsilon_0) = \mathbb{E}_{\epsilon_0} p(y|\phi_{BTBC}(\phi_{BTBC}(x,\, \alpha),\, \epsilon_0)) \tag{658}$$

$$\overset{(a.)}{=} \mathbb{E}_{\epsilon_0} p\Big(y|x;\, \alpha + ((\epsilon_0)_k, (\epsilon_0)_b / e^{\alpha_k}, (\epsilon_0)_{Tx}, (\epsilon_0)_{Ty}, (\epsilon_0)_B)^T\Big) \tag{659}$$

$$\overset{(b.)}{=} \mathbb{E}_{\epsilon_1} p\left(y|x; \alpha + \epsilon_1\right). \tag{660}$$

The step (a.) uses the property **(P1)** of transformation ϕ_{BTBC}, and the step (b.) follows the definition of ε_1 in 13 (we define $k := \alpha_k$ hereinafter for simplicity). Thus, $g(\phi_{BTBC}(x,\, \alpha); \epsilon_0) = g(x; \alpha + \epsilon_1)$, and the robustness condition is equivalent to $g(x; \alpha + \epsilon_1) = g(x; \epsilon_1) = y_A$.

According to 5, to prove the robustness, we only need to show that $\xi(\tilde{p}_A) + \xi(1 - \tilde{p}_B) > 1$ given $p'_A > p'_B$. Note that in the definition of ξ, the density functions f_0 and f_1 are for distributions of $\epsilon_1 \sim \mathbb{P}_0$ and $(\alpha + \epsilon_1) \sim \mathbb{P}_1$ respectively.

In the proof below, we will compute the closed-form solution of $\xi(p)$ for any $0 \leqslant p \leqslant 1$, and show that $\xi(\tilde{p}_A) + \xi(1 - \tilde{p}_B) > 1$ given $p'_A > p'_B$. To begin with, we write down f_0 and f_1.

$$f_0(z) = \frac{\lambda_B}{(2\pi)^2\sigma_k\sigma_b\sigma_T^2} \exp\left(-\lambda_B z_B - {}^{(z_{Tx}^2+z_{Ty}^2)}/_{2\sigma_T^2} - {}^{z_k^2}/_{2\sigma_k^2} - {}^{z_b^2}/_{2e^{-2k}\sigma_b^2}\right), \tag{661}$$

$$f_1(z) = \begin{cases} \dfrac{\lambda_B \exp(\lambda_B\alpha_B)}{(2\pi)^2\sigma_k\sigma_b\sigma_T^2} \exp\left(-\lambda_B z_B - {}^{(z_{Tx}-\alpha_{Tx})^2}/_{2\sigma_T^2} - {}^{(z_{Ty}-\alpha_{Ty})^2}/_{2\sigma_T^2} - {}^{(z_k-\alpha_k)^2}/_{2\sigma_k^2} - {}^{(z_b-\alpha_b)^2}/_{2e^{-2k}\sigma_b^2}\right), & \text{if } z_B \geqslant \alpha_B, \\ 0, & \text{otherwise,} \end{cases} \tag{662}$$

where $z = (z_k, z_b, z_{Tx}, z_{Ty}, z_B)^T \in \mathbb{R}^4 \times \mathbb{R}_{\geqslant 0}$. As a result, function $\Lambda = f_1/f_0$ in 5 writes as

$$\Lambda(z) = \begin{cases} \exp\left(\lambda_B\alpha_B - \dfrac{\alpha_{Tx}^2}{2\sigma_T^2} - \dfrac{\alpha_{Ty}^2}{2\sigma_T^2} - \dfrac{\alpha_k^2}{2\sigma_k^2} - \dfrac{\alpha_b^2}{2e^{-2k}\sigma_b^2} + \dfrac{\alpha_{Tx}z_{Tx}}{\sigma_T^2} + \dfrac{\alpha_{Ty}z_{Ty}}{\sigma_T^2} + \dfrac{\alpha_k z_k}{\sigma_k^2} + \dfrac{\alpha_b z_b}{e^{-2k}\sigma_b^2}\right), & \text{if } z_B \geqslant \alpha_B, \\ 0 & \text{otherwise.} \end{cases} \tag{663}$$

It turns out that for any $t > 0$,

$$\begin{aligned} \underline{S}_t &= \{f_1/f_0 < t\} \\ &= \big\{(\hat{z}_k\sigma_k, \hat{z}_b e^{-k}\sigma_b, \hat{z}_{Tx}\sigma_T, \hat{z}_{Ty}\sigma_T)^T \\ &\quad | \, \hat{\alpha}_{Tx}\hat{z}_{Tx} + \hat{\alpha}_{Ty}\hat{z}_{Ty} + \hat{\alpha}_k\hat{z}_k + \hat{\alpha}_b\hat{z}_b \\ &\quad < \ln t + \hat{\alpha}_{Tx}^2/2 + \hat{\alpha}_{Ty}^2/2 + \hat{\alpha}_k^2/2 + \hat{\alpha}_b^2/2 - \lambda_B\alpha_B\big\} \times [\alpha_B, +\infty) \\ &\quad \cup \mathbb{R}^4 \times [0, \alpha_B), \end{aligned} \tag{664}$$

$$\begin{aligned} \overline{S}_t &= \{f_1/f_0 \leqslant t\} \\ &= \big\{(\hat{z}_k\sigma_k, \hat{z}_b e^{-k}\sigma_b, \hat{z}_{Tx}\sigma_T, \hat{z}_{Ty}\sigma_T)^T \\ &\quad | \, \hat{\alpha}_{Tx}\hat{z}_{Tx} + \hat{\alpha}_{Ty}\hat{z}_{Ty} + \hat{\alpha}_k\hat{z}_k + \hat{\alpha}_b\hat{z}_b \\ &\quad \leqslant \ln t + \hat{\alpha}_{Tx}^2/2 + \hat{\alpha}_{Ty}^2/2 + \hat{\alpha}_k^2/2 + \hat{\alpha}_b^2/2 - \lambda_B\alpha_B\big\} \times [\alpha_B, +\infty) \\ &\quad \cup \mathbb{R}^4 \times [0, \alpha_B), \end{aligned} \tag{665}$$

where $\hat{\alpha}_{Tx} = \alpha_{Tx}/\sigma_T, \hat{\alpha}_{Ty} = \alpha_{Ty}/\sigma_T, \hat{\alpha}_k = \alpha_k/\sigma_k, \hat{\alpha}_b = \alpha_b/(e^{-k}\sigma_b)$. When $t = 0$, $\underline{S}_t = \emptyset$ and $\overline{S}_t = \mathbb{R}^4 \times [0, \alpha_B)$. Then, the probability integration shows that

$$\begin{aligned} \tau_p &= \inf\{t \geqslant 0 : \mathbb{P}_0(\overline{S}_t) \geqslant p\} \\ &= \begin{cases} 0, & \text{if } p \leqslant 1 - \exp(-\lambda_B\alpha_B), \\ \exp\left(\lambda_B\alpha_B + \|\hat{\alpha}_{:-1}\|\Phi^{-1}\left(1 - \exp(\lambda_B\alpha_B)(1-p)\right) - 1/2\|\hat{\alpha}_{:-1}\|^2\right), & \text{otherwise,} \end{cases} \end{aligned} \tag{666}$$

where $\|\hat{\alpha}_{:-1}\| = \sqrt{\hat{\alpha}_{T_x}^2 + \hat{\alpha}_{T_y}^2 + \hat{\alpha}_k^2 + \hat{\alpha}_b^2}$. Now we are ready to compute

$$\xi(p) = \sup\{\mathbb{P}_1(S) : \underline{S}_{\tau_p} \subset S \subset \overline{S}_{\tau_p}\}. \tag{667}$$

When $p \leqslant 1 - \exp(-\lambda_B \alpha_B)$, we have $S \subset \mathbb{R}^4 \times [0, \alpha_B)$ and $\mathbb{P}_1(S) = 0$ because $\mathbb{P}_1$ has zero mass for any $z_B < \alpha_B$ (see (662)). When $p > 1 - \exp(-\lambda_B \alpha_B)$, $\tau_p > 0$. Again, from probability integration, we get

$$\mathbb{P}_1(\underline{S}_{\tau_p}) = \mathbb{P}_1(\overline{S}_{\tau_p}) = \Phi\left(\frac{\ln \tau_p - \lambda_B \alpha_B}{\|\hat{\alpha}_{:-1}\|} - \frac{1}{2}\|\hat{\alpha}_{:-1}\|\right). \tag{668}$$

We inject the closed-form solution of τ_p in (666) and yield

$$\xi(p) = \mathbb{P}_1(S) = \Phi\left(\Phi^{-1}\left(1 - (1-p)\exp(\lambda_B \alpha_B)\right) - \|\hat{\alpha}_{:-1}\|\right) \tag{669}$$

for any S satisfying $\underline{S}_{\tau_p} \subset S \subset \overline{S}_{\tau_p}$. We summarize the above equations and write down the closed-form solution of $\xi(p)$ as such:

$$\xi(p) = \begin{cases} 0, & \text{if } p \leqslant 1 - \exp(-\lambda_B \alpha_B), \\ \Phi\left(\Phi^{-1}\left(1 - (1-p)\exp(\lambda_B \alpha_B)\right) - \|\hat{\alpha}_{:-1}\|\right), & \text{otherwise.} \end{cases} \tag{670}$$

We can easily observe that p'_A in lemma statement (643) is indeed $\xi(\tilde{p}_A)$, and p'_B (644) is indeed $1 - \xi(1 - \tilde{p}_B)$. Therefore,

$$\xi(\tilde{p}_A) + \xi(1 - \tilde{p}_B) > 1 \iff p'_A > p'_B \tag{671}$$

and using Theorem 5 concludes the proof. □

Proof. Proof of Corollary 14 Since $q(y_A|x; \varepsilon_1) \geqslant \tilde{p}_A$, according to the complement rule, $\max_{y \neq y_A} q(y|x; \varepsilon_1) < 1 - \tilde{p}_A =: \tilde{p}_B$. Use the $\tilde{p}_A$ and $\tilde{p}_B$ in Lemma 21 we find that $p'_A + p'_B = 1$ always hold. Therefore, $p'_A > 0.5$ guarantees that $p'_A > p'_B$ and thus the robustness. Indeed, from simple algebra, $p'_A > 0.5 \iff$ (646). □

D.3 PROOFS FOR DIFFERENTIALLY RESOLVABLE TRANSFORMATIONS

D.3.1 *Proof of Corollary 4*

Corollary 4 (restated). *Let $\psi(x, \delta) = x + \delta$ and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$. Furthermore, let ϕ be a transformation with parameters in $\mathcal{Z}_\phi \subseteq \mathbb{R}^m$ and let $\mathcal{S} \subseteq \mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. Let $y_A \in \mathcal{Y}$ and suppose that for any i, the ε-smoothed classifier defined by $q(y|x; \varepsilon) := \mathbb{E}(p(y|x + \varepsilon))$ has class probabilities that satisfy*

$$q(y_A|\phi(x, \alpha_i); \varepsilon) \geqslant p_A^{(i)} \geqslant p_B^{(i)} \geqslant \max_{y \neq y_A} q(y|\phi(x, \alpha_i); \varepsilon). \tag{672}$$

Then it is guaranteed that $\forall \alpha \in \mathcal{S}$: $y_A = \arg\max_y q(y|\phi(x, \alpha); \varepsilon)$ if the maximum interpolation error

$$M_{\mathcal{S}} := \max_{\alpha \in \mathcal{S}} \min_{1 \leqslant i \leqslant N} \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2 \tag{673}$$

satisfies

$$M_{\mathcal{S}} < R := \frac{\sigma}{2} \min_{1 \leqslant i \leqslant N} \left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right)\right). \tag{674}$$

Proof. Since the resolvable transformation ψ is given by $\psi(x, \delta) = x + \delta$ we can write

$$\phi(x, \alpha) = \phi(x, \alpha_i) + \underbrace{(\phi(x, \alpha) - \phi(x, \alpha_i))}_{=:\delta_x(\alpha, \alpha_i)}. \tag{675}$$

Furthermore, by assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$ and $g(\cdot; \varepsilon)$ is $(p_A^{(i)}, p_B^{(i)})$-confident at $\phi(x, \alpha_i)$ for y_A and for all i. Thus, by Corollary 8, if δ satisfies

$$\|\delta\|_2 < R_i := \frac{\sigma}{2}\left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right)\right) \tag{676}$$

then it is guaranteed that $y_A = \arg\max_y q(y|\, \phi(x, \alpha_i) + \delta; \varepsilon)$. Let $\Delta_i := B_{R_i}(0)$ and notice that $R \equiv \min_i R_i$ and thus

$$\bigcap_{i=1}^{N} B_{R_i}(0) = B_R(0) = \Delta^*. \tag{677}$$

To see that Δ^* has the desired property, consider

$$\forall \alpha \in \mathcal{S}\, \exists \alpha_i \colon\ \delta_x(\alpha, \alpha_i) \in \Delta^* \tag{678}$$

$$\iff\ \forall \alpha \in \mathcal{S}\, \exists \alpha_i \colon\ \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2 < R. \tag{679}$$

Since $R \leqslant R_i$ it follows that for $\delta_i = \phi(x, \alpha) - \phi(x, \alpha_i)$ it is guaranteed that

$$y_A = \arg\max_y q(y|\, \phi(x, \alpha_i) + \delta_i; \varepsilon) \tag{680}$$

$$= \arg\max_y q(y|\, \phi(x, \alpha); \varepsilon). \tag{681}$$

Thus, the set Δ^* has the desired property. In particular, since

$$\forall \alpha \in \mathcal{S}\, \exists \alpha_i \colon\ \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2 < R \tag{682}$$

$$\iff \max_{\alpha \in \mathcal{S}} \min_{1 \leqslant i \leqslant N} \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2 < R \tag{683}$$

the statement follows. □

D.3.2 *Composition of Rotation and Scaling with Brightness*

To certify the composition of scaling and brightness or rotation and brightness, we follow the same methodology as certifying scaling or rotation alone and reuse the computed interpolatation error $M_{\mathcal{S}}$. We only make the following two changes: (1) alter the smoothing distribution from additive Gaussian noise $\psi(x, \delta) = x + \delta$ where $\delta \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$ to additive Gaussian noise and Gaussian brightness change $\psi(x, \delta, \delta_b) = x + \delta + b \cdot \mathbb{1}_d$ where $\delta \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d), b \sim \mathcal{N}(0, \sigma_b^2)$; (2) change the robustness condition from $R > M_{\mathcal{S}}$ in Corollary 4 to $R > \sqrt{M_{\mathcal{S}}^2 + (\sigma^2/\sigma_b^2) b_0^2}$.

Corollary 15. *Let* $\psi_B(x, \delta, b) = x + \delta + b \cdot \mathbb{1}_d$ *and let* $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$, $\varepsilon_b \sim \mathcal{N}(0, \sigma_b^2)$. *Furthermore, let* ϕ *be a transformation with parameters in* $\mathcal{Z}_\phi \subseteq \mathbb{R}^m$ *and let* $\mathcal{S} \subseteq \mathcal{Z}_\phi$ *and* $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. *Let* $y_A \in \mathcal{Y}$ *and suppose that for any* i, *the* $(\varepsilon, \varepsilon_b)$*-smoothed classifier* $q(y|\, x; \varepsilon, \varepsilon_b) := \mathbb{E}(p(y|\, \psi_B(x, \varepsilon, \varepsilon_b))$ *satisfies*

$$q(y_A|\, x; \varepsilon, \varepsilon_b) \geqslant p_A^{(i)} > p_B^{(i)} \geqslant \max_{y \neq y_A} q(y|\, x; \varepsilon, \varepsilon_b). \tag{684}$$

for each i. *Let*

$$R := \frac{\sigma}{2} \min_{1 \leqslant i \leqslant N} \left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right) \right) \tag{685}$$

Then, $\forall \alpha \in \mathcal{S}$ *and* $\forall b \in [-b_0, b_0]$ *it is guaranteed that* $y_A = \arg\max_y q(y|\phi(x,\alpha) + b \cdot \mathbb{1}_d; \varepsilon, \varepsilon_b)$ *as long as*

$$R > \sqrt{M_S^2 + \frac{\sigma^2}{\sigma_b^2} b_0^2}, \tag{686}$$

where M_S *is defined as in 4.*

Proof. Since the resolvable transformation ψ_B is given by

$$\psi_B(x, \delta, b) = x + \delta + b \cdot \mathbb{1}_d, \tag{687}$$

we can write

$$\phi(x, \alpha) + b \cdot \mathbb{1}_d = \phi(x, \alpha_i) + \underbrace{(\phi(x, \alpha) - \phi(x, \alpha_i)) + b \cdot \mathbb{1}_d}_{=:\delta_x((\alpha,b),(\alpha_i,0))}. \tag{688}$$

Furthermore, by assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$, $\varepsilon_b \sim \mathcal{N}(0, \sigma_b^2)$ and $g(\cdot; \varepsilon, \varepsilon_b)$ is $(p_A^{(i)}, p_B^{(i)})$-confident at $\phi(x, \alpha_i)$ for y_A and all i. Thus, by Corollary 8, if δ and b satisfy

$$\sqrt{\frac{\|\delta\|_2^2}{\sigma^2} + \frac{b^2}{\sigma_b^2}} < \frac{1}{2}\left(\Phi^{-1}(p_A^{(i)}) - \Phi^{-1}(p_B^{(i)})\right), \tag{689}$$

then it is guaranteed that

$$y_A = \arg\max_y q(y|\phi(x, \alpha_i) + \delta + b \cdot \mathbb{1}_d; \varepsilon, \varepsilon_b). \tag{690}$$

Let

$$R_i := \frac{\sigma}{2}\left(\Phi^{-1}(p_A^{(i)}) - \Phi^{-1}(p_B^{(i)})\right) \tag{691}$$

and note that without loss of generality we can assume that $R_i > \sigma/\sigma_b b_0$, because otherwise the robustness condition is violated. Rearranging terms in (689) leads to the condition

$$\|\delta\|_2 < \sqrt{R_i^2 - \frac{\sigma^2}{\sigma_b^2} b^2} \tag{692}$$

that can be turned into a sufficient robustness condition holding for any $b \in [-b_0, b_0]$ simultaneously

$$\|\delta\|_2 < \sqrt{R_i^2 - \frac{\sigma^2}{\sigma_b^2} b_0^2} \tag{693}$$

Note that, without loss of generality For each i let Δ_i be the set defined as

$$\Delta_i := \left\{ \delta + b \cdot \mathbf{1}_d \in \mathbb{R}^d \colon \|\delta\|_2 < \sqrt{R_i^2 - \frac{\sigma^2}{\sigma_b^2} b_0^2},\ |b| \leqslant b_0 \right\} \tag{694}$$

and note that

$$\Delta^* := \bigcap_{i=1}^{N} \Delta_i \tag{695}$$

$$= \left\{ \delta + b \cdot \mathbf{1}_d \in \mathbb{R}^d \colon \|\delta\|_2 < \sqrt{R^2 - \frac{\sigma^2}{\sigma_b^2} b_0^2},\ |b| \leqslant b_0 \right\} \tag{696}$$

with $R := \min_i R_i$. Clearly, if $\forall \alpha \in \mathcal{S}$, $\forall b \in [-b_0, b_0]\ \exists i$ such that

$$\delta_x((\alpha, b), (\alpha_i, 0)) \in \Delta^* \tag{697}$$

then it is guaranteed that

$$y_A = \arg\max_y q(y|\, \phi(x, \alpha_i) + \delta_x((\alpha, b), (\alpha_i, 0))) \tag{698}$$

$$= \arg\max_y q(y|\, \phi(x, \alpha) + b \cdot \mathbb{1}_d). \tag{699}$$

We can thus reformulate the robustness condition as

$$\forall \alpha \in \mathcal{S},\ \forall b \in [-b_0, b_0]\ \exists i \text{ s.t. } \delta_x((\alpha, b), (\alpha_i, 0)) \in \Delta^* \tag{700}$$

$$\iff$$

$$\forall \alpha \in \mathcal{S},\ \forall b \in [-b_0, b_0]\ \exists i \text{ s.t. } \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2 < \sqrt{R^2 - \frac{\sigma^2}{\sigma_b^2} b_0^2} \tag{701}$$

$$\iff$$

$$\max_{\alpha \in \mathcal{S}} \min_{1 \leqslant i \leqslant N} \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2 < \sqrt{R^2 - \frac{\sigma^2}{\sigma_b^2} b_0^2} \tag{702}$$

that, written in terms of the maximum ℓ_2 interpolation error $M_{\mathcal{S}}$, is equivalent to

$$R > \sqrt{M_{\mathcal{S}}^2 + \frac{\sigma^2}{\sigma_b^2} b_0^2} \tag{703}$$

what concludes the proof. □

D.3.3 *Composition of Scaling and Rotation with Brightness and ℓ_2 Perturbations*

We use the same smoothing distribution as above, and the following corollary directly allows us to certify the robustness against the composition of scaling/rotation, brightness, and an additional ℓ_2-bounded perturbations—we only need to change the robustness condition from $R > \sqrt{M_{\mathcal{S}}^2 + (\sigma^2/\sigma_b^2) b_0^2}$ to $R > \sqrt{(M_{\mathcal{S}} + r)^2 + (\sigma^2/\sigma_b^2) b_0^2}$.

Corollary 16. *Under the same setting as in 15, for $\forall \alpha \in \mathcal{S}$, $\forall b \in [-b_0, b_0]$ and $\forall \delta \in \mathbb{R}^d$ such that $\|\delta\|_2 \leqslant r$, it is guaranteed that $y_A = \arg\max_k q(y|\, \phi(x, \alpha) + b \cdot \mathbb{1}_d + \delta;\ \varepsilon, \varepsilon_d)$ as long as*

$$R > \sqrt{(M_{\mathcal{S}} + r)^2 + \frac{\sigma^2}{\sigma_b^2} b_0^2}, \tag{704}$$

where $M_{\mathcal{S}}$ is defined as in 4.

Proof. Note that we can write the transformed input as

$$\begin{aligned}&\phi(x,\alpha)+b\cdot\mathbb{1}_d+\delta\\&\quad=\phi(x,\alpha_i)+\underbrace{(\phi(x,\alpha)-\phi(x,\alpha_i)+\delta)+b\cdot\mathbb{1}_d}_{=:\delta_x((\alpha,b,\delta),(\alpha_i,0,0))}.\end{aligned}\tag{705}$$

Since we use the same smoothing protocol as in 15, the general proof idea is similar to 15 — we use the same resolvable transformation ψ_B and define the same set Δ_i, namely

$$\begin{aligned}\Delta_i:=\Big\{&\delta'+b\cdot\mathbf{1}_d+\delta\in\mathbb{R}^d:\\&\|\delta'+\delta\|_2<\sqrt{R_i^2-\frac{\sigma^2}{\sigma_b^2}b_0^2},\,|b|\leqslant b_0,\,\|\delta\|_2\leqslant r\Big\}.\end{aligned}\tag{706}$$

and set

$$\Delta^*:=\bigcap_{i=1}^{N}\Delta_i\tag{707}$$

$$\begin{aligned}=\Big\{&\delta'+b\cdot\mathbf{1}_d+\delta\in\mathbb{R}^d:\\&\|\delta'+\delta\|_2<\sqrt{R^2-\frac{\sigma^2}{\sigma_b^2}b_0^2},\,|b|\leqslant b_0,\,\|\delta\|_2\leqslant r\Big\}\end{aligned}\tag{708}$$

with $R:=\min_i R_i$. Clearly, if $\forall\alpha\in\mathcal{S}$, $\forall b\in[-b_0,b_0]$, $\|\delta\|_2\leqslant r\,\exists i$ such that

$$\delta_x((\alpha,b,\delta),(\alpha_i,0))\in\Delta^*\tag{709}$$

then it is guaranteed that

$$y_A=\arg\max_y q(y|\,\phi(x,\alpha_i)+\delta_x((\alpha,b),(\alpha_i,0)))\tag{710}$$

$$=\arg\max_y q(y|\,\phi(x,\alpha)+b\cdot\mathbb{1}_d).\tag{711}$$

We can thus reformulate the robustness condition as

$$\forall\alpha\in\mathcal{S},\,\forall b\in[-b_0,b_0],\,\|\delta\|_2\leqslant r\,\exists i\,\text{s.t. }\delta_x((\alpha,b,\delta),(\alpha_i,0,0))\in\Delta^*\tag{712}$$

$$\iff$$

$$\begin{aligned}&\forall\alpha\in\mathcal{S},\,\forall b\in[-b_0,b_0],\,\|\delta\|_2\leqslant r\\&\qquad\exists i\text{ s.t. }\|\phi(x,\alpha)-\phi(x,\alpha_i)+\delta\|_2<\sqrt{R^2-\frac{\sigma^2}{\sigma_b^2}b_0^2}\end{aligned}\tag{713}$$

$$\iff$$

$$\max_{\alpha\in\mathcal{S}}\min_{1\leqslant i\leqslant N}\|\phi(x,\alpha)-\phi(x,\alpha_i)+\delta\|_2<\sqrt{R^2-\frac{\sigma^2}{\sigma_b^2}b_0^2}.\tag{714}$$

Note that by the triangle inequality have

$$\max_{\alpha\in\mathcal{S}}\min_{1\leqslant i\leqslant N}\|\phi(x,\alpha)-\phi(x,\alpha_i)+\delta\|_2\leqslant M_{\mathcal{S}}+\|\delta\|_2\leqslant M_{\mathcal{S}}+r\tag{715}$$

and thus, robustness is implied by

$$R > \sqrt{(M_{\mathcal{S}} + r)^2 + \frac{\sigma^2}{\sigma_b^2} b_0^2} \tag{716}$$

what concludes the proof. □

D.4 PROOFS FOR SCALING AND ROTATION TRANSFORMATIONS

In this section we state the proofs for the theoretical results governing our approach to certifying rotations and scaling transformations using randomized smoothing. We first define the maximum ℓ_2 interpolation error. First, let us recall the following definitions from the main part of this thesis.

Definition 11 (ℓ_2 interpolation error). *Let $x \in \mathcal{X}$, $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ a transformation, $\mathcal{S} = [a, b]$, $N \in \mathbb{N}$ and suppose $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. The maximum ℓ_2 interpolation error is defined as*

$$M_{\mathcal{S}} := \max_{a \leqslant \alpha \leqslant b} \min_{1 \leqslant i \leqslant N} \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2. \tag{717}$$

Definition 5 (restated). *For pixels $(i, j) \in \Omega$, we define the grid pixel generator G_{ij} as*

$$G_{ij} := \{(i, j), (i+1, j), (i, j+1), (i+1, j+1)\}. \tag{718}$$

Definition 6 (restated). *We define the operator that extracts the channel-wise maximum pixel wise on a grid $S \subseteq \Omega$ as the map $\overline{m}\colon \mathbb{R}^{K \times W \times H} \times \{0, \ldots, K-1\} \times 2^{\Omega} \to \mathbb{R}$ with*

$$\overline{m}(x, k, S) := \max_{(i,j) \in S} \left(\max_{(r,s) \in G_{ij}} x_{k,r,s} \right) \tag{719}$$

Definition 7 (restated). *We define the operator that extracts the channel-wise maximum change in color on a grid $S \subseteq \Omega$ as the map $m_\Delta\colon \mathbb{R}^{K \times W \times H} \times \{0, \ldots, K-1\} \times 2^{\Omega} \to \mathbb{R}$ with*

$$m_\Delta(x, k, S) := \max_{(i,j) \in S} \left(\max_{(r,s) \in G_{ij}} x_{k,r,s} - \min_{(r,s) \in G_{ij}} x_{k,r,s} \right) \tag{720}$$

The following auxiliary lemma is used for both rotation and scaling:

Lemma 22. *Let $x \in \mathbb{R}^{K \times W \times H}$, $-\infty < t_1 < t_2 < \infty$ and suppose $\rho\colon [t_1, t_2] \to [0, W-1] \times [0, H-1]$ is a curve of class C^1. Let*

$$\psi_k\colon [t_1, t_2] \to \mathbb{R}, \quad \psi_k(t) := Q_x(k, \rho_1(t), \rho_2(t)) \tag{721}$$

where $k \in \Omega_K$ and Q_x denotes bilinear interpolation. Then ψ_k is L_k-Lipschitz continuous with constant

$$L_k = \max_{t \in [t_1, t_2]} \left(\sqrt{2}\, \|\dot{\rho}(t)\|_2 \cdot m_\Delta(x, k, \lfloor \rho(t) \rfloor) \right) \tag{722}$$

Proof. Note that the function $t \mapsto \lfloor \rho(t) \rfloor$ is piecewise constant and let $t_1 =: u_1 < u_2 < \ldots < u_{N_0} := t_2$ such that $\lfloor \rho(t) \rfloor$ is constant on $[u_i, u_{i+1})$ for all $1 \leqslant i \leqslant N_0 - 1$ and $\dot{\cup}_{i=1}^{N_0} [u_i, u_{i+1}) = [t_1, t_2)$. We notice that ψ_k is a continuous real-valued function since it is the composition of the continuous Q_x and C^1-curve ρ. L_k-Lipschitz continuity

on $[t_1, t_2)$ thus follows if we show that ψ_k is L_k-Lipschitz on each interval in the partition. For that purpose, let $1 \leqslant i \leqslant N_0$ be arbitrary and fix some $t \in [u_i, u_{i+1})$. Let $(w, h) := \lfloor \rho(t) \rfloor$ and $\gamma(t) := \rho(t) - \lfloor \rho(t) \rfloor$ and notice that $\gamma(t) \in [0, 1)^2$. Let

$$V_1 := x_{k,w,h},\ V_2 := x_{k,w,h+1}, \tag{723}$$

$$V_3 := x_{k,w+1,h},\ V_4 := x_{k,w+1,h+1}, \tag{724}$$

Then, for any $u \in [u_i, u_{i+1})$

$$\psi_k(u) = Q_x(k, \rho_1(u), \rho_2(u)) \tag{725}$$

$$\begin{aligned} &= (1-\gamma_1(u)) \cdot ((1-\gamma_2(u)) \cdot V_1 + \gamma_2(u) \cdot V_2) \\ &\qquad + \gamma_1(u) \cdot ((1-\gamma_2(u) \cdot V_3 + \gamma_2(u) \cdot V_4). \end{aligned} \tag{726}$$

Let $m_\Delta := m_\Delta(x, k \lfloor \rho(t) \rfloor)$ and notice that by definition

$$m_\Delta = \max_i V_i - \min_i V_i \tag{727}$$

and in particular

$$\left|V_i - V_j\right| \leqslant m_\Delta \quad \forall i, j. \tag{728}$$

Since V_i is constant for each i and γ is differentiable, ψ_k is differentiable on $[u_i, u_{i+1})$ and hence

$$\dot\psi_k(u) = (\dot\gamma_1(u)\gamma_2(u) + \gamma_1(u)\dot\gamma_2(u))(V_1 - V_2 - V_3 + V_4) \tag{729}$$

$$+ \dot\gamma_1(u)(V_3 - V_1) + \dot\gamma_2(u)(V_2 - V_1). \tag{730}$$

Note that the derivative $\dot\psi_k$ is linear in γ_1 and γ_2 and hence its extreme values are bounded when evaluated at extreme values of γ, that is $(\gamma_1, \gamma_2) \in \{0, 1\}^2$. We treat each case separately:

- $\gamma_1 = \gamma_2 = 0$. Then,

$$\left|\dot\psi_k\right| \leqslant |\dot\gamma_1(V_3 - V_1) + \dot\gamma_2(V_2 - V_1)| \tag{731}$$

$$\leqslant |\dot\gamma_1| \cdot |V_3 - V_1| + |\dot\gamma_2| \cdot |V_2 - V_1| \tag{732}$$

$$\leqslant m_\Delta(|\dot\gamma_1| + |\dot\gamma_2|) \tag{733}$$

- $\gamma_1 = \gamma_2 = 1$. Then,

$$\left|\dot\psi_k\right| \leqslant |\dot\gamma_1(V_4 - V_2) + \dot\gamma_2(V_4 - V_3)| \tag{734}$$

$$\leqslant |\dot\gamma_1| \cdot |V_4 - V_2| + |\dot\gamma_2| \cdot |V_4 - V_3| \tag{735}$$

$$\leqslant m_\Delta(|\dot\gamma_1| + |\dot\gamma_2|) \tag{736}$$

- $\gamma_1 = 0$, $\gamma_2 = 1$. Then,

$$\left|\dot\psi_k\right| \leqslant |\dot\gamma_1(V_4 - V_2) + \dot\gamma_2(V_2 - V_1)| \tag{737}$$

$$\leqslant |\dot\gamma_1| \cdot |V_4 - V_2| + |\dot\gamma_2| \cdot |V_2 - V_1| \tag{738}$$

$$\leqslant m_\Delta(|\dot\gamma_1| + |\dot\gamma_2|) \tag{739}$$

- $\gamma_1 = 1$, $\gamma_2 = 0$. Then,

$$\begin{aligned} \left|\dot{\psi}_k\right| &\leqslant |\dot{\gamma}_1(V_3 - V_1) + \dot{\gamma}_2(V_4 - V_3)| && (740)\\ &\leqslant |\dot{\gamma}_1| \cdot |V_3 - V_1| + |\dot{\gamma}_2| \cdot |V_4 - V_3| && (741)\\ &\leqslant m_\Delta(|\dot{\gamma}_1| + |\dot{\gamma}_2|) && (742)\end{aligned}$$

Hence, for any $u \in [u_i, u_{i+1})$, the modulus of the derivative is bounded by $m_\Delta(|\dot{\gamma}_1| + |\dot{\gamma}_2|)$. We can further bound this by observing the following connection between ℓ_1 and ℓ_2 distance

$$\forall x \in \mathbb{R}^n : \quad \|x\|_1 = |\langle |x|, \mathbf{1}\rangle| \leqslant \|x\|_2 \|\mathbf{1}\|_2 = \sqrt{n}\|x\|_2 \tag{743}$$

and hence $\forall u \in [u_i, u_{i+1})$

$$\begin{aligned} |\psi_k(u)| &\leqslant m_\Delta \|\dot{\gamma}(u)\|_1 && (744)\\ &\leqslant m_\Delta\sqrt{2}\|\dot{\gamma}(u)\|_2 && (745)\\ &= m_\Delta\sqrt{2}\|\dot{\rho}(u)\|_2 . && (746)\end{aligned}$$

Since ψ_k is differentiable on $[u_i, u_{i+1})$, its Lipschitz constant is bounded by the maximum absolute value of its derivative. Hence

$$\begin{aligned} &\max_{u\in[u_i,u_{i+1})} m_\Delta\sqrt{2}\|\dot{\rho}(u)\|_2 \\ &\qquad = \max_{u\in[u_i,u_{i+1})} m_\Delta(x, k, \lfloor\rho(u)\rfloor)\sqrt{2}\|\dot{\rho}(u)\|_2 && (747)\\ &\qquad \leqslant \max_{u\in[t_1,t_2)} m_\Delta(x, k, \lfloor\rho(u)\rfloor)\sqrt{2}\|\dot{\rho}(u)\|_2 = L_k && (748)\end{aligned}$$

is a Lipschitz constant for ψ_k on $[u_i, u_{i+1})$. Note that L_k does not depend on i. Furthermore, i was chosen arbitrarily and hence L_k is a Lipschitz constant for ψ_k on $[t_1, t_2)$ and due to continuity on $[t_1, t_2]$, concluding the proof. □

D.4.1 *Bilinear Interpolation*

Let $\Omega_K := \{0, \dots, K-1\}$ and $\Omega := [0, W-1] \times [0, H-1]$. We define bilinear interpolation to be the map $Q\colon \mathbb{R}^{K\times W\times H} \to L^2(\Omega_K \times \mathbb{R}^2, \mathbb{R})$, $x \mapsto Q(x) =: Q_x$ where Q_x is given by

$$(k, i, j) \mapsto Q_x(k, i, j) := \begin{cases} 0 & (i, j) \notin \Omega \\ x_{k,i,j} & (i, j) \in \Omega \cap \mathbb{N}^2 \\ \tilde{x}_{k,i,j} & (i, j) \in \Omega \setminus \mathbb{N}^2. \end{cases} \tag{749}$$

and where

$$\begin{aligned} \tilde{x}_{k,i,j} := (1 - (i - \lfloor i\rfloor)) \cdot \big((1 - (j - \lfloor j\rfloor)) \cdot x_{k,\lfloor i\rfloor,\lfloor j\rfloor} \\ + (j - \lfloor j\rfloor) \cdot x_{k,\lfloor i\rfloor,\lfloor j\rfloor+1}\big) \\ + (i - \lfloor i\rfloor) \cdot \big((1 - (j - \lfloor j\rfloor)) \cdot x_{k,\lfloor i\rfloor+1,\lfloor j\rfloor} \\ + (j - \lfloor j\rfloor) \cdot x_{k,\lfloor i\rfloor+1,\lfloor j\rfloor+1}\big) . \end{aligned} \tag{750}$$

D.4.2 *Rotation*

The rotation transformation is denoted as $\phi_R\colon \mathbb{R}^{K\times W\times H}\times\mathbb{R}\to\mathbb{R}^{K\times W\times H}$ and acts on an image in three steps that we will highlight in greater detail. First, it rotates the image by α degrees counter-clockwise. After rotation, pixel values are determined using bilinear interpolation (749). Finally, we apply black padding to all pixels (i, j) whose ℓ_2-distance to the center pixel is larger than half of the length of the shorter side, and denote this operation by P. Let c_W and c_H be the center pixels

$$c_W := \frac{W-1}{2}, \qquad c_H := \frac{H-1}{2}. \tag{751}$$

and

$$\begin{aligned} d_{i,j} &= \sqrt{(i-c_W)^2+(j-c_H)^2}, \\ g_{i,j} &= \operatorname{arctan2}(j-c_H,\, i-c_W)\,. \end{aligned} \tag{752}$$

We write $\tilde{\phi}_R$ for the rotation transformation before black padding and decompose ϕ_R as $\phi_R = P\circ\tilde{\phi}_R$, where $\tilde{\phi}_R\colon \mathbb{R}^{K\times W\times H}\times\mathbb{R}\to\mathbb{R}^{K\times W\times H}$ is defined by

$$\begin{aligned} \tilde{\phi}_R(x,\,\alpha)_{k,i,j} := Q_x(k,\, &c_W + d_{i,j}\,\cos(g_{i,j}-\alpha), \\ &c_H + d_{i,j}\,\sin(g_{i,j}-\alpha)) \end{aligned} \tag{753}$$

and $P\colon \mathbb{R}^{K\times W\times H}\to\mathbb{R}^{K\times W\times H}$ by

$$f \mapsto P(f)_{k,i,j} = \begin{cases} f(k,\,i,\,j) & d_{i,j} < \min\{c_W,\, c_H\} \\ 0 & \text{otherwise} \end{cases}. \tag{754}$$

The rotation transformation in practice may use different padding mechanisms. For example, the rotation in the physical world may fill in boundary pixels with real elements captured by the camera. We remark that our TSS against the transformation ϕ_R implies the defense against rotation with *any other* padding mechanisms, because we first apply black-padding P to any such rotated input and then feed into TSS models so that TSS models always receive black-padded inputs. We now prove the following result:

Lemma 8 (restated). *Let $x\in\mathbb{R}^{K\times W\times H}$ be a K-channel image and let $\phi_R = P\circ I\circ\tilde{\phi}_R$ be the rotation transformation. Then, a global Lipschitz constant L for the functions $\{g_i\}_{i=1}^N$ is given by*

$$L_r = \max_{1\leqslant i\leqslant N-1}\sum_{k=0}^{K-1}\sum_{r,s\in V} 2d_{r,s}\cdot m_\Delta(x,k,\mathcal{P}^{(i)}_{r,s})\cdot\overline{m}(x,\,k,\,\mathcal{P}^{(i)}_{r,s}) \tag{755}$$

where $V = \left\{(r,s)\in\mathbb{N}^2 \middle|\, d_{r,s} < \frac{1}{2}(\min\{W,H\}-1)\right\}$. The set $\mathcal{P}^{(i)}_{r,s}$ is given by all integer grid pixels that are covered by the trajectory of source pixels of (r,s) when rotating from angle α_i to α_{i+1}.

Proof. Recall that ϕ_R acts on images $x\in\mathbb{R}^{K\times W\times H}$ and that g_i is defined as

$$\begin{aligned} g_i(\alpha) &= \|\phi_R(x,\,\alpha)-\phi_R(x,\,\alpha_i)\|_2^2 \\ &= \sum_{k=0}^{K-1}\sum_{r=0}^{W-1}\sum_{s=0}^{H-1}(\phi_R(x,\,\alpha)_{k,r,s}-\phi_R(x,\,\alpha_i)_{k,r,s})^2 \end{aligned} \tag{756}$$

Let c_W and c_H denote the center pixels

$$c_W := \frac{W-1}{2}, \qquad c_H := \frac{H-1}{2}. \tag{757}$$

and recall the following quantities from the definition of ϕ_R (D.4.2):

$$\begin{aligned} d_{r,s} &= \sqrt{(r-c_W)^2 + (s-c_H)^2}, \\ g_{r,s} &= \arctan 2\,(s-c_H,\, r-c_W) \end{aligned} \tag{758}$$

Note that

$$d_{r,s} \geqslant \min\{c_W,\, c_H\} \;\Rightarrow\; \phi_R(x,\, \alpha)_{k,r,s} = 0. \tag{759}$$

We thus only need to consider pixels that lie inside the centered disk. We call the collection of such pixels *valid* pixels, denoted by V:

$$V := \left\{(r,\, s) \in \mathbb{N}^2 \mid d_{r,s} < \min\{c_W,\, c_H\}\right\}. \tag{760}$$

Let $f_1^{r,s} \colon \mathbb{R} \to \mathbb{R}$ and $f_2^{r,s} \colon \mathbb{R} \to \mathbb{R}$ be functions defined as

$$\begin{aligned} f_1^{r,s}(\alpha) &= c_W + d_{r,s}\, \cos(g_{r,s} - \alpha), \\ f_2^{r,s}(\alpha) &= c_H + d_{r,s}\, \sin(g_{r,s} - \alpha). \end{aligned} \tag{761}$$

Then for any valid pixel $(r,\, s) \in V$, the value of the rotated image $\phi_R(x,\, \alpha)$ is given by

$$\phi_R(x,\, \alpha)_{k,r,s} = Q_x(k,\, f_1^{r,s}(\alpha),\, f_2^{r,s}(\alpha)) \tag{762}$$

where Q_x denotes bilinear interpolation. We define the shorthand

$$g_i^{k,r,s}(\alpha) := (\phi_R(x,\, \alpha)_{k,r,s} - \phi_R(x,\, \alpha_i)_{k,r,s})^2 \tag{763}$$

and denote by $L_i^{k,r,s}$ and $L_{i+1}^{k,r,s}$ the Lipschitz constants of $g_i^{k,r,s}$ and $g_{i+1}^{k,r,s}$ on $[\alpha_i,\, \alpha_{i+1}]$. We can write (756) as

$$\begin{aligned} g_i(\alpha) &= \sum_{k=0}^{K-1} \sum_{(r,\,s)\in V} g_i^{k,r,s}(\alpha), \\ g_{i+1}(\alpha) &= \sum_{k=0}^{K-1} \sum_{(r,\,s)\in V} g_{i+1}^{k,r,s}(\alpha) \end{aligned} \tag{764}$$

and note that Lipschitz constants of g_i and g_{i+1} on $[\alpha_i,\, \alpha_{i+1}]$ are given by

$$\max_{c,\, d \in [\alpha_i,\, \alpha_{i+1}]} \frac{|g_i(c) - g_i(d)|}{|c-d|} \leqslant \left(\sum_{k=0}^{K-1} \sum_{(r,\,s)\in V} L_i^{k,r,s} \right) =: L_i \tag{765}$$

$$\max_{c,\, d \in [\alpha_i,\, \alpha_{i+1}]} \frac{|g_{i+1}(c) - g_{i+1}(d)|}{|c-d|} \leqslant \left(\sum_{k=0}^{K-1} \sum_{(r,\,s)\in V} L_{i+1}^{k,r,s} \right) =: L_{i+1} \tag{766}$$

We can hence determine L according to equation (149) as

$$L = \max_i \{\max\{L_i,\, L_{i+1}\}\}. \tag{767}$$

Without loss of generality, consider $L_i^{k,r,s}$ and note that

$$\max_{c,d\in[\alpha_i,\,\alpha_{i+1}]} \left| \frac{g_i^{k,r,s}(c) - g_i^{k,r,s}(d)}{c-d} \right| \tag{768}$$

$$= \max_{c,d\in[\alpha_i,\,\alpha_{i+1}]} \left| \frac{\phi_R(x,\,c)_{k,r,s} - \phi_R(x,\,d)_{k,r,s}}{c-d} \right| \cdot \left|\phi_R(x,\,c)_{k,r,s} + \phi_R(x,\,d)_{k,r,s} - 2\phi_R(x,\,\alpha_i)_{k,r,s}\right| \tag{769}$$

$$\leqslant \underbrace{\max_{c,d\in[\alpha_i,\,\alpha_{i+1}]} \left| \frac{\phi_R(x,\,c)_{k,r,s} - \phi_R(x,\,d)_{k,r,s}}{c-d} \right|}_{\text{(I)}} \cdot 2 \max_{\theta\in[\alpha_i,\,\alpha_{i+1}]} \underbrace{\left|\phi_R(x,\,\theta)_{k,r,s} - \phi_R(x,\,\alpha_i)_{k,r,s}\right|}_{\text{(II)}}. \tag{770}$$

To compute a Lipschitz constant for $g_i^{k,r,s}$ on the interval $[\alpha_i,\,\alpha_{i+1}]$ we thus only need to compute a Lipschitz constant for $\phi_R(x,\,\cdot)$ on $[\alpha_i,\,\alpha_{i+1}]$ and an upper bound on (II). For that purpose, note that ϕ_R takes only positive values and consider

$$\text{(II)} \leqslant \max_{\theta\in[\alpha_i,\,\alpha_{i+1}]} \{\phi_R(x,\,\theta)_{k,r,s},\,\phi_R(x,\,\alpha_i)_{k,r,s}\} \tag{771}$$

$$= \max_{\theta\in[\alpha_i,\,\alpha_{i+1}]} \phi_R(x,\,\theta)_{k,r,s} \tag{772}$$

Notice that now both $L_i^{k,r,s}$ and $L_{i+1}^{k,r,s}$ share the same upper bound. Recall (762), i.e.,

$$\phi_R(x,\,\theta)_{k,r,s} = Q_x(k,\,f_1^{r,s}(\theta),\,f_2^{r,s}(\theta)). \tag{773}$$

Now, we upper bound (771) by finding all integer grid pixels that are covered by the trajectory $(f_1^{r,s}(\theta),\,f_2^{r,s}(\theta))$. Specifically, let

$$\mathcal{P}_{r,s}^{(i)} := \bigcup_{\theta\in[\alpha_i,\,\alpha_{i+1}]} \left(\lfloor f_1^{r,s}(\theta)\rfloor,\,\lfloor f_2^{r,s}(\theta)\rfloor\right). \tag{774}$$

Since ϕ_R is interpolated from integer pixels, we can consider the maximum over $\mathcal{P}_{r,s}^{(i)}$ in order to upper bound (771):

$$\max_{\theta\in[\alpha_i,\,\alpha_{i+1}]} \phi_R(x,\,\theta)_{k,r,s} = \max_{\theta\in[\alpha_i,\,\alpha_{i+1}]} Q_x(k,\,f_1^{r,s}(\theta),\,f_2^{r,s}(\theta)) \tag{775}$$

$$\leqslant \max_{(i,j)\in\mathcal{P}_{r,s}} \max\big\{x(k,\,i,\,j),\,x(k,\,i+1,\,j),\,x(k,\,i,\,j+1),\,x(k,\,i+1,\,j+1)\big\} \tag{776}$$

$$= \bar{m}(x,\,k,\,\mathcal{P}_{r,s}^{(i)}). \tag{777}$$

We now have to find an upper bound of (I), that is, a Lipschitz constant of $\phi_R(x, \cdot)_{k,r,s}$ on the interval $[\alpha_i, \alpha_{i+1}]$. For that purpose, consider the following. Note that the curve $\rho\colon [\alpha_i, \alpha_{i+1}] \to \mathbb{R}^2$, $\rho(t) := (f_1^{r,s}(t), f_2^{r,s}(t))$ is of class C^1 and

$$\begin{aligned}\frac{df_1^{r,s}(t)}{dt} &= \frac{d}{dt}\left(c_W + d_{r,s}\,\cos(g_{r,s} - t)\right) && (778)\\ &= d_{r,s}\,\sin(g_{r,s} - t) && (779)\\ \frac{df_2^{r,s}(t)}{dt} &= \frac{d}{dt}\left(c_H + d_{r,s}\,\sin(g_{r,s} - t)\right) && (780)\\ &= -d_{r,s}\,\cos(g_{r,s} - t) && (781)\end{aligned}$$

and hence

$$\|\dot\rho(t)\|_2 = \sqrt{\left(\frac{df_1^{r,s}(t)}{dt}\right)^2 + \left(\frac{df_2^{r,s}(t)}{dt}\right)^2} = \sqrt{2}\,d_{r,s}. \tag{782}$$

By Lemma 22 a Lipschitz constant for the function $\phi_R(x, \cdot)_{k,r,s}$ is thus given by

$$\begin{aligned}&\max_{c,d\in[\alpha_i,\alpha_{i+1}]}\left|\frac{\phi_R(x, c)_{k,r,s} - \phi_R(x, d)_{k,r,s}}{c - d}\right| \\ &\qquad\leqslant 2\,d_{r,s}\cdot m_\Delta(x, k, \mathcal{P}_{r,s}^{(i)}).\end{aligned} \tag{783}$$

We can thus upper bound (I) and (II) in (770) yielding a Lipschitz constant for $g_i^{k,r,s}$ and $g_{i+1}^{k,r,s}$ on $[\alpha_i, \alpha_{i+1}]$

$$\begin{aligned}&\max_{c,d\in[\alpha_i,\alpha_{i+1}]}\left|\frac{g_i^{k,r,s}(c) - g_i^{k,r,s}(d)}{c - d}\right| && (784)\\ &\qquad\leqslant 2\,d_{r,s}\cdot m_\Delta(x, k, \mathcal{P}_{r,s}^{(i)})\cdot \bar m(x, k, \mathcal{P}_{r,s}^{(i)}) \\ &\qquad= L_i^{k,r,s} (= L_{i+1}^{k,r,s}). && (785)\end{aligned}$$

Finally, we can compute L_r as

$$\begin{aligned}L &= \max_{1\leqslant i\leqslant N-1}\sum_{k=0}^{K-1}\sum_{(r,s)\in V} L_i^{k,r,s} \\ &= \max_{1\leqslant i\leqslant N-1}\sum_{k=0}^{K-1}\sum_{r,s\in V} 2d_{r,s}\cdot m_\Delta(x, k, \mathcal{P}_{r,s}^{(i)})\cdot \overline{m}(x, k, \mathcal{P}_{r,s}^{(i)}))\end{aligned} \tag{786}$$

what concludes the proof. □

D.4.3 *Scaling*

The scaling transformation is denoted as $\phi_S\colon \mathbb{R}^{K\times W\times H}\times\mathbb{R}\to\mathbb{R}^{K\times W\times H}$. Similar to rotations, ϕ_S acts on an image in three steps. First, it stretches height and width by a fixed ratio $\alpha\in\mathbb{R}$. Second, we determine missing pixel values with bilinear interpolation. Finally, we apply black padding to regions with missing pixel values if the image is scaled by a factor smaller than 1. Let c_W and c_H be the center pixels

$$c_W := \frac{W-1}{2}, \qquad c_H := \frac{H-1}{2}. \tag{787}$$

We notice that black padding is naturally applied during bilinear interpolation in cases where the scaling factor is smaller than 1 (that is, when we make images smaller). We can thus write the scaling operation as $\phi_S \colon \mathbb{R}^{K\times W\times H} \times \mathbb{R}_{>0} \to \mathbb{R}^{K\times W\times H}$, $(x, \alpha) \mapsto \phi(x, \alpha)$ where

$$\phi_S(x, \alpha)_{k,i,j} := Q_x\left(k, c_W + \frac{i - c_W}{\alpha}, c_H + \frac{j - c_H}{\alpha}\right). \tag{788}$$

When the scaling transformation in practice uses different padding mechanisms, we can simply apply black padding to the outer pixels during preprocessing. For example, if we know the semantic attacker could choose 0.7 as the smallest scaling ratio, we can apply black padding to all pixels that are out of canvas after 0.7 scaling. Therefore, we overwrite all different padding mechanisms and ensure the generalizability. As a trade-off, the classifier has a narrower reception field that affects the clean accuracy.

Due to black padding, the functions g_i (145) may contain discontinuities. To circumvent this issue, we enumerate all these discontinuities as $\mathcal{D}$. It can be shown that $\mathcal{D}$ contains at most $H + W$ elements. Hence, for large enough N, the interval $[\alpha_i, \alpha_{i+1}]$ contains at most one discontinuity. We thus modify the upper bounds M_i in (150) as

$$M_i := \begin{cases} \max\limits_{\alpha_i \leqslant \alpha \leqslant \alpha_{i+1}} \min\{g_i(\alpha), g_{i+1}(\alpha)\} & [\alpha_i, \alpha_{i+1}] \cap \mathcal{D} = \varnothing \\ \max\left\{ \max\limits_{\alpha_i \leqslant \alpha \leqslant t_i} g_{i+1}(\alpha), \max\limits_{t_i \leqslant \alpha \leqslant \alpha_{i+1}} g_i(\alpha) \right\} & [\alpha_i, \alpha_{i+1}] \cap \mathcal{D} = \{t_i\} \end{cases} \tag{789}$$

In either case, the quantity M_i can again be bounded by a Lipschitz constant. With this definition, the following lemma provides a closed form expression for the Lipschitz constant L in (150) for scaling.

Lemma 23. *Let $x \in \mathbb{R}^{K\times W\times H}$ be a K-channel image and let ϕ_S be the scaling transformation. Then, a global Lipschitz constant L for the functions $\{g_i\}_{i=1}^N$ is given by*

$$L_s = \max_{1 \leqslant i \leqslant N-1} \sum_{k=0}^{K-1} \sum_{r,s \in \Omega \cap \mathbb{N}^2} \frac{\sqrt{2} d_{r,s}}{a^2} \cdot m_\Delta(x, k, \mathcal{P}_{r,s}^{(i)}) \cdot \overline{m}(x, k, \mathcal{P}_{r,s}^{(i)}) \tag{790}$$

where $\Omega = [0, W-1] \times [0, H-1]$ and a is the lower boundary value in $S = [a, b]$. The set $\mathcal{P}_{r,s}^{(i)}$ is given by all integer grid pixels that are covered by the trajectory of source pixels of (r, s) when scaling with factors from α_{i+1} to α_i.

Proof. Recall the Definition of the Scaling transformation ϕ_S given by $\phi_S \colon \mathbb{R}^{K\times W\times H} \times \mathbb{R} \to \mathbb{R}^{K\times W\times H}$, where

$$\phi_S(x, \alpha)_{k,r,s} := Q_x\left(k, c_W + \frac{r - c_W}{s}, c_H + \frac{s - c_H}{s}\right). \tag{791}$$

Recall that the set Ω is given by $\Omega = [0, W-1] \times [0, H-1] = \{1, \ldots, K\}$ and let

$$\Omega_{\mathbb{N}} := \Omega \cap \mathbb{N}^2 \tag{792}$$

be the set of integers in Ω. Let $f_1^r \colon [a, b] \to \mathbb{R}$ and $f_2^{r,s} \colon [a, b] \to \mathbb{R}$ be functions defined as

$$\begin{aligned} f_1^r(\alpha) &:= c_W + \frac{r - c_W}{\alpha}, \\ f_2^s(\alpha) &:= c_H + \frac{s - c_H}{\alpha}. \end{aligned} \tag{793}$$

Then, the value of the scaled image $\phi_S(x, \alpha)$ is given by

$$\phi_S(x, \alpha)_{k,r,s} = Q_x(k, f_1^r(\alpha), f_2^s(\alpha)) \tag{794}$$

where Q_x denotes bilinear interpolation. Let

$$\psi_k \colon [a, b] \to \mathbb{R},\ \alpha \mapsto Q_x(k, f_1^r(\alpha), f_2^s(\alpha)). \tag{795}$$

We notice that, in contrast to rotations, ψ_k is *not* continuous at every $\alpha \in \mathbb{R}_{>0}$. Namely, when considering scaling factors in $(0, 1)$, bilinear interpolation applies black padding to some $(r, s) \in \Omega$ resulting in discontinuities of ψ_k. To see this, consider the following. The interval $[\alpha_{i+1}, \alpha_i]$ contains a discontinuity of ψ_k, if

$$\begin{cases} \alpha_{i+1} < \dfrac{r - c_W}{c_W} < \alpha_i, & r > c_W, \\ \alpha_{i+1} < \dfrac{c_W - r}{c_W} < \alpha_i, & r < c_W, \end{cases} \tag{796}$$

because then $\exists\, \alpha_0 \in [\alpha_{i+1}, \alpha_i]$ such that $f_1^r(\alpha_0) \in \{0, W-1\} \subseteq \Omega$ and hence

$$\phi_S(x, \alpha_0)_{k,r,s} \neq 0 \tag{797}$$

but, for $r > c_W$,

$$\phi_S(x, \alpha_0 + \varepsilon)_{k,r,s} = 0 \quad \forall \varepsilon > 0 \tag{798}$$

or, when $r < c_W$,

$$\phi_S(x, \alpha_0 - \varepsilon)_{k,r,s} = 0 \quad \forall \varepsilon > 0. \tag{799}$$

A similar reasoning leads to a discontinuity in the s-coordinates. We can thus define the set of discontinuities of ψ_k as

$$\mathcal{D} := \left(\bigcup_{r=0}^{W-1} \mathcal{D}_1^r \right) \cup \left(\bigcup_{s=0}^{H-1} \mathcal{D}_2^s \right) \tag{800}$$

where

$$\begin{aligned} \mathcal{D}_1^r &:= \{\alpha_0 \in [a, b] \mid f_1^r(\alpha_0) \in \{0, W-1\}\} \\ \mathcal{D}_2^s &:= \{\alpha_0 \in [a, b] \mid f_2^s(\alpha_0) \in \{0, H-1\}\}. \end{aligned} \tag{801}$$

We notice that $|\mathcal{D}| \leqslant H + W$ and hence for large enough N, each interval $[\alpha_i, \alpha_{i+1}]$ contains at most 1 discontinuity.

Due to these continuities, we need to modify the general upper bound M of the interpolation error M_S Recall that for $a < b$ and $\{\alpha_i\}_{i=1}^N$, the maximum L_2-sampling error $M_{a,b}$ is given by

$$M_S := \max_{a \leqslant \alpha \leqslant b} \min_{1 \leqslant i \leqslant N} \|\phi_S(x, \alpha) - \phi_S(x, \alpha_i)\|_2 . \tag{802}$$

In order to compute an upper bound on (802) for scaling, we are interested in finding $M \geqslant 0$ such that

$$M_S^2 \leqslant M \tag{803}$$

For scaling, similar as in the case for rotations, we sample α_i uniformly from $[a, b]$:

$$\alpha_i = a + \frac{b-a}{N-1}(i-1) \text{ for } 1 \leqslant i \leqslant N. \tag{804}$$

and note that $\alpha_1 = b$ and $\alpha_N = a$. For $1 \leqslant i \leqslant N$ Let g_i be the functions $g_i \colon [a, b] \to \mathbb{R}_{\geqslant 0}$ defined by

$$g_i(\alpha) := \|\phi_S(x, \alpha) - \phi_S(x, \alpha_i)\|_2^2 . \tag{805}$$

Note that $\forall\, \alpha \in [a, b]$, $\exists\, i$ such that $\alpha \in [\alpha_{i+1}, \alpha_i]$. Suppose that N is large enough such that $\forall\, i \colon |\mathcal{D} \cap [\alpha_{i+1}, \alpha_i]| \leqslant 1$ and denote the discontinuity in interval $[\alpha_{i+1}, \alpha_i]$ by t_i if it exists. Let

$$M_i := \begin{cases} \max\limits_{\alpha_i \leqslant \alpha \leqslant \alpha_{i+1}} \min\{g_i(\alpha), g_{i+1}(\alpha)\} & [\alpha_i, \alpha_{i+1}] \cap \mathcal{D} = \varnothing \\ \max\left\{ \max\limits_{\alpha_i \leqslant \alpha \leqslant t_i} g_{i+1}(\alpha), \max\limits_{t_i \leqslant \alpha \leqslant \alpha_{i+1}} g_i(\alpha) \right\} & [\alpha_i, \alpha_{i+1}] \cap \mathcal{D} = \{t_i\} \end{cases} \tag{806}$$

Similarly as in the case for rotations, we find

$$M_S^2 \leqslant \max_{1 \leqslant i \leqslant N-1} M_i. \tag{807}$$

For simplicity, we assume for the sequel that $\mathcal{D} = \varnothing$. The case where discontinuities exist can be treated analogously. We further divide each interval $[\alpha_i, \alpha_{i+1}]$ by sampling $n \in \mathbb{N}$ points $\{\gamma_{i,j}\}_{j=1}^n$ according to

$$\gamma_{i,j} := \alpha_i + \frac{\alpha_{i+1} - \alpha_i}{n-1}(j-1) \text{ for } 1 \leqslant j \leqslant n \tag{808}$$

and define

$$m_{i,j} := \max_{\gamma_{i,j} \leqslant \gamma \leqslant \gamma_{i,j+1}} \min\{g_i(\gamma), g_{i+1}(\gamma)\}. \tag{809}$$

We can thus upper bound each M_i by

$$M_i \leqslant \max_{1 \leqslant j \leqslant n-1} m_{i,j}. \tag{810}$$

In order to find an upper bound on M_S^2, we thus need to find an upper bound on $m_{i,j}$ and can proceed analogously to rotations. Namely, setting

$$M := \max_{1 \leqslant i \leqslant N-1} \left\{ \max_{1 \leqslant j \leqslant n-1} \left\{ \frac{1}{2} \cdot \big(\min\big\{ g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}), \; g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1}) \big\} \big) + L \cdot \frac{\gamma_{i,j+1} - \gamma_{i,j}}{2} \right\} \right\} \tag{811}$$

yields a computable upper bound of the maximum ℓ_2 interpolation error. Computing a Lipschitz constant for g_i and g_{i+1} is also analogous to rotations. The difference lies only in computing a Lipschitz constant for ϕ_S what we will explain in greater detail.

Recall that Lemma 22 provides a Lipschitz constant for the function $t \mapsto \psi_k(t) := Q_x(k, \rho_1(t), \rho_2(t))$ where ρ is a differentiable curve with values in $\mathbb{R}^2$. Namely, a Lipschitz constant for ψ_k is given by

$$L_k = \max_{t \in [t_1, t_2]} \left(\sqrt{2}\, \|\dot{\rho}(t)\|_2 \cdot m_\Delta(x, k, \lfloor \rho(t) \rfloor) \right). \tag{812}$$

Consider the curve

$$\rho(t) := (f_1^r(t),\, f_2^s(t)), \quad t > 0 \tag{813}$$

and note that it is differentiable with derivatives

$$\frac{df_1^r(t)}{dt} = \frac{d}{dt}\left(c_W + \frac{r - c_W}{t}\right) = \frac{c_W - r}{t^2} \tag{814}$$

$$\frac{df_2^s(t)}{dt} = \frac{d}{dt}\left(c_H + \frac{s - c_H}{t}\right) = \frac{c_H - s}{t^2} \tag{815}$$

and

$$\|\dot{\rho}(t)\|_2 = \frac{1}{t^2}\sqrt{(c_W - r)^2 + (c_H - s)^2}. \tag{816}$$

A Lipschitz constant for $\phi_S(x, \cdot)_{k,r,s}$ is thus given by

$$L_k^{r,s} = \max_{t\in[t_1,t_2]}\left(\frac{\sqrt{(c_W - r)^2 + (c_H - s)^2}}{t^2} \cdot \sqrt{2}\mathfrak{m}_\Delta(x, k, \lfloor\rho(t)\rfloor)\right) \tag{817}$$

$$\leqslant \frac{\sqrt{(c_W - r)^2 + (c_H - s)^2}}{t_1^2} \cdot \sqrt{2} \cdot \mathfrak{m}_\Delta(x, k, \mathcal{P}_{r,s}) \tag{818}$$

$$\leqslant \frac{\sqrt{(c_W - r)^2 + (c_H - s)^2}}{a^2} \cdot \sqrt{2} \cdot \mathfrak{m}_\Delta(x, k, \mathcal{P}_{r,s}) \tag{819}$$

where

$$\mathcal{P}_{r,s} = \bigcup_{\alpha\in[t_1,t_2]} \{(\lfloor f_1^r(t)\rfloor, \lfloor f_2^s(t)\rfloor)\}. \tag{820}$$

Finally, setting

$$L_i^{k,r,s} := L_k^{r,s} \cdot \bar{\mathfrak{m}}(x, k, \mathcal{P}_{r,s}^{(i)}) \tag{821}$$

and

$$\begin{aligned} L_s &= \max_{1\leqslant i\leqslant N-1} \sum_{k=0}^{K-1} \sum_{(r,s)\in\Omega_{\mathbb{N}}} L_i^{k,r,s} \\ &= \max_{1\leqslant i\leqslant N-1} \sum_{k=0}^{K-1} \sum_{r,s\in\Omega\cap\mathbb{N}^2} \frac{\sqrt{2}d_{r,s}}{a^2} \cdot \mathfrak{m}_\Delta(x, k, \mathcal{P}_{r,s}^{(i)}) \cdot \overline{\mathfrak{m}}(x, k, \mathcal{P}_{r,s}^{(i)}) \end{aligned} \tag{822}$$

yields the desired Lipschitz constant. □

D.4.4 *Discussion on More Transformations and Compositions*

Since the TSS framework is not limited to specific transformations, here we briefly discuss how to extend TSS to new transformations or to new compositions.

For a new transformation, we first identify the parameter space $\mathcal{Z}$, with the restriction that the parameter is required to completely and deterministically determine the output after the transformation for any given input. Then, we use 3 to check whether the

transformation is resolvable. If so, we can write down the function γ_α. In the next step, we choose a smoothing distribution, i.e., the distribution of the random variable ε_0, and identify the distribution of $\varepsilon_1 = \gamma_\alpha(\varepsilon_0)$. Finally, we use Theorem 5 to derive the robustness certificates and follow the two-step template outlined in Section 7.4.1.5 to compute the robustness certificate.

If the transformation is not resolvable, we identify a dimension in $\mathcal{Z}$ for which the transformation is *resolvable*. For example, the composition of rotation and brightness has a rotation and a brightness axis, where the brightness axis is itself resolvable. As a result, we can write the parameter space as Cartesian product of non-resolvable subspace and resolvable subspace: $\mathcal{Z} := \mathcal{Z}_{\text{no-resolve}} \times \mathcal{Z}_{\text{resolve}}$. We perform smoothing on the resolvable subspace and sample enough points in the non-resolvable subspace. Next, we bound the interpolation error between sampled points and arbitrary points in the non-resolvable subspace, using either ℓ_p difference as we did for rotation and scaling or other regimes. Specifically, Lemma 22 presented in Section D.4 is a useful tool to bound the ℓ_p difference stemming from interpolation errors. Finally, we instantiate 6 to compute the robustness certificate.

Theoretically, we can certify against the composition of all the discussed transformations: Gaussian blur, brightness, contrast, translation, rotation, and scaling. However, as justified in [92, Figure 3], the composition of more than two transformations leads to unrealistic images that are even hard to distinguish by humans. Moreover, if the composition contains too many transformations, the parameter space would no longer be low dimensional. Therefore, there would be much more axes that are differentially resolvable (instead of resolvable). As a consequence, much more samples are required to obtain a small bound on the interpolation error (which is necessary for a nontrivial robustness certification). Therefore, we focus on evaluating either single transformations, or the composition of two transformations to simulate a practical attack.

D.5 ALGORITHM FOR DIFFERENTIALLY RESOLVABLE TRANSFORMATIONS

In Algorithm 3 we present pseudo-code describing the computation of the interpolation error M, for rotations. This algorithm corresponds to to the description outlined in Section 7.5.2. In Algorithm 4 we present pseudo-code for progressive sampling, which was described in Section 7.5.4.2. We remark that in practice, we sample in mini-batches with batch size B. We set the error tolerance T to M_S (143) for rotation and scaling. For the composition of rotation or scaling with brightness within $[-b_0, b_0]$, then error tolerance T is set to

$$\sqrt{M_S^2 + \sigma^2/\sigma_b^2 \cdot b_0^2}. \tag{823}$$

Finally, for the composition of rotation or scaling, brightness change $[-b_0, b_0]$, and ℓ_2 bounded perturbations within r, the error tolerance T is set to

$$\sqrt{(M_S + r)^2 + \sigma^2/\sigma_b^2 \cdot b_0^2} \tag{824}$$

Jointly, these two algorithms make up the pipeline TSS-DR for certifying robustness against differentially resolvable transformations as shown in Figure 12a.

Algorithm 3 Interpolation Error M Computation for Rotation Transformation.

Input: clean input image x;
interval of rotation angle to certify $[a, b]$;
number of first-level samples N;
number of second-level samples n
Output: rotation angle samples $\{\alpha_i\}_{i=1}^{N}$;
upper bound M of squared ℓ_2-interpolation error

$$M_S^2 = \arg\max_{\alpha\in[a,b]} \min_{1\leqslant i\leqslant N} \|\tilde{\phi}_R(x,\alpha) - \tilde{\phi}_R(x,\alpha_i)\|_2^2.$$

/* *Compute Lipschitz constant* L_r (155) */
$\alpha_1 \leftarrow a$
for $i = 1,\dots,N-1$ **do**
  $\alpha_{i+1} \leftarrow a + (b-a)\cdot\frac{i}{N-1}$ (144)
  for all $(r,s) \in V$ **do**
    /* V *and* $\mathcal{P}_{r,s}^{(i)}$ *are defined in 8* */
    Compute trajectory covered grid pixels $\mathcal{P}_{r,s}^{(i)}$
    for $k = 0,\dots,K-1$ **do**
      Compute $2d_{r,s}\cdot m_\Delta(x,k,\mathcal{P}_{r,s}^{(i)})\cdot\bar{m}(x,k,\mathcal{P}_{r,s}^{(i)})$ (155)
    end for
  end for
  $L_{r,i} \leftarrow \sum_{k=0}^{K-1}\sum_{(r,s)\in V} 2d_{r,s}\cdot m_\Delta(x,k,\mathcal{P}_{r,s}^{(i)})\cdot\bar{m}(x,k,\mathcal{P}_{r,s}^{(i)})$.
end for
$L_r \leftarrow \max_{1\leqslant i\leqslant N-1} L_{r,i}$ (155)
/* *Compute interpolation error bound* M (147) *from stratified sampling* */
for $i = 1,\dots,N-1$ **do**
  for $j = 1,\dots,n$ **do**
    /* *Second-level sampling* */
    $\gamma_{i,j} \leftarrow \alpha_i + (\alpha_{i+1}-\alpha_i)\cdot\frac{j-1}{n-1}$ (148)
  end for
  $M_i \leftarrow 0$
  for $j = 1,\dots,n-1$ **do**
    Compute $g_i(\gamma_{i,j})$, $g_i(\gamma_{i,j+1})$, $g_{i+1}(\gamma_{i,j})$, and $g_{i+1}(\gamma_{i,j+1})$ (145)
    $M_i \leftarrow \max\{M_i, \min\{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}),$
      $g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})\}\}$
  end for
  $M_i \leftarrow \frac{1}{2}M_i + L\cdot\frac{b-a}{(N-1)(n-1)}$ (150)
end for
Return: $M \leftarrow \max_{1\leqslant i\leqslant N-1} M_i$ (147)

Algorithm 4 Progressive Sampling for Certification.

Input: clean input image x with true class k_A; first-level parameter samples $\{\alpha_i\}_{i=1}^{N}$; perturbation random variable ε with variance σ^2; ℓ_2 error tolerance T; batch size B; sampling size limit n_s; confidence level p.
Output: with probability $1-p$, whether $g(\cdot;\varepsilon)$ is certifiably robust at $\phi(x, \alpha)$.
for $i = 1, \ldots, N$ **do**
  $x^{(i)} \leftarrow \phi(x, \alpha_i)$
  $j \leftarrow 0$
  while $j \leqslant n_s$ **do**
    Sample B instances of $\phi(x^{(i)}, \varepsilon)$, and use them to update empirical mean $\hat{q}(y_A|x^{(i)}; \varepsilon)$.
    $j \leftarrow j + B$.
    /* *Lower confidence interval bound with these* j *samples* */
    $\underline{p_A}^{(i)} = \text{LowerConfBound}(\hat{q}(y_A|x^{(i)}; \varepsilon), j, 1-p/N)$.
    if $\underline{R_i} = \sigma\Phi^{-1}\left(\underline{p_A}^{(i)}\right) > T$ **then**
      /* *Already get the certification that* $R_i > T$, *break* */
      `Break`
    end if
  end while
  if $\underline{R_i} = \sigma\Phi^{-1}\left(\underline{p_A}^{(i)}\right) \leqslant T$ **then**
    /* *Cannot ensure that* $R_i > T$. *So cannot ensure that* $R = \min R_i > T$. *Early halt* */
    Return: `false`
  end if
end for
Return: `true`

D.6 ADDITIONAL DETAILS ABOUT EXPERIMENTS

Here we provide additional details about the experiment setup, implementations, baselines, evaluation protocols, results, findings, and analyses.

D.6.1 *Model Preparation and Hyperparameters*

As previous work shows, an undefended model is very vulnerable even under simple random semantic attacks. Therefore, to obtain nontrivial certified robustness, we require the model itself to be trained to be robust against semantic transformations. We apply data augmentation training [39] combined with Consistency regularization [106] to train the base classifiers. The data augmentation training randomly transforms the with the specified transformation using parameters drawn from the specified smoothing distribution/strategy. Consistency regularization further enhances the consistency of the base classifiers' prediction among the drawn parameters. Then, the base classifiers are used to construct smoothed classifiers by the specified smoothing distribution/strategy, and we compute its robustness certification with our approach.

On the relatively small datasets MNIST and CIFAR-10, the models are trained from scratch. On MNIST, we use a Convolutional Neural Network (CNN) composed of four convolutional layers and three fully connected layers. On CIFAR-10, we use ResNet-110, a 110-layer ResNet model [85]. These model architectures are the same as in the literature [39, 187, 264], enabling a direct comparison. On MNIST, we train for 100 epochs; on CIFAR-10, we train for 150 epochs. The batch sizes (B) are 400 and 256 on MNIST and CIFAR-10, respectively. The learning rate on both datasets is initialized to 0.01, and

after every 50 epochs, the learning rate is multiplied by 0.1. For resolvable transformations, the data augmentation usually uses the same smoothing distribution/strategy as we will use to construct the smoothed classifier. In particular, for brightness and contrast transformation, we empirically observe that a larger variance during inference time helps to improve the certified accuracy under large attack radius. For the composition of Gaussian blur, brightness, contrast, and translation, we additionally add small additive Gaussian noise to improve the ability to defend against other unforeseen attacks. For differentially resolvable transformations, since Gaussian noise is required in constructing the smoothed classifier, the data augmentation jointly combines additive Gaussian noise and the transformation to certify against. The detailed hyperparameters such as distribution type and variance are listed in Table 19 and Table 20. The weight of Consistency regularization is set to 10 throughout the training.

On the large ImageNet dataset, we finetune the existing trained models. For resolvable transformations, we finetune from the ResNet-50 model available in the `torchvision` library [176]. For differentially resolvable transformations, since the base classifier should also be robust under Gaussian noise, we finetune from the Resnet-50 model in [187] that achieves state-of-the-art robustness under Gaussian noise. In either case, we follow the same data augmentation scheme as on MNIST and CIFAR-10, and we finetune for two epochs with batch size (B) 128, learning rate 0.001, and Consistency regularization weight 10. During certification (e.g., Algorithm 4), we use the same batch sizes as during training on these datasets.

Channel-wise normalization is used for all models on these three datasets as in [39, 187]. On all three datasets, in each training epoch, we feed the entire training dataset without random shuffling.

We remark that, since our approach focuses on robustness certification and the smoothing strategy to improve certified robustness, we did not fully explore the potential of improving certified robustness from the training side, nor did we conduct any hyperparameter tuning. Therefore, even though we already achieved the state of the art using our robustness certification and smoothing strategies, we believe the results can potentially be further improved by more effective training approaches.

D.6.2 *Implementation Details*

Our implementation, including the training scripts is based on `PyTorch`. For resolvable transformations, we extend the smoothing module from [39] to accommodate various smoothing strategies and smoothing distributions. The predict and certify modules are kept the same. For differentially resolvable transformations, since the stratified sampling requires $N \times n$ transformations to compute the interpolation error bound (where N is the number of first-level samples and n the number of second-level samples), we implement a fast `C` module and integrate it in our `Python`-based tool. Empirically, this implementation achieves a $3-5$x speedup compared to the `OpenCV`[166]-based transformations. For Lipschitz upper bound computation, since the loop in `Python` is slow, we reformulate the computation using loop-free tensor computations in `numpy`. This empirically achieves $20-40$x speedups compared to the plain loop-based implementation. The full implementation of TSS along with all trained models is publicly available at `https://github.com/AI-secure/semantic-randomized-smoothing`.

D.6.3 *Attack Details*

To evaluate the empirical robust accuracy of both TSS models and undefended models we use Random Attack, Random+ Attack, and PGD Attack. The Random Attack is used in previous work [8, 65] but does not consider the intrinsic characteristics of semantic transformations. Thus, we propose Random+ Attack and PGD Attack as alternatives since they are adaptively designed for our smoothed TSS models and also consider the intrinsic characteristics of these transformations.

D.6.3.1 *Random Attack*

The random attack is used to evaluate the empirical robust accuracy, which is an upper bound of the certified robust accuracy. The random attack reads the clean input, and uniformly samples N parameters from the predefined transformation parameter space to transform the input. If the model makes a wrong prediction on any of these N transformed inputs, we treat this sample as being successfully attacked; otherwise, the sample counts toward the empirical robust accuracy. We denote by N the "number of initial starts". In the main experiments, we set $N = 100$, while in the ablation study presented in Section D.6.5.3, we also compare the behaviors of the three attacks under $N = 10/20/50$.

For transformations with a hyper-rectangle parameter space, including brightness, contrast, scaling, rotation, Gaussian blur, and their compositions, we uniformly sample transformation parameters for each coordinate. For transformations with a discrete parameter space, such as translation, we draw the parameter with equal probability. When the transformation is composed with ℓ_p-bounded perturbations, we additionally generate the perturbation vector using FGSM attack [212], where the precise gradient is used for vanilla models, and the empirical mean gradient over 100 samples is used for smoothed TSS models.

D.6.3.2 *Random+ adaptive Attack*

The Random+ attack follows the same procedure as the Random attack, with the difference that, instead of using uniform distribution for sampling transformation parameters, we use the Beta distribution $\mathrm{Beta}(0.5, 0.5)$.

Formally, suppose the transformation space is $[a, b]$. In the Random attack, the attack parameter ε is generated according to

$$\varepsilon' \sim \mathrm{Unif}(0, 1), \quad \varepsilon \leftarrow a + (b - a) \times \varepsilon'. \tag{825}$$

In contrast, in the Random+ attack, we generate the attack parameter δ randomly according to

$$\delta' \sim \mathrm{Beta}(0.5, 0.5), \quad \delta \leftarrow a + (b - a) \times \delta'. \tag{826}$$

We choose the Beta distribution because, intuitively, an adversarial example would be more likely to exist at the boundary, i.e., closer to a or b, making the attack more powerful. For example, suppose that a rotation attacker can use angles in $[-r, r]$. Then, successful adversarial examples are more likely to have large rotation angle. As shown in Figure 32, the Beta distribution assigns more probability mass on to the boundaries,

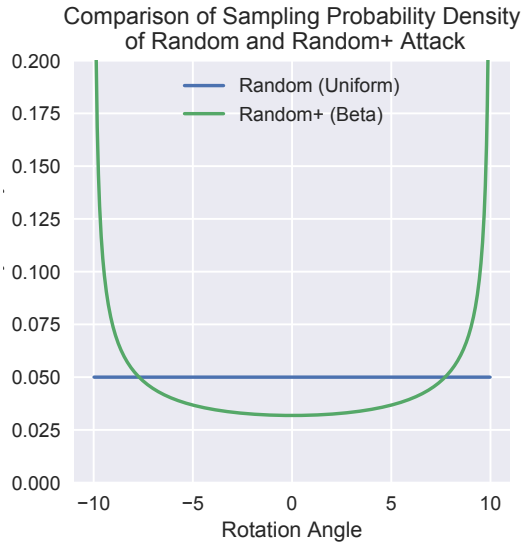


Figure 32: Comparison of probability density of Random and Random+ attack when attacking the rotation transformation with rotation angle between $-10°$ and $+10°$.

compared to the uniform distribution. Choosing other Beta distribution hyperparameters could control the trade-off on sampling weights over the boundary or over center, and we empirically find $\text{Beta}(0.5, 0.5)$ already works very well as shown in additional experiments in Section D.6.5.3.

D.6.3.3 *PGD adaptive Attack*

To get a more powerful attack, here we propose the semantic transformation version of PGD attack in the following procedure.

(1) Initialize the transformation parameter following the same process as in the Random+ Attack.

(2) Suppose the current parameter is $(\alpha_1, \cdots, \alpha_z)$. The attack slightly perturbs each coordinate from α_i to $\alpha_i \pm \tau_i$ where $\tau_i = l_i/10$ and l_i is the length of the specified interval on the i-th coordinate. This yields $2z$ perturbed candidates.

(3) Clip each coordinate to be within the specified range, and choose the candidate that yields the largest increase in cross-entropy loss (for vanilla models) or empirical mean cross-entropy loss (for TSS models) to update the current parameter.

(4) Repeat steps (2) and (3) for 10 iterations for each sample obtained in step (1).

Finally, if the transformation is composed with ℓ_p-bounded perturbations, we additionally generate the perturbation vector in the same way as in Random attack.

Note that, for each sample generated in Step (1) we would have one output and resulting in a total of N outputs. If any of these outputs fool the target model, we treat this sample as being successfully attacked; otherwise, the sample counts toward the empirical robust accuracy. Note that since the translation transformation has a discrete parameter space, the PGD attack is not applicable.

We refer to the attack as the semantic transformation version of PGD attack because: (1) It involves multiple initial starts; (2) It leverages the local landscape information

to maximize the loss function iteratively. (3) It clips (i.e., "projects") the parameters to be within the perturbation range. Compared to the classical PGD attack under ℓ_p-norm constraint, we use coordinate-wise perturbations to probe the local landscape to circumvent the hardness of obtaining the gradient with respect to transformation parameters.

D.6.4 *Baseline Details*

Here we provide additional about the baselines used in the comparison with TSS.

DeepG [8] is based on linear relaxations. The code is open-sourced, and we utilize it to provide a direct comparison. The code provides trained models on MNIST and CIFAR-10, while on ImageNet the method is too slow and memory-consuming to run. On both MNIST and CIFAR-10, we use the provided trained models from the code. In terms of computation time, since our approach uses far less than 1000 s for certification per input on MNIST and CIFAR-10, we tune the hyperparameters to let the code spend roughly 1000 s for the certification.

Interval [202] is based on interval bound propagation. We also utilize the open-source code to provide a direct comparison. The settings are the same as in DeepG.

VeriVis [171] provides an enumeration-based solution when the number of possible transformation parameters or the number of possible transformed images is finite. In our evaluation, only translation satisfies this property. Therefore, as the baseline, we implement the enumeration-based robustness certification algorithm for our trained robust models.

Semantify-NN [156] proposes to insert a preprocessing layer with the goal of reducing the problem of verification against semantic transformations to the problem of verification against classical ℓ_p noise. To our knowledge, the code has not been open-sourced yet. Therefore, we directly compare with the numbers reported in their paper. Since they report the average of certified robust magnitude, we apply Markov's inequality to obtain an upper bound of their certified robust accuracy. For example, they report 46.24 degrees as the average certified robust rotation angles. This means that $\mathbb{P}[r \geqslant 50^\circ] \leqslant \mathbb{E}[r]/50 = 92.48\%$, i.e., the certified robust accuracy is no larger than 92.48% when fixing the rotation angle to be 50°. For brightness and contrast changes, Semantify-NN considers first applying the change and then clipping to $[0, 1]$, while our TSS considers only brightness and contrast changes. This makes a one-to-one comparison with [156] difficult, but since other baselines (e.g., [8]) consider the same setting as we do, and to align with most baselines, we slightly sacrifice comparability in this special case. For interested readers who would like to have an absolutely fair comparison with Semantify-NN on brightness and contrast changes, they can extend our TSS by modeling Semantify-NN's tranformation by $\phi_{BC}(x, (b, c)) \circ \phi_{clip}(x, t_l, t_h)$, where ϕ_{clip} clips the pixel intensities lower than t_l and higher than t_h. Applying TSS-R on transformation parameters (b, c, t_l, t_h) then derives the robustness certification under the same threat model as Semantify-NN.

DistSPT [65] combines randomized smoothing and interval bound propagation to provide certified robustness against semantic transformations. Concretely, the approach leverages interval bound propagation to compute the upper bound of interpolation error and then applies randomized smoothing. On small datasets such as MNIST and CIFAR-10, the approach is able to provide nontrivial robustness certification. Neverthe-

less, the certified robust accuracy is lower than TSS as shown in Table 5. We use their reported numbers in [65, Table 4] for DistSPT^x for comparison, since the certification goal and evaluation protocol are the same as ours. On ImageNet, as described in [65, Section 7.4], the interval bound propagation is computationally expensive and loose. Therefore, they use sampling to estimate the interpolation error, which makes the robustness certification no longer hold against arbitrary attacks but just a certain random attack ("worst-of-10" attack).

IndivSPT [65] provides a different certification goal from the above approaches. At a high level, the approach uses a transformed image as the input where the transformation parameter is within a predefined threshold. Then the approach certifies whether the prediction for the transformed image and the prediction for the original image are the same. In contrast, TSS and other baseline approaches take the original image as the input and certify whether there exists no transformed image that can mislead the model. Due to these different certification goals, TSS is not comparable with IndivSPT.

D.6.5 *Additional Results*

D.6.5.1 *Benign Accuracy*

In Table 18 we show the benign accuracy of our models corresponding to Table 5. For comparison, the vanilla trained models have benign accuracy 98.6% on MNIST, 88.6% on CIFAR-10, and 74.4% on ImageNet. We observe that, even though the trade-off between accuracy and (certified) robustness is widely reported both theoretically [154, 264, 272] and empirically (e.g., [39, 106, 271]) for the classical ℓ_p threat models, this trade-off does not always exist in the semantic defense setting. Specifically, for resolvable transformations, we do not observe an apparent loss of benign accuracy for our certifiably robust models; while for differentially resolvable transformations (i. e., those involving scaling and rotation), there is no loss in accuracy on MNIST, slight losses on CIFAR-10, and apparent losses on ImageNet. In cases where there does exist a trade-off between benign accuracy and certified robust accuracy, we show in Section D.6.5.6 that this is largely controlled by the smoothing variance.

D.6.5.2 *Smoothing Distributions and Running Time Statistics*

In Table 19 and Table 20, we present the smoothing distributions with specific parameters and average certification computing time per sample. This accompanies the results present in Table 5. In the tables, α corresponds to the squared kernel radius for Gaussian blur; Δx and Δy correspond to translation displacement on horizontal and vertical direction; b and c are for brightness shift and contrast change respectively as in $x \mapsto (1+c)x + b$; r is for rotation angle; s is for size scaling ratio; ε is for additive noise vector; and $\|\delta\|_2$ for ℓ_2 norm of permitted additional perturbations. Specifically, "Training Distribution" corresponds to the distributions used for data augmentation during training of the base classifiers; "Smoothing Distribution" on the other hand stands for the distributions used for the smoothed classifiers during certification. We select these distributions according to the principles outlined in Section D.6.1.

Table 18: Benign accuracy of our TSS models corresponding to those in Table 5. Certified robust accuracy shown as reference.

Transformation	Dataset	Attack Radius	Certified Robust Accuracy	Benign Accuracy
Gaussian Blur	MNIST	Squared Radius $\alpha \leqslant 36$	90.6%	96.8%
	CIFAR-10	Squared Radius $\alpha \leqslant 16$	63.6%	76.2%
	ImageNet	Squared Radius $\alpha \leqslant 36$	51.6%	59.2%
Translation (Reflection Pad.)	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 8$	99.6%	99.6%
	CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 20$	80.8%	87.0%
	ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 100$	50.0%	73.0%
Brightness	MNIST	$b \pm 50\%$	98.2%	98.2%
	CIFAR-10	$b \pm 40\%$	87.0%	87.8%
	ImageNet	$b \pm 40\%$	70.0%	72.2%
Contrast and Brightness	MNIST	$c \pm 50\%, b \pm 50\%$	97.6%	98.0%
	CIFAR-10	$c \pm 40\%, b \pm 40\%$	82.4%	86.8%
	ImageNet	$c \pm 40\%, b \pm 40\%$	61.4%	72.2%
Gaussian Blur, Translation, Brightness, and Contrast	MNIST	$\alpha \leqslant 1, c, b \pm 10\%, \sqrt{\Delta x^2 + \Delta y^2} \leqslant 5$	90.2%	98.2%
	CIFAR-10	$\alpha \leqslant 1, c, b \pm 10\%, \sqrt{\Delta x^2 + \Delta y^2} \leqslant 5$	58.2%	77.6%
	ImageNet	$\alpha \leqslant 10, c, b \pm 20\%, \sqrt{\Delta x^2 + \Delta y^2} \leqslant 10$	32.8%	61.6%
Rotation	MNIST	$r \pm 50°$	97.4%	99.4%
	CIFAR-10	$r \pm 10°$	70.6%	83.2%
		$r \pm 30°$	63.6%	82.6%
	ImageNet	$r \pm 30°$	30.4%	46.2%
Scaling	MNIST	$s \pm 30\%$	99.0%	99.4%
	CIFAR-10	$s \pm 30\%$	58.8%	79.8%
	ImageNet	$s \pm 30\%$	26.4%	50.8%
Rotation and Brightness	MNIST	$r \pm 50°, b \pm 20\%$	97.0%	99.4%
	CIFAR-10	$r \pm 10°, b \pm 10\%$	70.2%	83.0%
		$r \pm 30°, b \pm 20\%$	61.4%	82.6%
	ImageNet	$r \pm 30°, b \pm 20\%$	26.8%	45.8%
Scaling and Brightness	MNIST	$s \pm 50\%, b \pm 50\%$	96.6%	99.4%
	CIFAR-10	$s \pm 30\%, b \pm 30\%$	54.2%	79.6%
	ImageNet	$s \pm 30\%, b \pm 30\%$	23.4%	50.8%
Rotation, Brightness, and ℓ_2	MNIST	$r \pm 50°, b \pm 20\%, \|\delta\|_2 \leqslant .05$	96.6%	99.4%
	CIFAR-10	$r \pm 10°, b \pm 10\%, \|\delta\|_2 \leqslant .05$	64.2%	83.0%
		$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leqslant .05$	55.2%	82.6%
	ImageNet	$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leqslant .05$	26.6%	45.8%
Scaling, Brightness, and ℓ_2	MNIST	$s \pm 50\%, b \pm 50\%, \|\delta\|_2 \leqslant .05$	96.4%	99.4%
	CIFAR-10	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leqslant .05$	51.2%	79.6%
	ImageNet	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leqslant .05$	22.6%	50.8%

Table 19: Detailed smoothing distributions and running time statistics for TSS. $\mathcal{N}(\mu, \Sigma)$ is the normal distribution, $\exp(\lambda)$ is the exponential distribution, $\mathcal{U}([a, b])$ is the uniform distribution. The random variable ϵ is elementwise noise as in Corollary 4. "Cert." means certification.

Transformation	Dataset	Attack Radius	Training Distribution	Smoothing Distribution	Avg. Cert. Time per Sample
Gaussian Blur	MNIST	Squared Radius $\alpha \leqslant 36$	$\alpha \sim \text{Exp}(1/10)$		7.9 s
	CIFAR-10	Squared Radius $\alpha \leqslant 16$	$\alpha \sim \text{Exp}(1/5)$		30.9 s
	ImageNet	Squared Radius $\alpha \leqslant 36$	$\alpha \sim \text{Exp}(1/10)$		45.7 s
Translation (Reflection Pad.)	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 8$	$(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$		10.2 s
	CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 20$	$(\Delta x, \Delta y) \sim \mathcal{N}(0, 15^2 I)$		39.4 s
	ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 100$	$(\Delta x, \Delta y) \sim \mathcal{N}(0, 30^2 I)$		161.9 s
Brightness	MNIST	$b \pm 50\%$	$b \sim \mathcal{N}(0, 0.6^2)$		2.1 s
	CIFAR-10	$b \pm 40\%$	$b \sim \mathcal{N}(0, 0.3^2)$		4.4 s
	ImageNet	$b \pm 40\%$	$b \sim \mathcal{N}(0, 0.4^2)$		45.1 s
Contrast and Brightness	MNIST	$c \pm 50\%, b \pm 50\%$	$(c, b) \sim \mathcal{N}(0, 0.6^2 I)$	$(c, b) \sim \mathcal{N}(0, 1.0^2 I)$	9.8 s
	CIFAR-10	$c \pm 40\%, b \pm 40\%$	$(c, b) \sim \mathcal{N}(0, 0.4^2 I)$	$(c, b) \sim \mathcal{N}(0, 0.6^2 I)$	45.0 s
	ImageNet	$c \pm 40\%, b \pm 40\%$	$(c, b) \sim \mathcal{N}(0, 0.4^2 I)$		325.6 s
Gaussian Blur, Translation, Brightness, and Contrast	MNIST	$\alpha \leqslant 5, c, b \pm 10\%,$ $\sqrt{\Delta x^2 + \Delta y^2} \leqslant 5$	$\alpha \sim \text{Exp}(1/10)$ $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$ $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$ $\epsilon \sim \mathcal{N}(0, 0.05^2 I)$	$\alpha \sim \text{Exp}(1/10)$ $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$ $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$	12.9 s
	CIFAR-10	$\alpha \leqslant 1, c, b \pm 10\%,$ $\sqrt{\Delta x^2 + \Delta y^2} \leqslant 5$	$\alpha \sim \text{Exp}(1)$ $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$ $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$ $\epsilon \sim \mathcal{N}(0, 0.01^2 I)$	$\alpha \sim \text{Exp}(1)$ $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$ $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$	43.1 s
	ImageNet	$\alpha \leqslant 10, c, b \pm 20\%,$ $\sqrt{\Delta x^2 + \Delta y^2} \leqslant 10$	$\alpha \sim \text{Exp}(1/5)$ $(\Delta x, \Delta y) \sim \mathcal{N}(0, 20^2 I)$ $(c, b) \sim \mathcal{N}(0, 0.4^2 I)$ $\epsilon \sim \mathcal{N}(0, 0.01^2 I)$	$\alpha \sim \text{Exp}(1/5)$ $(\Delta x, \Delta y) \sim \mathcal{N}(0, 20^2 I)$ $(c, b) \sim \mathcal{N}(0, 0.4^2 I)$	238.1 s

Table 20: Detailed smoothing distributions and running time statistics for TSS. $\mathcal{N}(\mu, \Sigma)$ is the normal distribution, $\exp(\lambda)$ is the exponential distribution, $\mathcal{U}([a, b])$ is the uniform distribution. The random variable ϵ is elementwise noise as in Corollary 4. "Cert." means certification.

Transformation	Dataset	Attack Radius	Training Distribution	Smoothing Distribution	Avg. Cert. Time per Sample
Rotation and Brightness	MNIST	$r \pm 50°, b \pm 20\%$	$r \sim \mathcal{U}([-55, 55])$ $\epsilon \sim \mathcal{N}(0, 0.12^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	$\epsilon \sim \mathcal{N}(0, 0.12^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	31.4 s
	CIFAR-10	$r \pm 10°, b \pm 10\%$	$r \sim \mathcal{U}([-12.5, 12.5])$ $\epsilon \sim \mathcal{N}(0, 0.05^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	$\epsilon \sim \mathcal{N}(0, 0.05^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	62.3 s
		$r \pm 30°, b \pm 20\%$	$r \sim \mathcal{U}([-35, 35])$ $\epsilon \sim \mathcal{N}(0, 0.05^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	$\epsilon \sim \mathcal{N}(0, 0.05^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	157.0 s
	ImageNet	$r \pm 30°, b \pm 20\%$	$r \sim \mathcal{U}([-35, 35])$ $\epsilon \sim \mathcal{N}(0, 0.5^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	$\epsilon \sim \mathcal{N}(0, 0.5^2 I)$ $b \sim \mathcal{N}(0, 0.2^2)$	2475.6 s
Scaling and Brightness	MNIST	$s \pm 50\%, b \pm 50\%$	$s \sim \mathcal{U}([0.45, 1.55])$ $\epsilon \sim \mathcal{N}(0, 0.12^2 I)$ $b \sim \mathcal{N}(0, 0.5^2)$	$\epsilon \sim \mathcal{N}(0, 0.12^2 I)$ $b \sim \mathcal{N}(0, 0.5^2)$	74.9 s
	CIFAR-10	$s \pm 30\%, b \pm 30\%$	$s \sim \mathcal{U}([0.65, 1.35])$ $\epsilon \sim \mathcal{N}(0, 0.12^2 I)$ $b \sim \mathcal{N}(0, 0.3^2)$	$\epsilon \sim \mathcal{N}(0, 0.12^2 I)$ $b \sim \mathcal{N}(0, 0.3^2)$	44.5 s
	ImageNet	$s \pm 30\%, b \pm 30\%$	$s \sim \mathcal{U}([0.65, 1.35])$ $\epsilon \sim \mathcal{N}(0, 0.5^2 I)$ $b \sim \mathcal{N}(0, 0.3^2)$	$\epsilon \sim \mathcal{N}(0, 0.5^2 I)$ $b \sim \mathcal{N}(0, 0.3^2)$	1401.6 s
Rotation	MNIST	$r \pm 50°$	Same as Rotation and Brightness	$\epsilon \sim \mathcal{N}(0, 0.12^2 I)$	20.1 s
	CIFAR-10	$r \pm 10°$		$\epsilon \sim \mathcal{N}(0, 0.05^2 I)$	52.8 s
		$r \pm 30°$		$\epsilon \sim \mathcal{N}(0, 0.05^2 I)$	141.0 s
	ImageNet	$r \pm 30°$		$\epsilon \sim \mathcal{N}(0, 0.5^2 I)$	2358.1 s
Scaling	MNIST	$s \pm 30\%$	Same as Scaling and Brightness	$\epsilon \sim \mathcal{N}(0, 0.12^2 I)$	17.7 s
	CIFAR-10	$s \pm 30\%$		$\epsilon \sim \mathcal{N}(0, 0.12^2 I)$	42.2 s
	ImageNet	$s \pm 30\%$		$\epsilon \sim \mathcal{N}(0, 0.5^2 I)$	1201.2 s
Rotation, Brightness, and ℓ_2	MNIST	$r \pm 50°, b \pm 20\%, \|\delta\|_2 \leqslant .05$	Same as Rotation and Brightness	Same as Rotation and Brightness	35.1 s
	CIFAR-10	$r \pm 10°, b \pm 10\%, \|\delta\|_2 \leqslant .05$			132.5 s
		$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leqslant .05$			520.2 s
	ImageNet	$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leqslant .05$			3463.8 s
Scaling, Brightness, and ℓ_2	MNIST	$s \pm 50\%, b \pm 50\%, \|\delta\|_2 \leqslant .05$	Same as Scaling and Brightness	Same as Scaling and Brightness	75.1 s
	CIFAR-10	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leqslant .05$			50.0 s
	ImageNet	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leqslant .05$			1657.7 s

D.6.5.3 *Comparison of Random Attack and Adaptive Attacks*

In Table 5, we compare the empirical robust accuracy of vanilla models and TSS models under random attacks and two adaptive attacks: Random+ and PGD. However, in the main table, we have omitted the empirical accuracy of each adaptive attack and only presented the minimum empirical accuracy among them. In Table 21, Table 22, Table 23, Table 24, and Table 25, Table 26, we present the detailed empirical accuracy for each attack under different number of initial starts N = 10, 20, 50, and 100. The results are shown for all transformations on MNIST, CIFAR-10, and ImageNet. Note that Table 5 shows the empirical accuracy with N = 100 for all attacks. From these three tables, we cross-validate the findings shown in the main part of this thesis. Namely, the adaptive attack decreases the empirical accuracy of TSS models slightly, while it decreases that of vanilla models more. Moreover, when comparing these three attacks, we find that with a small number of initial starts (e.g., N = 10), the PGD attack is typically the most powerful. However, with a large number of initial starts (e.g., N = 100), the Random+ attack sometimes becomes better. We conjecture that the optimization goal of the PGD attack—maximization of cross-entropy loss—might be sub-optimal in terms of increasing the misclassification rate. Thus, with a small number of initial starts, PGD is better than Random/Random+ attack due to the iterative ascending. However, with a large number of initial starts, both PGD and Random+ attack can sufficiently explore the adversarial region, and PGD may be misled by the optimization goal to a benign region. We believe it is interesting future work to study these intriguing properties of semantic attacks.

D.6.5.4 *Empirical Robustness on CIFAR-10-C and ImageNet-C*

In Section 7.6.2.4, we have shown that TSS generalizes to defend against unknown physical attacks by evaluating on CIFAR-10-C and ImageNet-C. Here, we first introduce the detailed evaluation protocol, then a breakdown of the result in Table 6 and show the empirical accuracy on each type of corruption.

EVALUATED MODELS On either CIFAR-10-C and ImageNet-C, we choose three models for evaluation: the vanilla model, the AugMix [92] trained model, and our TSS model for defending the composition of Gaussian blur, translation, brightness, and contrast. The vanilla models and TSS models are the same models as used in the main experiments. AugMix is the state-of-the-art empirical defense on the CIFAR-10-C and ImageNet-C dataset according to [43]. For AugMix, on CIFAR-10-C, since the model weights are not released, we use the official implementation of AugMix [1] and extend the code to support our used model architecture (ResNet-110) for a fair comparison. We run the code with the suggested hyperparameters and achieve similar performance as reported in their paper. On ImageNet-C, we directly use the officially released model weights. The model has the same architecture (ResNet-50) as ours so the comparison is naturally fair. Note that all these models are trained only on the clean CIFAR-10 or ImageNet training set.

EMPIRICAL ACCURACY COMPUTATION We compute the empirical accuracy (on CIFAR-10-C/ImageNet-C) as the ratio of correctly predicted samples among the test

1 See https://github.com/google-research/augmix

Table 21: Comparison between empirical robust accuracy against random and adaptive attacks and certified robust accuracy on MNIST. The attack radii are consistent with Table 5. The most powerful attack in each setting is highlighted in bold font. Random+ and PGD attacks are adaptive. Note that PGD attack cannot apply to translation transformation because the parameter space is discrete.

Transformation	Attack Radius	Model	Attack	Empirical Robust Accuracy				Certified robust Accuracy
				N = 10	N = 20	N = 50	N = 100	
Gaussian Blur	Squared Radius $\alpha \leqslant 36$	**TSS**	Random	93.2%	92.2%	92.0%	91.4%	90.6%
			Random+	92.4%	92.2%	**91.2%**	**91.2%**	
			PGD	**91.6%**	**91.6%**	91.6%	91.6%	
		Vanilla	Random	14.0%	12.4%	**12.2%**	**12.2%**	-
			Random+	12.4%	12.4%	**12.2%**	**12.2%**	
			PGD	**12.2%**	**12.2%**	**12.2%**	**12.2%**	
Translation (Reflection Pad.)	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 8$	**TSS**	Random	**99.6%**	**99.6%**	**99.6%**	**99.6%**	99.6%
			Random+	**99.6%**	**99.6%**	**99.6%**	**99.6%**	
			PGD	-	-	-	-	
		Vanilla	Random	**0.0%**	**0.0%**	**0.0%**	**0.0%**	-
			Random+	**0.0%**	**0.0%**	**0.0%**	**0.0%**	
			PGD	-	-	-	-	
Brightness	$b \pm 50\%$	**TSS**	Random	**98.2%**	**98.2%**	**98.2%**	**98.2%**	98.2%
			Random+	**98.2%**	**98.2%**	**98.2%**	**98.2%**	
			PGD	**98.2%**	**98.2%**	**98.2%**	**98.2%**	
		Vanilla	Random	97.2%	**96.6%**	**96.6%**	**96.6%**	-
			Random+	96.8%	**96.6%**	**96.6%**	**96.6%**	
			PGD	**96.6%**	**96.6%**	**96.6%**	**96.6%**	
Contrast and Brightness	$c \pm 50\%$, $b \pm 50\%$	**TSS**	Random	**98.0%**	**98.0%**	**98.0%**	**98.0%**	97.6%
			Random+	**98.0%**	**98.0%**	**98.0%**	**98.0%**	
			PGD	**98.0%**	**98.0%**	**98.0%**	**98.0%**	
		Vanilla	Random	96.8%	95.8%	95.0%	94.6%	-
			Random+	95.8%	94.4%	93.8%	93.6%	
			PGD	**93.6%**	**93.4%**	**93.2%**	**93.2%**	
Gaussian Blur, Translation Contrast, and Brightness	$\alpha \leqslant 5$, $c \pm 10\%$, $b \pm 10\%$, $\sqrt{\Delta x^2 + \Delta y^2} \leqslant 5$	**TSS**	Random	97.6%	97.6%	97.6%	97.2%	90.2%
			Random+	97.6%	**97.2%**	**97.0%**	**97.0%**	
			PGD	**97.4%**	97.4%	97.2%	**97.0%**	
		Vanilla	Random	10.0%	4.4%	1.4%	**0.4%**	-
			Random+	**6.8%**	2.4%	1.2%	**0.4%**	
			PGD	7.0%	**1.4%**	**0.8%**	**0.4%**	
Rotation	$r \pm 50°$	**TSS**	Random	98.6%	**98.4%**	**98.2%**	98.4%	97.4%
			Random+	98.6%	98.6%	98.4%	**98.2%**	
			PGD	**98.2%**	**98.4%**	98.4%	**98.2%**	
		Vanilla	Random	27.2%	17.4%	13.8%	12.2%	-
			Random+	**15.4%**	**13.0%**	**11.0%**	**11.0%**	
			PGD	16.4%	15.6%	15.4%	15.2%	

Table 22: Comparison between empirical robust accuracy against random and adaptive attacks and certified robust accuracy on MNIST. The attack radii are consistent with Table 5. The most powerful attack in each setting is highlighted in bold font. Random+ and PGD attacks are adaptive. Note that PGD attack cannot apply to translation transformation because the parameter space is discrete.

Transformation	Attack Radius	Model	Attack	Empirical Robust Accuracy				Certified robust Accuracy
				N = 10	N = 20	N = 50	N = 100	
Scaling	$s \pm 30\%$	**TSS**	Random	**99.2%**	**99.2%**	**99.2%**	**99.2%**	97.2%
			Random+	**99.2%**	**99.2%**	**99.2%**	**99.2%**	
			PGD	**99.2%**	**99.2%**	**99.2%**	**99.2%**	
		Vanilla	Random	92.0%	91.4%	90.2%	90.2%	-
			Random+	**90.0%**	**89.4%**	**89.2%**	**89.2%**	
			PGD	90.4%	90.2%	90.2%	90.2%	
Rotation and Brightness	$r \pm 50\%$, $b \pm 20\%$	**TSS**	Random	98.8%	98.4%	98.2%	98.2%	97.0%
			Random+	98.6%	98.2%	**98.0%**	98.2%	
			PGD	**98.2%**	**98.0%**	**98.0%**	**98.0%**	
		Vanilla	Random	28.8%	17.8%	12.6%	11.0%	-
			Random+	16.6%	**11.6%**	**10.4%**	**10.4%**	
			PGD	**13.4%**	13.6%	13.0%	12.6%	
Scaling and Brightness	$s \pm 50\%$, $b \pm 50\%$	**TSS**	Random	98.6%	98.6%	98.4%	**97.8%**	96.6%
			Random+	98.4%	98.0%	**97.8%**	**97.8%**	
			PGD	**98.2%**	**97.8%**	**97.8%**	**97.8%**	
		Vanilla	Random	57.4%	46.0%	31.0%	24.8%	-
			Random+	40.4%	28.0%	**19.8%**	**15.6%**	
			PGD	**29.0%**	**25.2%**	25.0%	24.0%	
Rotation, Brightness, and ℓ_2	$r \pm 50\%$, $b \pm 20\%$, $\|\delta\|_2 \leqslant .05$	**TSS**	Random	98.2%	97.8%	**97.6%**	97.6%	96.6%
			Random+	98.4%	98.0%	97.8%	97.6%	
			PGD	**97.6%**	**97.6%**	**97.6%**	**97.4%**	
		Vanilla	Random	27.6%	17.2%	11.4%	10.8%	-
			Random+	15.2%	**11.2%**	**9.4%**	**9.0%**	
			PGD	**13.4%**	11.8%	12.0%	11.8%	
Scaling, Brightness, and ℓ_2	$s \pm 50\%$, $b \pm 50\%$, $\|\delta\|_2 \leqslant .05$	**TSS**	Random	98.4%	98.4%	**97.6%**	**97.6%**	96.4%
			Random+	**97.8%**	97.8%	**97.6%**	**97.6%**	
			PGD	**97.8%**	**97.6%**	**97.6%**	**97.6%**	
		Vanilla	Random	50.4%	38.2%	28.2%	22.2%	-
			Random+	34.4%	23.2%	**13.4%**	**12.2%**	
			PGD	**23.4%**	**22.0%**	21.6%	20.8%	

Table 23: Comparison between empirical robust accuracy against random and adaptive attacks and certified robust accuracy on CIFAR-10. The attack radii are consistent with Table 5. The most powerful attack in each setting is highlighted in bold font. Random+ and PGD attacks are adaptive. Note that the PGD attack cannot apply to translation transformation because the parameter space is discrete.

Transformation	Attack Radius	Model	Attack	Empirical Robust Accuracy				Certified robust Accuracy
				N = 10	N = 20	N = 50	N = 100	
Gaussian Blur	Squared Radius $\alpha \leqslant 16$	**TSS**	Random	66.4%	66.4%	**65.8%**	**65.8%**	63.6%
			Random+	66.8%	66.0%	**65.8%**	**65.8%**	
			PGD	**65.8%**	**65.8%**	**65.8%**	**65.8%**	
		Vanilla	Random	4.8%	4.2%	**3.4%**	**3.4%**	-
			Random+	4.6%	4.0%	3.6%	**3.4%**	
			PGD	**3.4%**	**3.4%**	**3.4%**	**3.4%**	
Translation (Reflection Pad.)	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 20$	**TSS**	Random	**86.2%**	**86.0%**	86.2%	86.2%	80.8%
			Random+	86.4%	**86.0%**	**86.0%**	**86.0%**	
			PGD	-	-	-	-	
		Vanilla	Random	**8.0%**	**7.0%**	4.4%	**4.2%**	-
			Random+	8.2%	7.2%	**4.2%**	**4.2%**	
			PGD	-	-	-	-	
Brightness	$b \pm 40\%$	**TSS**	Random	87.2%	87.2%	87.4%	**87.2%**	87.0%
			Random+	**87.0%**	**87.0%**	**87.0%**	87.4%	
			PGD	87.4%	87.4%	87.4%	87.4%	
		Vanilla	Random	57.8%	51.2%	45.8%	44.4%	-
			Random+	**49.8%**	**44.2%**	**42.8%**	**42.6%**	
			PGD	52.4%	51.0%	50.8%	50.8%	
Contrast and Brightness	$c \pm 40\%$, $b \pm 40\%$	**TSS**	Random	86.2%	86.2%	86.2%	86.0%	82.4%
			Random+	**85.8%**	86.2%	86.0%	**85.8%**	
			PGD	**85.8%**	**85.8%**	**85.8%**	**85.8%**	
		Vanilla	Random	48.0%	40.0%	27.2%	21.0%	-
			Random+	32.0%	23.2%	14.8%	**9.6%**	
			PGD	**17.0%**	**13.0%**	**12.2%**	11.8%	
Gaussian Blur, Translation, Contrast, and Brightness	$\alpha \leqslant 1$, $c \pm 10\%$, $b \pm 10\%$, $\sqrt{\Delta x^2 + \Delta y^2} \leqslant 5$	**TSS**	Random	71.0%	**69.2%**	**68.0%**	**67.6%**	58.2%
			Random+	70.6%	69.8%	68.4%	67.8%	
			PGD	**69.8%**	69.8%	69.0%	68.0%	
		Vanilla	Random	21.2%	16.6%	12.0%	9.6%	-
			Random+	18.6%	14.2%	9.0%	7.2%	
			PGD	**12.8%**	**9.8%**	**6.8%**	**5.6%**	
Rotation	$r \pm 10°$	**TSS**	Random	78.0%	77.0%	76.8%	76.6%	70.6%
			Random+	77.4%	**76.8%**	**76.4%**	**76.4%**	
			PGD	**76.8%**	**76.8%**	76.8%	76.6%	
		Vanilla	Random	69.2%	68.0%	**65.6%**	65.6%	-
			Random+	68.4%	67.2%	66.0%	65.6%	
			PGD	**66.4%**	**66.0%**	**65.6%**	**65.4%**	
	$r \pm 30°$	**TSS**	Random	71.8%	70.2%	69.8%	**69.2%**	63.6%
			Random+	71.0%	**69.4%**	**69.2%**	69.4%	
			PGD	**70.4%**	70.0%	70.0%	69.8%	
		Vanilla	Random	31.6%	27.4%	22.6%	21.6%	-
			Random+	32.2%	27.2%	23.8%	**21.4%**	
			PGD	**25.2%**	**23.8%**	**23.2%**	23.2%	

Table 24: Comparison between empirical robust accuracy against random and adaptive attacks and certified robust accuracy on CIFAR-10. The attack radii are consistent with Table 5. The most powerful attack in each setting is highlighted in bold font. Random+ and PGD attacks are adaptive. Note that the PGD attack cannot apply to translation transformation because the parameter space is discrete.

Transformation	Attack Radius	Model	Attack	Empirical Robust Accuracy				Certified robust Accuracy
				N = 10	N = 20	N = 50	N = 100	
Scaling	$s \pm 30\%$	**TSS**	Random	69.6%	67.8%	67.8%	67.2%	58.8%
			Random+	69.2%	68.4%	67.4%	**67.0%**	
			PGD	**67.8%**	**67.6%**	**67.2%**	**67.0%**	
		Vanilla	Random	60.0%	54.6%	52.8%	51.6%	-
			Random+	56.6%	53.8%	52.2%	**51.2%**	
			PGD	**53.2%**	**52.4%**	**52.0%**	52.0%	
Rotation and Brightness	$r \pm 10°$, $b \pm 10\%$	**TSS**	Random	77.2%	76.8%	77.0%	76.6%	70.6%
			Random+	77.2%	**76.6%**	**76.4%**	**76.0%**	
			PGD	**76.6%**	**76.6%**	**76.4%**	76.4%	
		Vanilla	Random	67.2%	64.8%	60.6%	59.4%	-
			Random+	66.0%	63.0%	59.4%	57.8%	
			PGD	**57.8%**	**57.6%**	**57.0%**	**56.8%**	
	$r \pm 30°$, $b \pm 20\%$	**TSS**	Random	72.0%	70.2%	68.8%	68.4%	61.4%
			Random+	70.6%	68.8%	**68.0%**	**68.2%**	
			PGD	**69.2%**	**68.6%**	68.6%	68.6%	
		Vanilla	Random	26.6%	20.2%	15.8%	13.0%	-
			Random+	18.8%	16.0%	11.6%	9.4%	
			PGD	**12.2%**	**10.4%**	**9.2%**	**9.0%**	
Scaling and Brightness	$s \pm 30\%$, $b \pm 30\%$	**TSS**	Random	68.6%	68.6%	67.4%	67.2%	54.2%
			Random+	68.4%	68.0%	67.0%	**66.8%**	
			PGD	**67.4%**	**67.4%**	**66.8%**	**66.8%**	
		Vanilla	Random	39.2%	30.6%	20.0%	17.4%	-
			Random+	30.4%	19.4%	15.4%	**11.6%**	
			PGD	**16.0%**	**14.4%**	**13.0%**	13.0%	
Rotation, Brightness, and ℓ_2	$r \pm 10°$, $b \pm 10\%$, $\|\delta\|_2 \leqslant .05$	**TSS**	Random	74.2%	72.8%	71.8%	71.6%	64.2%
			Random+	72.8%	72.2%	71.8%	**71.2%**	
			PGD	**71.6%**	**71.6%**	**71.6%**	71.6%	
		Vanilla	Random	40.4%	35.8%	34.4%	31.8%	-
			Random+	36.4%	**34.6%**	**30.8%**	**29.6%**	
			PGD	**36.0%**	35.0%	34.6%	34.6%	
	$r \pm 30°$, $b \pm 20\%$, $\|\delta\|_2 \leqslant .05$	**TSS**	Random	67.6%	66.2%	64.8%	65.2%	55.2%
			Random+	65.6%	65.6%	65.2%	64.4%	
			PGD	**65.2%**	**64.6%**	**64.0%**	**64.0%**	
		Vanilla	Random	7.6%	5.4%	2.6%	0.8%	-
			Random+	3.8%	2.4%	1.2%	**0.4%**	
			PGD	**1.2%**	**0.6%**	**0.6%**	0.6%	
Scaling, Brightness, and ℓ_2	$s \pm 30\%$, $b \pm 30\%$, $\|\delta\|_2 \leqslant .05$	**TSS**	Random	67.6%	66.8%	65.2%	65.0%	51.2%
			Random+	66.0%	66.2%	64.6%	64.4%	
			PGD	**64.2%**	**62.2%**	**61.8%**	**61.8%**	
		Vanilla	Random	15.6%	11.4%	5.8%	4.4%	-
			Random+	8.2%	5.0%	**2.0%**	**2.0%**	
			PGD	**3.8%**	**2.8%**	2.8%	2.6%	

Table 25: Comparison between empirical robust accuracy against random and adaptive attacks and certified robust accuracy on ImageNet. The attack radii are consistent with Table 5. The most powerful attack in each setting is highlighted in bold font. Random+ and PGD attacks are adaptive. Note that PGD attack cannot apply to translation transformation because the parameter space is discrete.

Transformation	Attack Radius	Model	Attack	Empirical Robust Accuracy				Certified robust Accuracy
				N = 10	N = 20	N = 50	N = 100	
Gaussian Blur	Squared Radius $\alpha \leqslant 36$	**TSS**	Random	53.2%	**52.8%**	**52.8%**	52.8%	51.6%
			Random+	53.2%	**52.8%**	**52.8%**	52.8%	
			PGD	**52.8%**	**52.8%**	**52.8%**	**52.6%**	
		Vanilla	Random	9.6%	8.6%	8.4%	8.4%	-
			Random+	8.8%	**8.2%**	**8.2%**	**8.2%**	
			PGD	**8.4%**	**8.2%**	**8.2%**	**8.2%**	
Translation (Reflection Pad.)	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 100$	**TSS**	Random	70.0%	69.6%	**69.2%**	**69.2%**	50.0%
			Random+	**69.4%**	**69.2%**	**69.2%**	**69.2%**	
			PGD	-	-	-	-	
		Vanilla	Random	**55.8%**	**53.4%**	**48.8%**	46.6%	-
			Random+	57.2%	54.6%	50.6%	**46.2%**	
			PGD	-	-	-	-	
Brightness	$b \pm 40\%$	**TSS**	Random	70.8%	**70.4%**	**70.4%**	**70.4%**	70.0%
			Random+	**70.4%**	**70.4%**	**70.4%**	**70.4%**	
			PGD	**70.4%**	**70.4%**	**70.4%**	**70.4%**	
		Vanilla	Random	31.6%	26.6%	21.6%	19.6%	-
			Random+	22.8%	**19.8%**	**18.4%**	**18.4%**	
			PGD	**22.0%**	22.4%	21.8%	21.8%	
Contrast and Brightness	$c \pm 40\%$, $b \pm 40\%$	**TSS**	Random	70.4%	69.2%	**68.4%**	**68.4%**	61.4%
			Random+	69.2%	68.8%	**68.4%**	**68.4%**	
			PGD	**68.4%**	**68.4%**	**68.4%**	**68.4%**	
		Vanilla	Random	20.8%	10.4%	3.6%	1.2%	-
			Random+	8.0%	2.0%	0.4%	**0.0%**	
			PGD	**1.8%**	**0.2%**	**0.0%**	**0.0%**	
Gaussian Blur, Translation Contrast, and Brightness	$\alpha \leqslant 10$, $c \pm 20\%$, $b \pm 20\%$, $\sqrt{\Delta x^2 + \Delta y^2} \leqslant 10$	**TSS**	Random	51.8%	50.2%	49.2%	48.8%	32.8%
			Random+	51.4%	**49.6%**	**48.0%**	48.2%	
			PGD	**49.6%**	**49.6%**	48.2%	**47.4%**	
		Vanilla	Random	20.6%	17.4%	12.0%	9.4%	-
			Random+	15.2%	12.8%	7.8%	6.6%	
			PGD	**11.2%**	**8.0%**	**6.0%**	**4.0%**	
Rotation	$r \pm 30\%$	**TSS**	Random	40.2%	**38.4%**	38.4%	**37.8%**	30.4%
			Random+	**39.0%**	38.6%	**38.0%**	**37.8%**	
			PGD	40.4%	39.8%	39.8%	39.4%	
		Vanilla	Random	47.8%	44.4%	41.4%	40.0%	-
			Random+	45.0%	43.6%	40.6%	38.8%	
			PGD	**39.6%**	**38.4%**	**37.8%**	**37.0%**	

Table 26: Comparison between empirical robust accuracy against random and adaptive attacks and certified robust accuracy on ImageNet. The attack radii are consistent with Table 5. The most powerful attack in each setting is highlighted in bold font. Random+ and PGD attacks are adaptive. Note that PGD attack cannot apply to translation transformation because the parameter space is discrete.

Transformation	Attack Radius	Model	Attack	Empirical Robust Accuracy				Certified robust Accuracy
				N = 10	N = 20	N = 50	N = 100	
Scaling	$s \pm 30\%$	**TSS**	Random	40.2%	38.0%	37.4%	37.4%	26.4%
			Random+	**38.8%**	**37.2%**	**36.8%**	**36.4%**	
			PGD	39.0%	37.8%	37.6%	37.8%	
		Vanilla	Random	55.2%	53.0%	51.2%	50.0%	-
			Random+	55.6%	52.8%	50.6%	50.0%	
			PGD	**50.6%**	**49.8%**	**49.4%**	**49.8%**	
Rotation and Brightness	$r \pm 30°$ $b \pm 20\%$	**TSS**	Random	**38.8%**	**38.0%**	37.2%	37.4%	26.8%
			Random+	39.0%	38.2%	**37.0%**	**36.8%**	
			PGD	39.6%	39.4%	38.6%	38.8%	
		Vanilla	Random	40.4%	35.4%	29.2%	22.4%	-
			Random+	35.2%	31.2%	25.2%	**21.2%**	
			PGD	**25.0%**	**23.2%**	**22.2%**	21.4%	
Scaling and Brightness	$s \pm 30\%$, $b \pm 30\%$	**TSS**	Random	40.2%	38.0%	**36.4%**	36.4%	23.4%
			Random+	38.0%	**37.0%**	36.6%	**36.0%**	
			PGD	**37.0%**	**37.0%**	36.6%	36.6%	
		Vanilla	Random	34.4%	26.2%	19.4%	16.0%	-
			Random+	21.0%	**15.0%**	**12.4%**	**8.8%**	
			PGD	**17.6%**	15.2%	13.8%	13.4%	
Rotation, Brightness, and ℓ_2	$r \pm 30°$, $b \pm 20\%$, $\|\delta\|_2 \leqslant .05$	**TSS**	Random	39.4%	38.2%	37.8%	37.0%	26.6%
			Random+	**38.2%**	**37.8%**	**36.6%**	**36.4%**	
			PGD	38.8%	38.8%	38.4%	38.0%	
		Vanilla	Random	26.0%	23.2%	19.8%	17.6%	-
			Random+	21.4%	18.4%	16.0%	14.4%	
			PGD	**16.6%**	**14.6%**	**14.2%**	**14.0%**	
Scaling, Brightness, and ℓ_2	$s \pm 30\%$, $b \pm 30\%$, $\|\delta\|_2 \leqslant .05$	**TSS**	Random	40.2%	38.2%	37.2%	36.0%	22.6%
			Random+	38.0%	**36.4%**	**35.8%**	**35.6%**	
			PGD	**36.8%**	**36.4%**	36.4%	36.0%	
		Vanilla	Random	24.4%	17.2%	11.4%	7.4%	-
			Random+	13.8%	**8.4%**	**5.8%**	**4.8%**	
			PGD	**9.8%**	8.8%	7.4%	7.4%	

Table 27: Comparison of Empirical Accuracy for each corruption evaluated from the highest severity level (5) of CIFAR-10-C and ImageNet-C.

Corruption		CIFAR-10			ImageNet		
Category	Type	Vanilla	AugMix [92]	TSS	Vanilla	AugMix [92]	TSS
Weather	Snow	68.2%	**75.6%**	69.4%	16.0%	**22.6%**	13.8%
	Fog	63.4%	**65.4%**	62.0%	**24.0%**	22.2%	18.0%
	Frost	59.2%	67.8%	**73.8%**	21.6%	**24.8%**	22.6%
	Brightness	**82.4%**	**82.4%**	71.8%	**56.8%**	56.6%	35.8%
Blur	Zoom Blur	52.6%	70.8%	**75.2%**	21.4%	**31.0%**	20.4%
	Glass Blur	46.6%	50.2%	**72.2%**	8.0%	**14.0%**	13.8%
	Motion Blur	54.8%	68.6%	**70.2%**	14.2%	**25.2%**	11.4%
	Defocus Blur	49.0%	72.2%	**75.6%**	14.0%	22.6%	**25.6%**
Noise	Impulse Noise	29.8%	**51.0%**	46.2%	4.0%	9.8%	**12.0%**
	Gaussian Noise	34.8%	56.4%	**62.8%**	4.4%	9.6%	**12.8%**
	Shot Noise	43.0%	**63.4%**	62.6%	4.0%	13.0%	**14.0%**
Digital	Pixelate	42.0%	59.0%	**76.0%**	19.6%	39.2%	**55.6%**
	Elastic Transform	71.4%	65.2%	**74.4%**	14.8%	**23.8%**	23.6%
	Contrast	23.8%	26.0%	**49.8%**	4.2%	**11.6%**	5.0%
	JPEG Compression	70.8%	**73.0%**	71.8%	33.6%	**45.4%**	31.6%
Extra	Saturate	79.6%	**83.4%**	63.6%	41.6%	**43.4%**	25.8%
	Spatter	72.8%	**82.0%**	69.0%	22.4%	**30.6%**	17.6%
	Speckle Noise	45.2%	**64.0%**	58.8%	11.4%	**27.4%**	23.6%
	Gaussian Blur	34.6%	67.4%	**75.8%**	11.2%	15.2%	**33.0%**
	Average	53.89%	65.46%	**67.42%**	18.27%	**25.68%**	21.89%

samples, where the test samples are all corrupted at the highest severity level (level 5) to model the strongest unforeseen semantic attacker. For each corruption type, there is a full test set generated by 1-to-1 mapping from the original clean test set samples processed with the corruption. Being consistent with the main experiment protocol, for each corruption type, we uniformly pick 500 samples from the corresponding test set. Then, we compute the average empirical accuracy among all 19 corruptions and report it in Table 6 in main text. For reference, we also include their certified accuracy against the composition of Gaussian blur, brightness, contrast, and translation. Since TSS provides robustness certification only for smoothed models, we apply the same smoothing strategy as our TSS models, hoping for providing robustness certificates for baseline models. As shown in Table 6, only TSS models can be certified with nontrivial certified robust accuracy.

BREAKDOWN In Table 27, we show the breakdown of empirical accuracy for all models evaluated in Table 6. Note that the TSS models are trained using only four of these 19 corruptions (brightness, contrast, Gaussian blur, and additive Gaussian noise). Almost on all the corruptions, TSS has higher accuracy than vanilla models and sometimes higher than the state-of-the-art defense—AugMix. Interestingly, we find TSS models have different generalization abilities on these corruptions. The additive Gaussian noise has the best generalization ability, because TSS model also achieves much higher accuracy against impulse noise and shot noise than all the baselines. The Gaussian blur also generalizes well, because we can see significantly higher accuracy of TSS models against zoom blur, glass blur, motion blur, and defocus blur especially on CIFAR-10-C. Finally, brightness and contrast, even though they seem to be among the simplest transformations, have the poorest generalization ability. For example, under severe corruptions, the empirical accuracy of brightness is even below than that of vanilla models. Manual inspection of the corrupted images showed that corrupted brightness or contrast images are severely altered so that they are hard to be distinguished even by humans, giving a hint for possible reasons for the poor performance on these images. We thus conjecture that overly severe corruption could be the reason, and we think that it would be an interesting future direction to study these different generalization abilities in depth.

D.6.5.5 *Certified Accuracy for larger Certification Radii*

In Figure 33 and Figure 34, on MNIST and CIFAR-10, the purple vertical dotted lines stand for the predefined certified radii that the models aim to defend, and the blue curves show the certified robust accuracy (y axis) with respect to attack radii (x axis). The figures show that the TSS models that aim to defend against transformations within certain thresholds still maintain high certified accuracy when the transformation parameters go even far beyond the preset thresholds. In Table 28, Table 29, and Table 30, we further show the empirical robust accuracy of TSS and vanilla models when the attacker goes beyond the predefined certified radius. The empirical robust accuracy is computed as the minimum among all three attacks: Random, Random+, and PGD. We observe that the empirical robust accuracy follows the same tendency. For example, on the CIFAR-10 dataset, the TSS model is trained to defend against the rotation transformation within 30° where it achieves 69.2%/63.6% empirical/certified accuracy. When the rotation angle goes up to 60° the model still preserves 46.8%/37.4% empirical/cer-

Table 28: Empirical and certified robust accuracy on MNIST when attack radii go beyond the predefined one. The predefined certified radii are consistent with 5. For the contrast transformation we allow additional ±50% brightness change since a single Contrast attack is not powerful enough.

			Beyond Predefined Radius			
Gaussian Blur		Radius:	36	40	44	48
	TSS	Empirical	91.2%	90.8%	90.4%	89.2%
		Certified	90.6%	90.0%	89.6%	88.8%
	Vanilla	Empirical	12.2%	11.8%	11.2%	11.2%
Translation (Reflection Pad.)		Radius:	8	10	12	14
	TSS	Empirical	99.6%	99.6%	99.6%	99.6%
		Certified	99.6%	99.6%	99.4%	99.0%
	Vanilla	Empirical	0.0%	0.0%	0.0%	0.0%
Brightness		Radius:	50%	52%	55%	60%
	TSS	Empirical	98.2%	98.2%	98.2%	98.2%
		Certified	98.2%	98.2%	98.2%	98.2%
	Vanilla	Empirical	96.6%	96.2%	95.6%	94.4%
Contrast*		Radius:	50%	52%	55%	60%
	TSS	Empirical	98.0%	98.0%	98.0%	98.0%
		Certified	97.6%	97.2%	96.8%	96.2%
	Vanilla	Empirical	93.2%	93.2%	93.2%	93.0%
Rotation		Radius:	50°	52°	55°	60°
	TSS	Empirical	98.2%	98.2%	98.2%	97.8%
		Certified	97.4%	97.4%	97.4%	96.6%
	Vanilla	Empirical	11.0%	9.8%	8.4%	7.2%
Scaling		Radius:	30%	35%	40%	50%
	TSS	Empirical	99.2%	98.8%	98.8%	98.6%
		Certified	97.2%	96.8%	96.8%	96.0%
	Vanilla	Empirical	89.2%	82.6%	72.8%	45.4%

Table 29: Empirical and certified robust accuracy on CIFAR-10 when the attack radii go beyond the predefined one. The predefined certified radii are consistent with Table 5. For the contrast transformation we allow additional 40% brightness change since a single Contrast attack is not powerful enough.

			Beyond Predefined Radius			
Gaussian Blur		Radius:	16	20	24	28
	TSS	Empirical	65.8%	63.4%	61.2%	56.4%
		Certified	63.6%	60.8%	56.0%	52.6%
	Vanilla	Empirical	3.4%	3.2%	3.0%	2.8%
Tranlation (Reflection Pad.)		Radius:	20	25	30	35
	TSS	Empirical	86.0%	86.0%	85.8%	85.8%
		Certified	80.8%	77.4%	74.8%	70.6%
	Vanilla	Empirical	4.2%	3.6%	3.6%	2.8%
Brightness		Radius:	40%	45%	50%	55%
	TSS	Empirical	87.2%	87.0%	87.0%	87.0%
		Certified	87.0%	87.0%	87.0%	87.0%
	Vanilla	Empirical	42.6%	32.8%	21.2%	14.0%
Contrast*		Radius	40%	45%	50%	55%
	TSS	Empirical	85.8%	85.8%	85.4%	85.2%
		Certified	82.4%	80.8%	79.2%	71.8%
	Vanilla	Empirical	9.6%	7.8%	5.6%	4.8%
Rotation		Radius:	30°	40°	50°	60°
	TSS	Empirical	69.2%	64.0%	57.8%	46.8%
		Certified	63.6%	57.6%	48.2%	37.4%
	Vanilla	Empirical	21.4%	8.8%	5.0%	3.2%
Scaling		Radius:	30%	35%	40%	50%
	TSS	Empirical	67.0%	65.0%	60.6%	54.8%
		Certified	58.8%	53.6%	51.0%	43.4%
	Vanilla	Empirical	51.2%	43.0%	34.8%	21.2%

Table 30: Empirical and certified robust accuracy on ImageNet when attack radii go beyond the predefined one. The predefined certified radii are consistent with Table 5. For Contrast/Rotation/Scaling we allow additional ±40%/±20%/±30% brightness change since a single Contrast/Rotation/Scaling attack is not powerful enough.

			Beyond Predefined Radius		
Gaussian Blur		Radius:	36	40	44
	TSS	Empirical	52.6%	51.2%	49.8%
		Certified	51.6%	50.0%	48.8%
	Vanilla	Empirical	8.2%	7.2%	6.2%
Translation (Reflection Pad.)		Radius:	100	105	110
	TSS	Empirical	69.2%	69.2%	69.0%
		Certified	50.0%	49.4%	46.8%
	Vanilla	Empirical	46.2%	37.6%	36.6%
Brightness		Radius:	40%	45%	50%
	TSS	Empirical	70.4%	70.2%	70.0%
		Certified	70.0%	69.8%	69.6%
	Vanilla	Empirical	18.4%	10.0%	5.2%
Contrast*		Radius:	40%	45%	50%
	TSS	Empirical	68.4%	68.2%	67.6%
		Certified	61.4%	55.8%	45.0%
	Vanilla	Empirical	0.0%	0.0%	0.0%
Rotation*		Radius:	30°	35°	45°
	TSS	Empirical	36.8%	36.4%	33.4%
		Certified	26.8%	26.2%	21.8%
	Vanilla	Empirical	21.2%	19.4%	16.2%
Scaling*		Radius:	30%	40%	50%
	TSS	Empirical	36.0%	32.4%	26.6%
		Certified	23.4%	18.4%	11.6%
	Vanilla	Empirical	8.8%	8.8%	7.0%

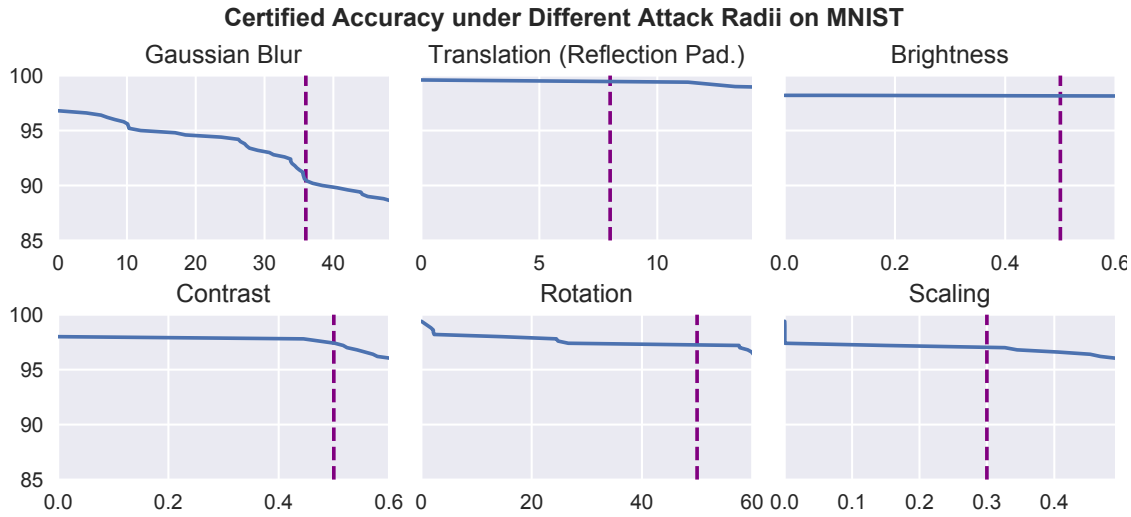


Figure 33: The blue curves show the certified robust accuracy on MNIST. The predefined certified radii are shown as purple vertical dotted lines. We observe no significant degradation after exceeding the predefined radii. For the Contrast transformation, we allow additional 50% brightness change.

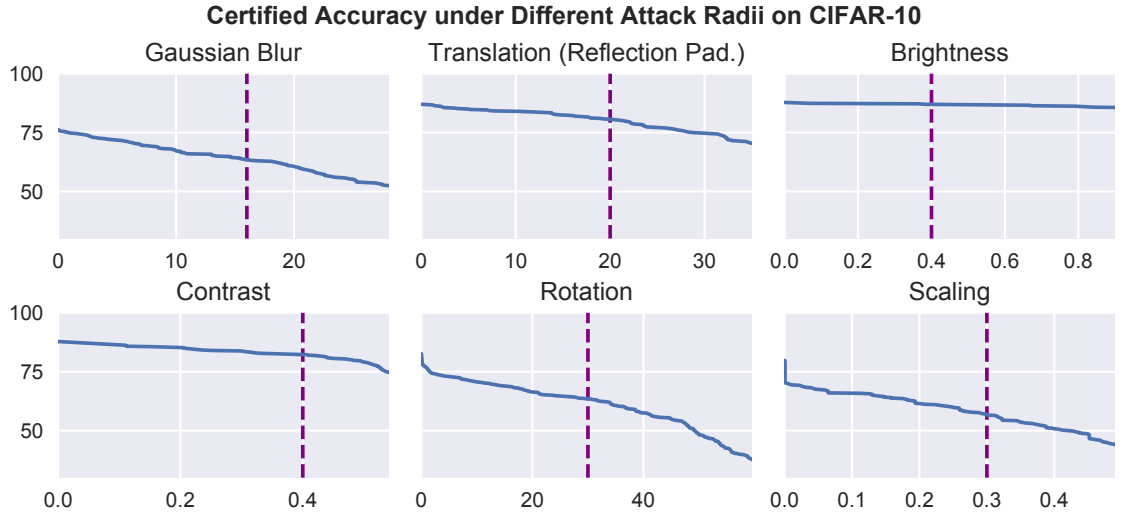


Figure 34: The blue curves show the certified robust accuracy on CIFAR-10. The predefined certified radii are shown as purple vertical dotted lines. We observe no significant degradation after exceeding the predefined certified radii. For the contrast transformation, we allow additional 40% brightness change.

tified accuracy. On the contrary, the vanilla model's empirical accuracy is reduced from 21.4% (30° rotation) to 3.2% (60° rotation).

D.6.5.6 *Additional Smoothing Variance Levels*

In Section 7.6.3.2 we have shown the study on smoothing variance levels on ImageNet (Table 7). Here, we present further results with additional smoothing variance levels on MNIST and CIFAR-10 in Table 31 and Table 32 respectively. The smoothing variances shown in the two tables are for both training and inference-time smoothing. Except for smoothing variance, all other hyperparameters for training and certification are kept

Table 31: Study of the impact of different smoothing variance levels on certified robust accuracy and benign accuracy on MNIST for TSS. The attack radii are consistent with Table 5. The "Dist." refers to both training and smoothing distribution.

Transformation	Attack Radius	Certified Accuracy and Benign Accuracy under Different Variance Levels			
Gaussian Blur	$\alpha \leqslant 36$	Dist. of α	Exp(1/5)	Exp(1/10)	Exp(1/20)
		Cert. Rob. Acc.	90.4%	**90.6%**	89.2%
		Benign Acc.	**97.0%**	96.8%	93.4%
Translation (Reflection Pad.)	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 8$	Dist. of $(\Delta x, \Delta y)$	$\mathcal{N}(0, 5^2 I)$	$\mathcal{N}(0, 10^2 I)$	$\mathcal{N}(0, 15^2 I)$
		Cert. Rob. Acc.	99.0%	**99.6%**	99.4%
		Benign Acc.	**99.6%**	**99.6%**	**99.6%**
Brightness	$b \pm 50\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.5^2 I)$	$\mathcal{N}(0, 0.6^2 I)$	$\mathcal{N}(0, 0.7^2 I)$
		Cert. Rob. Acc.	**98.4%**	98.2%	**98.4%**
		Benign Acc.	**98.4%**	**98.4%**	**98.4%**
Contrast	$c \pm 50\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.5^2 I)$	$\mathcal{N}(0, 0.6^2 I)$	$\mathcal{N}(0, 0.7^2 I)$
		Cert. Rob. Acc.	0.0%	98.0%	**98.4%**
		Benign Acc.	**98.4%**	**98.4%**	**98.4%**
Rotation	$r \pm 50°$	Dist. of ϵ	$\mathcal{N}(0, 0.05^2 I)$	$\mathcal{N}(0, 0.12^2 I)$	$\mathcal{N}(0, 0.20^2 I)$
		Cert. Rob. Acc.	**97.6%**	97.4%	**97.6%**
		Benign Acc.	99.2%	**99.4%**	99.2%
Scaling	$s \pm 30\%$	Dist. of ϵ	$\mathcal{N}(0, 0.05^2 I)$	$\mathcal{N}(0, 0.12^2 I)$	$\mathcal{N}(0, 0.20^2 I)$
		Cert. Rob. Acc.	96.6%	**97.2%**	96.0%
		Benign Acc.	**99.4%**	**99.4%**	99.0%

Table 32: Study of the impact of different smoothing variance levels on certified robust accuracy and benign accuracy on CIFAR-10 for TSS. The attack radii are consistent with Table 5. The "Dist." refers to both training and smoothing distribution.

Transformation	Attack Radius	Certified Accuracy and Benign Accuracy under Different Variance Levels			
Gaussian Blur	$\alpha \leqslant 16$	Dist. of α	Exp(1/5)	Exp(1/10)	Exp(1/20)
		Cert. Rob. Acc.	**63.6%**	60.6%	53.0%
		Benign Acc.	**76.2%**	68.0%	57.4%
Translation (Reflection Pad.)	$\sqrt{\Delta x^2 + \Delta y^2} \leqslant 20$	Dist. of $(\Delta x, \Delta y)$	$\mathcal{N}(0, 10^2 I)$	$\mathcal{N}(0, 15^2 I)$	$\mathcal{N}(0, 20^2 I)$
		Cert. Rob. Acc.	76.2%	**80.8%**	74.4%
		Benign Acc.	**89.0%**	87.0%	84.6%
Brightness	$b \pm 40\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.2^2 I)$	$\mathcal{N}(0, 0.3^2 I)$	$\mathcal{N}(0, 0.4^2 I)$
		Cert. Rob. Acc.	**87.4%**	87.0%	86.2%
		Benign Acc.	**87.8%**	**87.8%**	86.4%
Contrast	$c \pm 40\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.2^2 I)$	$\mathcal{N}(0, 0.3^2 I)$	$\mathcal{N}(0, 0.4^2 I)$
		Cert. Rob. Acc.	0.0%	**82.4%**	**82.4%**
		Benign Acc.	**87.8%**	**87.8%**	86.4%
Rotation	$r \pm 30°$	Dist. of ϵ	$\mathcal{N}(0, 0.05^2 I)$	$\mathcal{N}(0, 0.09^2 I)$	$\mathcal{N}(0, 0.12^2 I)$
		Cert. Rob. Acc.	**63.6%**	62.0%	59.0%
		Benign Acc.	**82.0%**	78.6%	72.2%
Scaling	$s \pm 30\%$	Dist. of ϵ	$\mathcal{N}(0, 0.05^2 I)$	$\mathcal{N}(0, 0.09^2 I)$	$\mathcal{N}(0, 0.12^2 I)$
		Cert. Rob. Acc.	59.0%	**59.4%**	58.8%
		Benign Acc.	**85.4%**	81.6%	79.2%

the same and consistent with the results shown in Table 5. As we can observe, the same conclusion still holds. Usually, when the smoothing variance increases, the benign accuracy drops and the certified robust accuracy first rises and then drops. The reason is that larger smoothing variance makes the input more severely transformed so that the benign accuracy becomes smaller. On the other hand, larger smoothing variance makes the robustness easier to be certified as we can observe in various robustness conditions in Section 7.4.2, where the required lower bound of p_A becomes smaller. This is the reason for the "first rise" on certified accuracy. However, when the smoothing variance becomes too large, the benign accuracy becomes too low, and according to our definition, the certified accuracy is upper bounded by the benign accuracy (precondition of robustness is correctness). This is the reason for the "then drop" on certified accuracy. We again observe that the range of acceptable variance is usually wide. For example, on CIFAR-10, for rotation transformation, the certified robust accuracy is 63.6% / 62.0% / 59.0% across a wide range of smoothing variance: 0.05, 0.09, 0.12. Thus, even in the presence of such trade-off, without fine-tuning the smoothing variances, we can still obtain high certified robust accuracy and high benign accuracy as reported in Table 5 and Table 18 respectively.

D.6.5.7 *Tightness-Efficiency Trade-Off*

We notice that as we increase the number of samples when estimating the interpolation error in Eq. (147) and Eq. (150), the interpolation error M_S and the upper bound $\sqrt{M} \geqslant M_S$ become smaller and the certification becomes tighter, leading to higher certified robust accuracy. However, the computation time is also increased, resulting in a trade-off between speed and accuracy. In Table 33 and Table 34, we illustrate this trade-off on two differentially resolvable transformations: composition of rotation and brightness on CIFAR-10, and composition of scaling and brightnes on MNIST. From the tables, we find that, for these compositions, as the sample numbers N and n increase, the interpolation error decreases and computing time increases (linearly with N and n). As a consequence, if using a large number of samples, we can decrease the smoothing noise level σ and achieve both higher certified accuracy and higher benign accuracy at the cost of larger computation time.

Table 33: Average interpolation upper bound $\sqrt{M}$ (147), average computation time, and "Certified accuracy (average certification time)" for varying number of samples and smoothing noise levels. Results on CIFAR-10 against the composition of rotation $\pm 10^\circ$ and brightness change $\pm 10\%$.

Number of Samples		Interpolation		Smoothing Noise Level σ		
First-Level	Second-Level	Avg. $\sqrt{M}$	Avg. Comp. Time	0.05	0.09	0.12
$N = 556$	$n = 2,000$	0.050	22.50 s	70.2% (62.32 s)	65.2% (86.60 s)	61.2% (53.73 s)
$N = 556$	$n = 200$	0.131	1.97 s	42.0% (490.21 s)	59.2% (93.19 s)	60.4% (86.60 s)
$N = 56$	$n = 2,000$	0.322	1.90 s	1.2% (6.18 s)	12.6% (16.64 s)	29.2% (25.77 s)
$N = 56$	$n = 200$	0.499	0.27 s	0.0% (5.22 s)	1.2% (5.68 s)	3.4% (8.49 s)
			Benign Accuracy:	83.0%	79.2%	79.6%

Table 34: Average interpolation upper bound $\sqrt{M}$ (147), average bound computation time, and "Certified robust accuracy (average certification time)" when using different number of samples and various smoothing noise levels. Data is collected on MNIST dataset against the composition of scaling $\pm 50\%$ and brightness change $\pm 50\%$.

Number of Samples		Interpolation		Smoothing Noise Level σ		
First-Level	Second-Level	Avg. $\sqrt{M}$	Avg. Comp. Time	0.05	0.09	0.12
$N = 2,500$	$n = 500$	0.064	10.52 s	97.2% (92.36 s)	97.4% (76.25 s)	96.6% (67.44 s)
$N = 2,500$	$n = 50$	0.163	0.90 s	18.8% (157.48 s)	97.0% (217.97 s)	95.0% (97.91 s)
$N = 250$	$n = 500$	0.441	0.74 s	0.0% (0.80 s)	6.0% (4.91 s)	16.2% (12.48 s)
$N = 250$	$n = 50$	0.641	0.13 s	0.0% (0.79 s)	0.0% (0.71 s)	0.6% (1.60 s)
			Benign Accuracy:	99.4%	99.6%	99.4%

E

ADDITIONAL RESULTS IN CERTIFYING OUT-OF-DOMAIN GENERALIZATION

E.1 FINITE SAMPLING ERRORS

Here we explain the reasoning behind the finite-sampling version of our main Theorem stated in Corollary 5. Let us first recall a version of Hoeffding's inequality, formulated in terms of our setting.

Theorem 13 ([94]). *Let $Z_1, \ldots, Z_n$ be independent random variables drawn from P and taking values in $\mathcal{Z}$. Let $\ell\colon \mathcal{Z} \to [0, M]$ be a loss function and let $\hat{L}_n := \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i)$ be the mean under the empirical distribution $\hat{P}_n$. Then, for $\delta > 0$, with probability at least $1-\delta$,*

$$\mathbb{E}_P[\ell(Z)] \leqslant \hat{L}_n + M\sqrt{\frac{\ln 1/\delta}{2n}}. \tag{827}$$

We remark that one could in principle different concentration inequalities at this stage which can potentially improve upon Hoeffding's inequality. For example, [148] present a finite sampling version of Bennett's inequality which is known to be an improvement over Hoeffding's inequality in the low variance regime. We leave such considerations for interesting future work. Recall that the certificate (196) is monotonically increasing in the variance. For this reason, we are interested in an upper bound on the population variance which can be computed from finite samples. To achieve this, we use the variance bound presented in Theorem 10 in [148] which we state here for completeness and adapt it to our use case.

Theorem 14 ([148], Theorem 10). *Let $Z_1, \ldots, Z_n$ be independent random variables drawn from P and taking values in $\mathcal{Z}$. For a loss function $\ell\colon \mathcal{Z} \to [0, M]$, let $S_n^2 := \frac{1}{n(n-1)}\sum_{1\leqslant i<j\leqslant n}(\ell(Z_i) - \ell(Z_j))^2$ be the unbiased estimator of the variance of the random variable $\ell(Z)$, $Z \sim P$. Then, for $\delta > 0$, with probability at least $1-\delta$,*

$$\sqrt{\mathbb{V}_P[\ell(Z)]} \leqslant \sqrt{S_n^2} + M\sqrt{\frac{2\ln 1/\delta}{n-1}} \tag{828}$$

Finally, we employ the union bound to upper bound both expectation and variance simultaneously with high probability. Thus, for any $\delta > 0$, we have with probability at least $1-\delta$

$$\begin{aligned} \mathbb{E}_P[\ell(Z)] &\leqslant \hat{L}_n + M\sqrt{\frac{\ln 2/\delta}{2n}}, \\ \sqrt{\mathbb{V}_P[\ell(Z)]} &\leqslant \sqrt{S_n^2} + M\sqrt{\frac{2\ln 2/\delta}{n-1}}. \end{aligned} \tag{829}$$

Finally, plugging in these upper bounds for the population quantities in Theorem 7 leads to the desired finite sampling bound. Getting the finite sampling version of the lower bound in Theorem 15 is analogous by using the corresponding lower bound variant of Hoeffding, but still the same upper bound for the variance.

E.2 A LOWER BOUND VERSION OF THEOREM 7

Given the proof of Theorem 7, it is straightforward to adapt it to get a *lower* bound on expectation values using Theorem 1. By following the analogous reasoning as in the proof of Theorem 7, we obtain the following result.

Theorem 15 (Lower bound). *Let $\ell\colon \mathcal{Z} \to \mathbb{R}_+$ be a non-negative function taking values in $\mathcal{Z}$. Then, for any probability measure P on $\mathcal{Z}$ and $\rho > 0$ we have*

$$\begin{aligned}\inf_{Q\in B_\rho(P)} \mathbb{E}_Q[\ell(Z)] \geqslant \mathbb{E}_P[\ell(Z)] - 2C_\rho\sqrt{\mathbb{V}_P[\ell(Z)]} \\ - \rho^2(2-\rho^2)\left[\mathbb{E}_P[\ell(Z)] - \frac{\mathbb{V}_P[\ell(Z)]}{\mathbb{E}_P[\ell(Z)]}\right]\end{aligned} \tag{830}$$

where $C_\rho = \sqrt{\rho^2(1-\rho^2)^2(2-\rho^2)}$ and $B_\rho(P) = \{Q \in \mathcal{P}(\mathcal{Z})\colon H(P, Q) \leqslant \rho\}$ is the Hellinger ball of radius ρ centered at P. The radius ρ is required to be small enough such that

$$\rho^2 \leqslant 1 - \left[1 + \frac{\mathbb{E}_P[\ell(Z)]^2}{\mathbb{V}_P[\ell(Z)]}\right]^{-1/2}. \tag{831}$$

E.3 SYNTHETIC DATASET

We consider a binary classification task with covariates $X \in \mathbb{R}^2$ and labels $Y \in \pm 1$, where the data is distributed according to the Gaussian mixture

$$X|\,Y = y \sim \mathcal{N}(y \cdot \mu, \mathbb{1}_2). \tag{832}$$

with $p(y) = 1/2$ and $\mu = (2, 0)^T \in \mathbb{R}^2$. When considering the distribution shift $P \to Q$ arising from perturbations $X \mapsto X + \delta$ for a fixed $\delta \in \mathbb{R}^2$, both the Wasserstein distance and Hellinger distance can be evaluated as functions of the L_2-norm of the perturbation:

$$W_2(P, Q) = \|\delta\|_2\,, \quad H(P, Q) = \sqrt{1 - e^{-\|\delta\|_2^2/8}}. \tag{833}$$

For our classification model, we use a small neural network with ELU activations and 2 hidden layers of size 4 and 2. The ELU activations, in combination with spectral normalization of the weights, enforce the model to be smooth and hence satisfy the assumptions required for the certificate from [203].

E.4 LIPSCHITZ CONSTANT FOR GRADIENTS OF NEURAL NETWORKS WITH JENSEN-SHANNON DIVERGENCE LOSS

Let us first recall the dual reformulation of the Wasserstein worst-case risk, which is the central result that underpins the distributional robustness certificate presented in [203].

Proposition 1 ([203], Proposition 1). *Let $\ell\colon \Theta \times \mathcal{Z} \to \mathbb{R}$ and $c\colon \Theta \times \mathcal{Z} \to \mathbb{R}_+$ be continuous, and let $\phi_\gamma(\theta; z_0) := \sup_{z\in\mathcal{Z}}\{\ell(\theta; z) - \gamma c(z, z_0)\}$. Then, for any distribution P and any $\rho > 0$,*

$$\sup_{Q\colon W_c(P,Q)\leqslant\rho} \mathbb{E}_Q[\ell(\theta; Z)] = \inf_{\gamma\geqslant 0}\{\gamma\rho + \mathbb{E}_P[\phi_\gamma(\theta; Z)]\}. \tag{834}$$

where $W_c(P, Q) := \inf_{\pi\in\Pi(P,Q)} \int_{\mathcal{Z}} c(z, z')\,d\pi(z, z')$ is the 1-Wasserstein distance between P and Q.

From this result, [203] derive a robustness certificate which can be instantiated to hold uniformly over a function of families parametrized by $\theta \in \Theta$, but also a certificate that holds pointwise, that is, for a single model $\ell(\theta_0; \cdot)$. One requirement for this certificate to be tractable is that the surrogate function ϕ_γ be concave in z. As shown in [203], this is the case when γ is larger than the Lipschitz constant L of the gradient of ℓ with respect to z. Thus one needs to compute L and choose $\gamma \geqslant L$ so that the inner maximization in (834) is guaranteed to converge and hence a robustness certificate can be calculated.

Here, we present the calculation of the Lipschitz constant for the gradient of the Jensen-Shannon divergence loss with respect to input features. For the remainder of this section, we set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with a binary label space $|\mathcal{Y}| = C = 2$. We will always write vectors in bold roman letters, for example $\mathbf{p} = (p_1, \ldots, p_C) \in \mathbb{R}^C$ and $\mathbf{e}_y \in \mathbb{R}^C$ denotes a standard basis vector with zeros everywhere except 1 at position y. We consider a feedforward neural network with L layers and ELU activation functions, denoted by σ:

$$F_L(\theta; x) := \sigma(\theta_l \cdot \sigma_{L-1}(\theta_{L-1} \cdots \sigma(\theta_1 \cdot x) \cdots)) \tag{835}$$

and we are interested in calculating $L > 0$ such that

$$\left\|\nabla \ell(F_L(\theta; x), y) - \nabla \ell(F_L(\theta; x'), y)\right\|_* \leqslant L \left\|x - x'\right\|_2 \tag{836}$$

where the gradient is taken with respect to x and where ℓ is the Jensen-Shannon divergence. To achieve this, we apply Proposition 5 in [203] which states that the Jacobian of F_L is $\beta_L(\theta)$-Lipschitz with respect to the operator norm induced by $\|\cdot\|_2$ with

$$\beta_l(\theta) = \alpha_l(\theta) \sum_{j=1}^{l} \left\{ \frac{L_j^1}{(L_j^0)^2} \alpha_j(\theta) \right\}, \qquad \alpha_l(\theta) = \prod_{j=1}^{l} L_j^0 \left\|\theta_j\right\|_{op}. \tag{837}$$

where L_j^0 is the Lipschitz constant of each activation function σ_j and L_j^1 is the Lipschitz constant of its Jacobian. It is useful to write this recursively as

$$\begin{aligned} \alpha_{l+1}(\theta) &= L_{l+1}^0 \left\|\theta_{l+1}\right\|_{op} \alpha_l(\theta), \\ \beta_{l+1}(\theta) &= L_{l+1}^0 \left\|\theta_{l+1}\right\|_{op} \beta_l(\theta) + L_{l+1}^1 \left\|\theta_{l+1}\right\|_{op}^2 \alpha_l(\theta)^2 \end{aligned} \tag{838}$$

In our case, since we have ELU activations, we have $L_j^0 = L_j^1 = 1$ for all j ([203], example 3). Finally, viewing $\ell(\mathbf{p}, \mathbf{e}_y)$ as an $L+1$ layer neural network with a single output dimension, we have that $\nabla_z \ell(\mathbf{p}(x), y)$ is L^*-Lipschitz continuous with constant

$$L^* = L_{L+1}^0 \beta_L(\theta) + L_{L+1}^1 \alpha_L(\theta)^2 \tag{839}$$

where we have used that $\|\theta_{L+1}\|_{op} = \|\mathbb{1}\|_{op} = 1$ and where L_{L+1}^0 is the Lipschitz constant of the function $z \mapsto \ell(\mathbf{p}(z), y)$ and L_{L+1}^1 is the Lipschitz constant of $z \mapsto \nabla_z \ell(\mathbf{p}(z), y)$ and $\mathbf{p}(z)$ is the softmax probability vector

$$\mathbf{p}(z) = \left(\frac{e^{z_1}}{\sum_j e^{z_j}}, \ldots, \frac{e^{z_C}}{\sum_j e^{z_j}} \right)^T \in \mathbb{R}^C. \tag{840}$$

We now show the calculation of L_{L+1}^0 and L_{L+1}^1. Fix $z \in \mathbb{R}^C$ and $y \in \mathcal{Y}$, and let $\mathbf{e}_y \mathbb{R}^K$ be the one hot encoded label vector with zero everywhere except at position y. The Jensen-Shannon divergence loss between a vector of predicted class probabilities $\mathbf{p}$ and the class label $\mathbf{e}_y$ is given by

$$\ell(\mathbf{p}, \mathbf{e}_y) = \frac{1}{2} \left(D_{KL}(\mathbf{p} \| \mathbf{m}) + D_{KL}(\mathbf{e}_y \| \mathbf{m}) \right) \tag{841}$$

with $\mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{e}_y)$. The Kullback Leibler divergences are

$$\begin{aligned} D_{KL}(\mathbf{p} \| \mathbf{m}) &= 1 + p_y \log\left(\frac{p_y}{1+p_y}\right) \\ D_{KL}(\mathbf{e}_y \| \mathbf{m}) &= 1 + \log\left(\frac{1}{1+p_y}\right) \end{aligned} \tag{842}$$

where $\log = \log_2$ is the logarithm with base 2. The Jensen-Shannon divergence loss is thus given by

$$\ell(\mathbf{p}, \mathbf{e}_y) = 1 + \frac{1}{2}\left(p_y \log(p_y) - (1+p_y)\log(1+p_y)\right). \tag{843}$$

The gradient $\nabla_z \ell(\hat{p}, e_y)$ of the loss with respect to the input x is given by

$$\begin{aligned} \nabla_z \ell(\mathbf{p}, \mathbf{e}_y) &= \frac{1}{2}\nabla_z \left(p_y \log(p_y) - (1+p_y)\log(1+p_y)\right) \\ &= \frac{1}{2}\nabla_z \left(p_y \log(p_y)\right) - \frac{1}{2}\nabla_z \left((1+p_y)\log(1+p_y)\right) \\ &= \frac{1}{2}(1+\log(p_y))\nabla_z p_y - \frac{1}{2}(1+\log(1+p_y))\nabla_z p_y \end{aligned} \tag{844}$$

Noting that

$$\nabla_z p_y(x) = p_y(\mathbf{e}_y - \mathbf{p}) \tag{845}$$

yields the expression

$$\nabla_z \ell(\mathbf{p}, \mathbf{e}_y) = \frac{1}{2}\log\left(\frac{p_y}{1+p_y}\right) p_y(\mathbf{e}_y - \mathbf{p}). \tag{846}$$

Thus,

$$\begin{aligned} L^0_{L+1} &= \sup_z \|\nabla_z \ell(\mathbf{p}, \mathbf{e}_y)\|_2 = \frac{1}{2}\sup_z \left(\log\left(\frac{1+p_y}{p_y}\right) p_y \|\mathbf{e}_y - \mathbf{p}\|_2\right) \\ &= \sup_z \|\nabla_z \ell(\mathbf{p}, \mathbf{e}_y)\|_2 = \frac{1}{\sqrt{2}}\sup_z \left(\log\left(\frac{1+p_y}{p_y}\right) p_y(1-p_y)\right) \approx 0.314568. \end{aligned} \tag{847}$$

We will now calculate $L^1_{L+1} = \sup_x \|J\|_2$ where $J \equiv J_{\ell(\mathbf{p},\mathbf{e}_y)}$, is the Jacobian of $\ell(\mathbf{p}, \mathbf{e}_y)$ and $\|J\|_2$ is given by the largest singular value of J. For ease of notation, let $f_i(x) \equiv (\nabla_z \ell)_i$ and recall that J is defined by

$$J = \begin{pmatrix} \nabla_z^T f_1 \\ \vdots \\ \nabla_z^T f_C \end{pmatrix}. \tag{848}$$

Note that

$$\begin{aligned} \nabla_z f_i &= \nabla_z \frac{1}{2}\log\left(\frac{p_y}{1+p_y}\right) p_y(\delta_{iy} - p_i) \\ &= \frac{1}{2}\left(\frac{1+p_y}{p_y}\left[\frac{\nabla_z p_y}{1+p_y} - \frac{p_y}{(1+p_y)^2}\nabla_z p_y\right]\right) p_y(\delta_{iy} - p_i) + \\ &\quad + \frac{1}{2}\log\left(\frac{p_y}{1+p_y}\right)(\delta_{iy} - p_i)\nabla_z p_y - \frac{1}{2}\log\left(\frac{p_y}{1+p_y}\right) p_y \nabla_z p_i \\ &= \frac{1}{2}\left(\frac{1}{1+p_y} + \log\left(\frac{p_y}{1+p_y}\right)\right)(\delta_{iy} - p_i)\nabla_z p_y - \frac{1}{2}\log\left(\frac{p_y}{1+p_y}\right) p_y \nabla_z p_i \end{aligned} \tag{849}$$

and hence, using $\nabla_z p_y = p_y(\mathbf{e}_y - \mathbf{p})$,

$$\begin{aligned}\nabla_z f_i = \frac{1}{2}\left(\frac{p_y}{1+p_y} + p_y \log\left(\frac{p_y}{1+p_y}\right)\right)(\delta_{iy} - p_i)(\mathbf{e}_y - \mathbf{p}) \\ - \frac{1}{2} p_y \log\left(\frac{p_y}{1+p_y}\right) p_i (\mathbf{e}_i - \mathbf{p})\end{aligned} \tag{850}$$

It follows that the Jacobian is given by

$$\begin{aligned}J = \frac{1}{2}\left(\frac{p_y}{1+p_y} + p_y \log\left(\frac{p_y}{1+p_y}\right)\right)(\mathbf{e}_y - \mathbf{p}) \cdot (\mathbf{e}_y - \mathbf{p})^T \\ + \frac{1}{2} p_y \log\left(\frac{1+p_y}{p_y}\right)(\operatorname{diag}(\mathbf{p}) - \mathbf{p} \cdot \mathbf{p}^T).\end{aligned} \tag{851}$$

Since we are only interested in the binary case $C = 2$, we see that

$$A := (\mathbf{e}_y - \mathbf{p}) \cdot (\mathbf{e}_y - \mathbf{p})^T = (1 - p_y)^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{852}$$

with spectrum $\sigma(A) = \{0, 2(1-p_y)^2\}$. The eigenvalues of $\operatorname{diag}(\mathbf{p})$ are p_i and hence $\lambda(\operatorname{diag}(\mathbf{p})) \subseteq [0, 1]$, and $\sigma(\mathbf{p} \cdot \mathbf{p}^T) = \{0, \|\mathbf{p}\|_2^2\}$. It follows that $\sigma(\operatorname{diag}(\mathbf{p}) - \mathbf{p} \cdot \mathbf{p}^T) \subseteq [-\|\mathbf{p}\|_2^2, 1]$. Thus, by Weyl's inequality and noting that the term in front of $(\mathbf{e}_y - \mathbf{p}) \cdot (\mathbf{e}_y - \mathbf{p})^T$ is always negative, we have for any eigenvalue λ of J that

$$\begin{aligned}(1-p_y)^2\left(\frac{p_y}{1+p_y} + p_y \log\left(\frac{p_y}{1+p_y}\right)\right) - \frac{1}{2} p_y \log\left(\frac{1+p_y}{p_y}\right)\|\mathbf{p}\|_2^2 \\ \leqslant \lambda \leqslant \frac{1}{2} p_y \log\left(\frac{1+p_y}{p_y}\right)\end{aligned} \tag{853}$$

Note that J is symmetric, and hence its largest singular value is given by the largest absolute value of its eigenvalues. Taking the infimum (supremum) of the LHS (RHS) with respect to z yields the bounds

$$-\frac{1}{2} \leqslant \lambda \leqslant \frac{1}{2} \tag{854}$$

and hence

$$L^1_{L+1} = \sup_z \|J\|_2 \leqslant \frac{1}{2}. \tag{855}$$

It follows that $\nabla_x \ell(\mathbf{p}(F_L(\theta; x)), y)$ is L^*-Lipschitz with

$$L^* = L^0_{L+1}\beta_L(\theta) + \frac{1}{2}\alpha_L(\theta)^2 \tag{856}$$

and $L^0_{L+1} = 0.314568$. Finally, choosing $\gamma \geqslant L^*$ in (834) makes the objective in the surrogate loss ϕ_γ concave and hence enables the certificate

$$\begin{aligned}\sup_{Q\,:\,W_c(P,Q)\leqslant\rho} \mathbb{E}_Q[\ell(\theta; Z)] &\leqslant \gamma\rho + \mathbb{E}_P[\phi_\gamma(\theta; Z)] \\ &= \gamma\rho + \mathbb{E}_{(X,Y)\sim P}[\sup_{x\in\mathcal{X}} \ell(F_L(\theta; x), Y) - \gamma\|x - X\|_2^2].\end{aligned} \tag{857}$$

E.5 HELLINGER DISTANCE FOR MIXTURES OF DISTRIBUTIONS WITH DISJOINT SUPPORT

Consider two joint (feature, label)-distributions $P, Q \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with densities f_P and f_Q with respect to a suitable measure. P and Q have disjoint support if

$$\forall x \in \mathcal{X}, y \in \mathcal{Y}: \qquad f_Q(x, y) > 0 \iff f_P(x, y) = 0. \tag{858}$$

In this case, for $\gamma \in (0, 1)$, we define the mixture measure as $\Pi_\gamma := \gamma P + (1-\gamma)Q$ with density

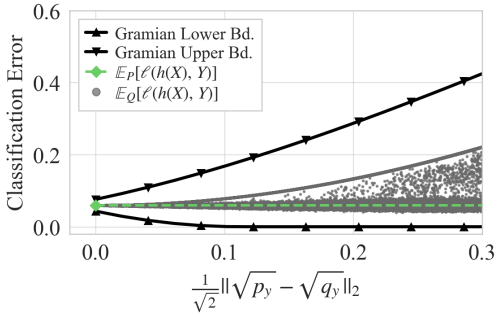
$$\pi_\gamma(x, y) = \gamma f_P(x, y) + (1-\gamma) f_Q(x, y). \tag{859}$$

We can calculate the squared Hellinger distance between P and Π_γ as

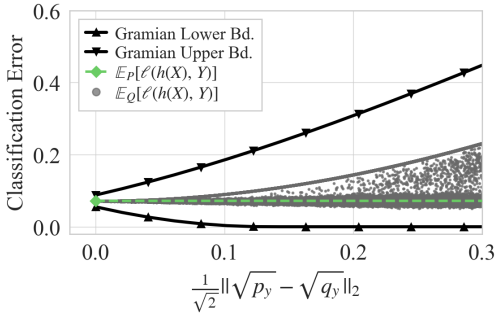
$$\begin{aligned} H^2(P, \Pi_\gamma) &= 1 - \int\int_{\mathcal{X}\times\mathcal{Y}} \sqrt{f_P(x, y)}\sqrt{\gamma f_P(x, y) + (1-\gamma) f_Q(x, y)}\, dx\, dy \\ &= 1 - \sqrt{\gamma}\int\int_{f_P>0} f_P(x, y)\sqrt{1 + \frac{1-\gamma}{\gamma}\frac{f_Q(x, y)}{f_P(x, y)}}\, dx\, dy \\ &= 1 - \sqrt{\gamma}\int\int_{f_P>0} f_P(x, y)\, dx\, dy \\ &= 1 - \sqrt{\gamma}. \end{aligned} \tag{860}$$

E.6 ADDITIONAL EXPERIMENTS

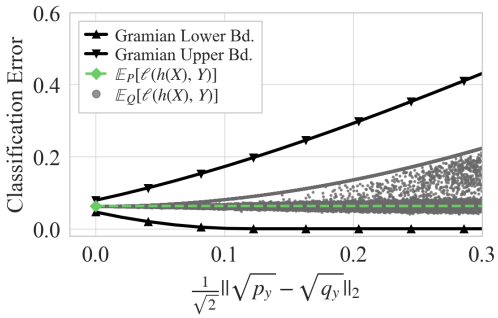
Here, we present results for a diverse set of model architectures and loss functions.



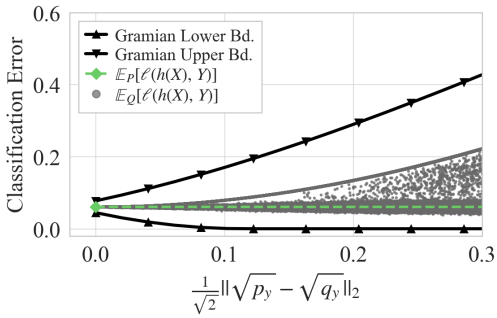
(a) DenseNet-169



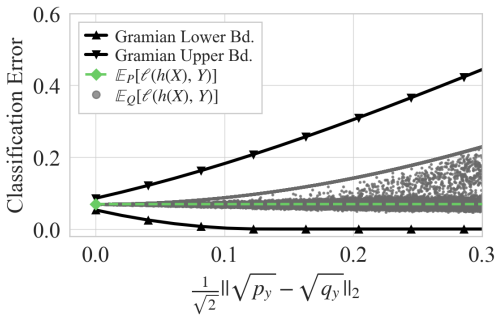
(b) GoogleNet



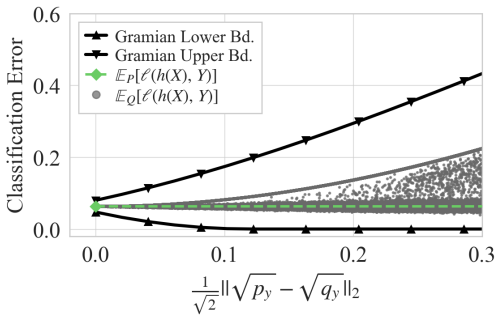
(c) Inception-V3



(d) MobileNet-V2

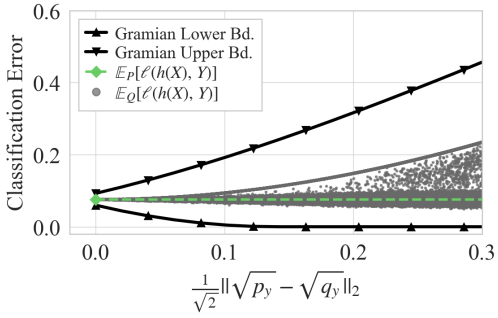


(e) ResNet-18

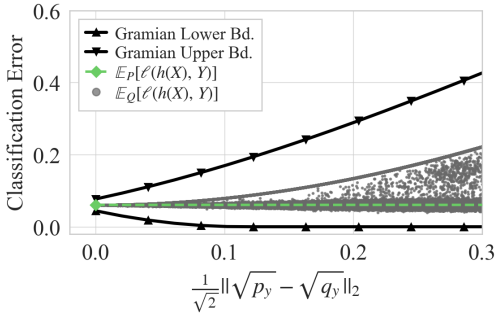


(f) ResNet-50

Figure 35: Certified classification error with label distribution shifts on CIFAR-10.

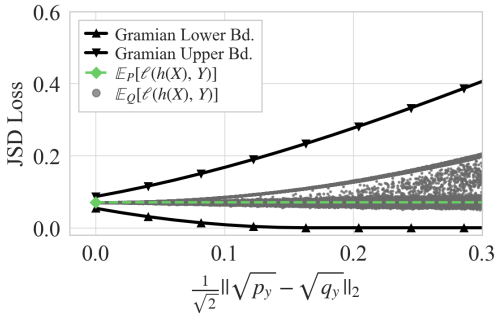


(a) VGG11-BN

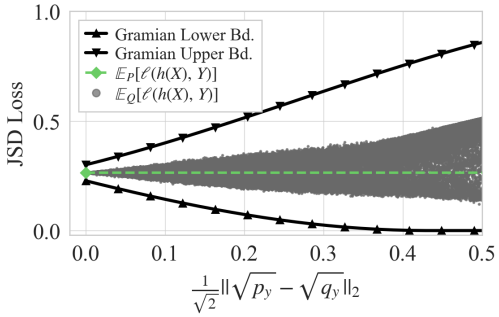


(b) VGG19-BN

Figure 36: Certified classification error with label distribution shifts on CIFAR-10.

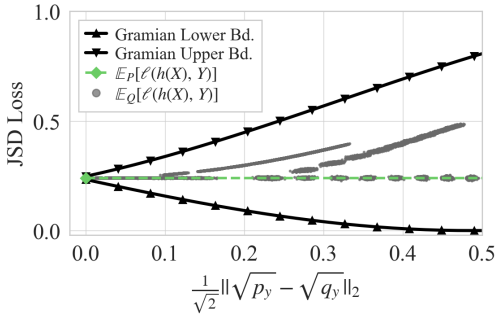


(a) DenseNet-121 on CIFAR-10

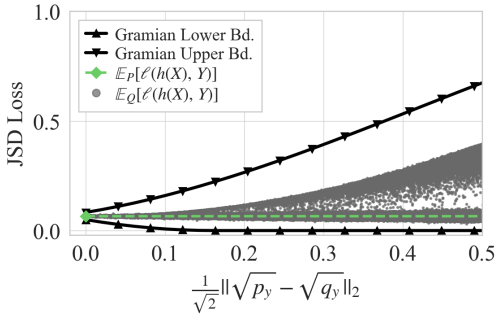


(b) BERT on Yelp

Figure 37: Certified JSD Loss with label distribution shifts.

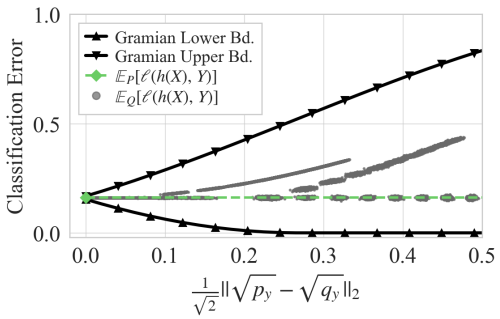


(a) EfficientNet-B7 on ImageNet-1k

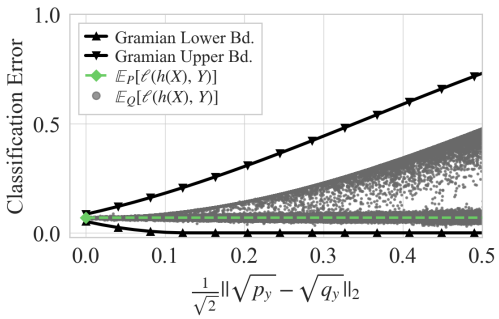


(b) DeBERTa on SNLI

Figure 38: Certified JSD Loss with label distribution shifts.



(a) EfficientNet-B7 on ImageNet-1k



(b) DeBERTa on SNLI

Figure 39: Certified classification error with label distribution shifts.

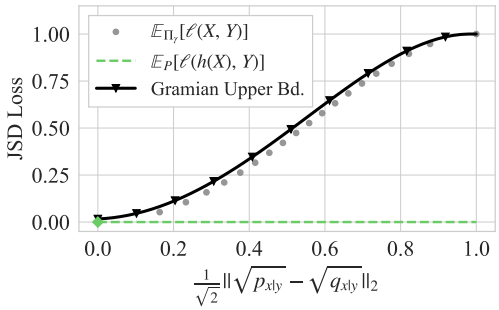


Figure 40: Certified Jensen-Shannon divergence loss for the colored MNIST dataset.

F

ADDITIONAL RESULTS IN CERTIFIED ROBUSTNESS VIA QUANTUM HYPOTHESIS TESTING

F.1 PROOFS

Here, we provide the proofs for Corollary 6 and Corollary 7.

F.1.1 *Proof of Corollary 6*

Corollary 6 (restated). *Let $\sigma, \rho \in \mathcal{S}(\mathcal{H})$ and suppose that $\sigma = |\psi_\sigma\rangle\langle\psi_\sigma|$ is pure. Let $\mathcal{A}$ be a quantum classifier and suppose that for $k_A \in \mathcal{C}$ and $p_A, p_B \in [0, 1]$, we have $k_A = \mathcal{A}(\sigma)$ and suppose that the score function $\mathbf{y}$ satisfies (215). Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with*

$$T(\rho, \sigma) < \delta(p_A, p_B)\left(1 - \sqrt{1 - \delta(p_A, p_B)^2}\right) \tag{861}$$

where $\delta(p_A, p_B) = [\frac{1}{2}(1 - g(p_A, p_B))]^{\frac{1}{2}}$.

Proof. We denote the convex hull enclosed by the set of robust pure states as $\mathcal{C} := \mathrm{Conv}(\{|\psi\rangle\langle\psi| : \||\psi\rangle\langle\psi| - \sigma\|_1 < \delta(P_A, P_B)\})$. Observe that any convex mixture $\rho = \sum_i p_i|\psi_i\rangle\langle\psi_i|$ with $\sum_i p_i = 1$ of any sets of robust pure states $\{|\psi_i\rangle\langle\psi_i|\} \in \mathcal{C}$ must also be robust. Thus it suffices to prove condition (241) implies $\rho \in \mathcal{C}$. Note that the boundary consisting of non-extreme points (which correspond to mixed-states) of $\mathcal{C}$ interfaces with the set $\mathcal{C}^* = \mathrm{Conv}(\{|\psi^*\rangle\langle\psi^*| : \||\psi^*\rangle\langle\psi^*| - \sigma\|_1 = \delta(P_A, P_B)\})$. Thus, it suffices to compute the shortest distance r from σ to $\mathcal{C}^*$, such that $r = \min_{\rho^*} \|\rho^* - \sigma\|_1$ where $\rho^* \in \mathcal{C}^*$, then $\|\rho - \sigma\|_1 < r$ guarantees robustness. Further note that for every ρ^*, $\exists D_\sigma \rho^* D_\sigma^\dagger \in \mathcal{C}^*$, where $D_\sigma = 2\sigma - \mathbb{1}$, such that $\|\rho^* - \sigma\|_1 = \|D_\sigma \rho^* D_\sigma^\dagger - \sigma\|_1$, and $\|p_1\rho^* + p_2 D_\sigma \rho^* D_\sigma^\dagger - \sigma\|_1 < \|\rho^* - \sigma\|_1$ for $p_1 + p_2 = 1$, and $p_1 \neq 0$, $p_2 \neq 0$. Therefore to minimise the distance to σ, it suffices to require $\rho^* = D_\sigma \rho^* D_\sigma^\dagger$, a valid of which is $\rho^* = \frac{1}{2}(|\psi^*\rangle\langle\psi^*| + D_\sigma|\psi^*\rangle\langle\psi^*|D_\sigma^\dagger)$. As such we have

$$\begin{aligned} r &= \|\sigma - \frac{1}{2}(|\psi^*\rangle\langle\psi^*| + D_\sigma|\psi^*\rangle\langle\psi^*|D_\sigma^\dagger)\|_1 \\ &= \||\psi^*\rangle\langle\psi^*| - \sigma + 2|\langle\psi_\sigma|\psi^*\rangle|^2\sigma - \langle\psi_\sigma|\psi^*\rangle|\psi_\sigma\rangle\langle\psi| - \langle\psi|\psi_\sigma\rangle|\psi\rangle\langle\psi_\sigma|\|_1 \end{aligned} \tag{862}$$

Note that we have $\||\psi^*\rangle\langle\psi^*| - \sigma\|_1 = \delta(P_A, P_B)$ by definition and that

$$\begin{aligned} &\|2|\langle\psi_\sigma|\psi^*\rangle|^2\sigma - \langle\psi_\sigma|\psi^*\rangle|\psi_\sigma\rangle\langle\psi| - \langle\psi|\psi_\sigma\rangle|\psi\rangle\langle\psi_\sigma\|_1 = \\ &\quad = |\langle\psi_\sigma|\psi^*\rangle|\mathrm{Tr}\left[|\psi_\sigma\rangle\langle\psi_\sigma| + |\psi^*\rangle\langle\psi^*| - \langle\psi_\sigma|\psi^*\rangle|\psi_\sigma\rangle\langle\psi^*| - \langle\psi^*|\psi_\sigma\rangle|\psi^*\rangle\langle\psi_\sigma|\right] \\ &\quad = |\langle\psi_\sigma|\psi^*\rangle|\mathrm{Tr}\left[(\sigma - |\psi^*\rangle\langle\psi^*|)(\sigma - |\psi^*\rangle\langle\psi^*|)^\dagger\right] \\ &\quad = |\langle\psi_\sigma|\psi^*\rangle|\||\psi^*\rangle\langle\psi^*| - \sigma\|_1 \\ &\quad = \delta(P_A, P_B)\sqrt{1 - \frac{\delta(P_A, P_B)^2}{4}}. \end{aligned} \tag{863}$$

Applying the reversed triangle inequality, we finally arrive at

$$\begin{aligned} r &\geqslant \left| \|\sigma - |\psi^*\rangle\langle\psi^*|\|_1 - \|2|\langle\psi_\sigma|\psi^*\rangle|^2\sigma - \langle\psi_\sigma|\psi^*\rangle|\psi_\sigma\rangle\langle\psi| - \langle\psi|\psi_\sigma\rangle|\psi\rangle\langle\psi_\sigma\|_1 \right| \\ &= \delta(P_A, P_B)\left(1 - \sqrt{1 - \frac{\delta(P_A, P_B)^2}{4}}\right). \end{aligned} \tag{864}$$

□

F.1.2 *Proof of Corollary 7*

Corollary 7 (restated). *Let $|\psi_\sigma\rangle, |\psi_\rho\rangle \in \mathbb{C}^2$ be single-qubit pure sates and let $\mathcal{E}_p^{\text{dep}}$ be a depolarising channel with noise parameter $p \in (0, 1)$. Then, if $p_A > 1/2$ and $p_B = 1 - p_A$, the robustness condition (216) for $\mathcal{E}_p^{\text{dep}}(\sigma)$ and $\mathcal{E}_p^{\text{dep}}(\rho)$ is equivalent to*

$$\frac{1}{2}\||\psi_\sigma\rangle\langle\psi_\sigma| - |\psi_\rho\rangle\langle\psi_\rho|\|_1 < r_Q(p) \tag{865}$$

where

$$r_Q(p) = \begin{cases} \sqrt{\frac{1}{2} - \frac{\sqrt{g(p,p_A)}}{1-p}}, & p_A < \frac{1+3(1-p)^2}{2+2(1-p)^2} \\ \sqrt{\frac{p\cdot(2-p)\cdot(1-2p_A)^2}{8(1-p)^2\cdot(1-p_A)}}, & p_A \geqslant \frac{1+3\cdot(1-p)^2}{2+2\cdot(1-p)^2} \end{cases} \tag{866}$$

with $g(p, p_A) = \frac{1}{2}\left(2p_A(1-p_A) - p(1-\frac{p}{2})\right)$.

Proof. In order to prove the corollary, we proceed in a manner analogous to the proof of Lemma 10. Specifically, we show that the condition on the trace distance in eq. (865) is equivalent to the SDP robustness condition (216) from Theorem 8 expressed in terms of type-II error probabilities. Let $\sigma = |\psi_\sigma\rangle\langle\psi_\sigma|$, $\rho = |\psi_\rho\rangle\langle\psi_\rho|$ and recall that the type-I and type-II error probabilities are given by

$$\alpha(M; \mathcal{E}_p^{\text{dep}}(\sigma)) = \text{Tr}\left[M\mathcal{E}_p^{\text{dep}}(\sigma)\right], \quad \beta(M; \mathcal{E}_p^{\text{dep}}(\rho)) = \text{Tr}\left[(\mathbb{1} - M)\mathcal{E}_p^{\text{dep}}(\rho)\right] \tag{867}$$

with $0 \leqslant M \leqslant \mathbb{1}_d$. Let $\sigma' := \mathcal{E}_p^{\text{dep}}(\sigma)$ and $\rho' := \mathcal{E}_p^{\text{dep}}(\rho)$ and recall that a Helstrom operator for testing the null σ' against the alternative ρ' with type-I error probability α_0 takes the form (372)

$$M_{\tau(\alpha_0)} := P_{\tau(\alpha_0),+} + q_0 P_{\tau(\alpha_0),0}, \quad q_0 := \begin{cases} \frac{\alpha_0 - \alpha(P_{\tau(\alpha_0),+})}{\alpha(P_{\tau(\alpha_0),0})} & \text{if } \alpha\left(P_{\tau(\alpha_0),0}\right) \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{868}$$

where $\tau(\alpha_0) := \inf\{t \geqslant 0 \colon \alpha(P_{t,+}) \leqslant \alpha_0\}$. Let $M_A^\star := M_{\tau(1-p_A)}$ and $M_B^\star := M_{\tau(p_B)}$ and note that by assumption $p_B = 1 - p_A$ and hence $M_A^\star = M_B^\star$. The SDP robustness condition then simplifies to $\beta^*_{1-p_A}(\sigma', \rho') > 1/2$. We now proceed as follows. We first compute the spectral decomposition of the operator $\rho' - t\sigma'$ as a function of t and relate it to the fidelity between σ and ρ. With this, we derive an expression for $\alpha(P_{t,+})$ and subsequently compute $\tau(\alpha_0)$. This yields an expression for the Helstrom operator with type-I error probability $1 - p_A$ which can then be used to solve inequality (216) for the fidelity. We thus start by solving the eigenvalue problem

$$(\rho' - t\sigma')|\mu\rangle = \mu|\mu\rangle \tag{869}$$

which can be rewritten as

$$\left((1-p)\cdot(\rho-t\sigma)+\frac{p(1-t)}{2}\mathbb{1}_2\right)|\eta\rangle. \tag{870}$$

We notice that the operators $\rho'-t\sigma'$ and $\rho-t\sigma$ share the same set of eigenvectors. Furthermore, if η is an eigenvalue of $\rho-t\sigma$ with eigenvector $|\eta\rangle$, then the corresponding eigenvalue μ of $\rho'-t\sigma'$ is given by

$$\mu=(1-p)\eta+\frac{p\cdot(1-t)}{2}. \tag{871}$$

From the proof of Lemma 10, we know that the eigenvalues of $\rho-t\cdot\sigma$ are given by

$$\begin{aligned}\eta_0=\frac{1}{2}(1-t)+R>0,\ \ \eta_1=\frac{1}{2}(1-t)-R\leqslant 0\\ R=\sqrt{\frac{1}{4}(1-t)^2+t(1-|\gamma|^2)},\ \ \gamma=\langle\psi_\rho|\psi_\sigma\rangle\end{aligned} \tag{872}$$

with eigenvectors

$$\begin{aligned}|\eta_0\rangle=-\gamma A_0|\psi_\rho\rangle+(1-\eta_0)A_0|\psi_\sigma\rangle,\ \ |\eta_2\rangle=-\gamma A_2|\psi_\rho\rangle+(1-\eta_2)A_2|\psi_\sigma\rangle\\ |A_k|^{-2}=2R\left|\eta_k-1+|\gamma|^2\right|.\end{aligned} \tag{873}$$

With this, we can compute the eigenvalues μ_k and eigenprojections P_k of $\rho'-t\sigma'$ as

$$\begin{aligned}\mu_0=(1-p)\eta_0+p\cdot\frac{1-t}{2},\ \ \mu_1=(1-p)\eta_1+p\cdot\frac{1-t}{2},\\ P_0=|\eta_0\rangle\langle\eta_0|,\ \ P_1=|\eta_1\rangle\langle\eta_1|.\end{aligned} \tag{874}$$

Since $\eta_0>0\geqslant\eta_1$ for any $t\geqslant 0$, we have $\mu_0\geqslant\mu_1$ and furthermore, the eigenvalues are monotonically decreasing functions of t for $|\gamma|^2<1$. To see this, consider

$$\frac{dR}{dt}=\frac{1+t-2|\gamma|^2}{2\sqrt{(1+t)^2-4t|\gamma|^2}} \tag{875}$$

and thus for $\forall\, t\geqslant 0$ and $|\gamma|^2<1$

$$\frac{d\mu_0}{dt}=\frac{dR}{dt}-\frac{1}{2}<0\ \text{ and }\ \frac{d\mu_1}{dt}=-\frac{dR}{dt}-\frac{1}{2}<0. \tag{876}$$

Hence, since both eigenvalues are strictly positive at $t=0$, there exists exactly one ξ_k such that μ_k vanishes at ξ_k, $k=0,1$. Algebra shows that these zeroes are given by

$$\begin{aligned}\xi_0=1+\frac{2(1-|\gamma|^2)(1-p)^2}{p(2-p)}\left(1+\sqrt{1+\frac{p(2-p)}{(1-|\gamma|^2)(1-p)^2}}\right)>1,\\ \xi_1=1+\frac{2(1-|\gamma|^2)(1-p)^2}{p(2-p)}\left(1-\sqrt{1+\frac{p(2-p)}{(1-|\gamma|^2)(1-p)^2}}\right)<1.\end{aligned} \tag{877}$$

We define the functions

$$g_0(t) := \langle\eta_0|\sigma'|\eta_0\rangle = \frac{1}{2}\left(1 + \frac{(1-p)(2|\gamma|^2 - 1 - t)}{\sqrt{(1+t)^2 - 4t|\gamma|^2}}\right), \tag{878}$$

$$g_1(t) := \langle\eta_1|\sigma'|\eta_1\rangle = \frac{1}{2}\left(1 - \frac{(1-p)(2|\gamma|^2 - 1 - t)}{\sqrt{(1+t)^2 - 4t|\gamma|^2}}\right), \tag{879}$$

$$f_0(t) := \langle\eta_0|\rho'|\eta_0\rangle = \frac{1}{2}\left(1 + \frac{(1-p)(1 + t\cdot(1-2|\gamma|^2)}{\sqrt{(1+t)^2 - 4t|\gamma|^2}}\right), \tag{880}$$

$$f_1(t) := \langle\eta_0|\rho'|\eta_0\rangle = \frac{1}{2}\left(1 - \frac{(1-p)(1 + t\cdot(1-2|\gamma|^2)}{\sqrt{(1+t)^2 - 4t|\gamma|^2}}\right). \tag{881}$$

With this, we now compute $t \mapsto \alpha(P_{t,+})$ as

$$\alpha(P_{t,+}) = \mathrm{Tr}\left[\sigma' P_{t,+}\right] = \begin{cases} 1 & 0 \leqslant t < \xi_1 \\ g_0(t) & \xi_1 \leqslant t < \xi_0 \\ 0 & \xi_0 \leqslant t \end{cases} \tag{882}$$

For $\alpha_0 \in [0, 1]$, we compute $\tau(\alpha_0) := \inf\{t \geqslant 0\colon\ \alpha(P_{t,+}) \leqslant \alpha_0\}$ as

$$\tau(\alpha_0) = \begin{cases} \xi_0 & 0 \leqslant \alpha_0 \leqslant g_0(\xi_0) \\ g_0^{-1}(\alpha_0) & g_0(\xi_0) < \alpha_0 < g_0(\xi_1) \\ \xi_1 & g_0(\xi_1) \leqslant \alpha_0 < 1 \\ 0 & \alpha_0 = 1 \end{cases} \tag{883}$$

where

$$g_0^{-1}(\alpha_0) = 2|\gamma|^2 - 1 + 2(1-2\alpha_0)\sqrt{\frac{|\gamma|^2(1-|\gamma|^2)}{p(2-p) - 4\alpha_0(1-\alpha_0)}}. \tag{884}$$

To solve condition (216) we now have to distinguish different cases, depending on which interval $1 - p_A$ falls into. Firstly, if $1 - p_A = 1$, then $\tau(1-p_A) = 0$ and thus $\beta(M_A^\star) = 0$ in which case the condition can not be satisfied. If $1 - p_A \in [g_0(\xi_1), 1)$, then we have $\tau(1-p_A) = \xi_1$. In this case, it holds that $\mu_0 > 0$ and $\mu_1 = 0$ and the Helstrom operator is given by

$$M_A^\star = |\eta_0\rangle\langle\eta_0| + \frac{1 - p_A - g_0(\xi_1)}{g_1(\xi_1)}|\eta_1\rangle\langle\eta_1| \tag{885}$$

and the robustness condition reads

$$\beta(M_A^\star) = 1 - f_0(\xi_1) - \frac{1 - p_A - g_0(\xi_1)}{g_1(\xi_1)} f_1(\xi_1) > \frac{1}{2} \tag{886}$$

which cannot to be satisfied simultaneously with $1 - p_A \in [g_0(\xi_1), 1)$. If, on the other hand $1 - p_A \in (g_0(\xi_0), g_0(\xi_1))$, then $\tau(1-p_A) = g_0^{-1}(1-p_A)$ and $\mu_0 > 0 > \mu_1$.

The Helstrom operator then is given by $M_A^\star = |\eta_0\rangle\langle\eta_0|$ which leads to the robustness condition

$$1 - f_0(g_0^{-1}(1-p_A)) > \frac{1}{2} \tag{887}$$

which, together with $1-p_A \in (g_0(\xi_0),\, g_0(\xi_1))$, is equivalent to

$$|\gamma|^2 > \frac{1}{2}\left(1+\sqrt{\frac{4p_A(1-p_A)-p(2-p)}{(1-p)^2}}\right),\quad \frac{1}{2} \leqslant p_A \leqslant \frac{4-6p+3p^2}{4-4p+2p^2}. \tag{888}$$

In the last case where $1-p_A \leqslant g_0(\xi_0)$, we have $\tau(1-p_A) = \xi_0$ and thus $\mu_0 = 0 > \mu_1$. The Helstrom operator is then given by $\frac{1-p_A}{g_0(\xi_0)}|\eta_0\rangle\langle\eta_0|$, leading to the robustness condition

$$1 - \frac{1-p_A}{g_0(\xi_0)} f_0(\xi_0) > \frac{1}{2}. \tag{889}$$

Together with $1-p_A \leqslant g_0(\xi_0)$ this is equivalent to

$$|\gamma|^2 > \begin{cases} \frac{4p_A(1-p_A)-p(2-p)}{(1-p)^2(4(1-p_A)-p(2-p))}, & \text{if } \frac{1+(1-p)^2}{2} < p_A \leqslant \frac{4-6p+3p^2}{4-4p+2p^2} \\ \frac{(4-3p-2p_A(2-p))(2-p(3-2p_A))}{8(1-p)^2(1-p_A)}, & \text{if } \frac{4-6p+3p^2}{4-4p+2p^2} < p_A \leqslant \frac{4-3p}{4-2p}, \\ 0, & \text{if } p_A > \frac{4-3p}{4-2p}. \end{cases} \tag{890}$$

Finally, combining together conditions (888) and (890) leads to

$$|\gamma|^2 > \begin{cases} \frac{1}{2}\left(1+\sqrt{\frac{4p_A(1-p_A)-p(2-p)}{(1-p)^2}}\right), & \text{if } \frac{1}{2} < p_A \leqslant \frac{4-6p+3p^2}{4-4p+2p^2} \\ \frac{(4-3p-2p_A(2-p))(2-p(3-2p_A))}{8(1-p)^2(1-p_A)}, & \text{if } \frac{4-6p+3p^2}{4-4p+2p^2} < p_A \leqslant \frac{4-3p}{4-2p}, \\ 0, & \text{if } p_A > \frac{4-3p}{4-2p}. \end{cases} \tag{891}$$

Since by assumption ρ and σ are pure states the proof is completed by noting that we have

$$T(\rho,\,\sigma) = \sqrt{1-F(\rho,\,\sigma)} \tag{892}$$

by the Fuchs-van de Graaf inequality. □

F.2 PSEUDOCODE FOR ROBUSTNESS CERTIFICATION

Here we provide pseudocode for the algorithm presented in Section 9.5.1 for certifying robustness.

Algorithm 5 Robustness Certification$(\sigma, N, \alpha, \mathcal{A})$

Require: Quantum state $\sigma \in \mathcal{S}(\mathcal{H})$, number of measurement shots N, error tolerance α, a quantum classifier $\mathcal{A} = (\mathcal{E}, \{\Pi_k\}_{k\in\mathcal{C}})$.

Ensure: Predicted class k_A, prediction score p_A and robust radius r_F according to Eq. (237) in terms of fidelity.

Set counter $\mathbf{n}_k \leftarrow 0$ for every $k \in \mathcal{C}$.
for $k = 1, \ldots N$ **do**
    Apply quantum circuit $\mathcal{E}$ to initial state σ.
    Perform $|\mathcal{C}|$-outcome measurement $\{\Pi_k\}_{k\in\mathcal{C}}$ on the evolved state $\mathcal{E}(\sigma)$.
    Record measurement outcome k by setting $\mathbf{n}_k \leftarrow \mathbf{n}_k + 1$.
end for
Calculate empirical probability distribution $\hat{\mathbf{y}}_k^{(N)} \leftarrow \mathbf{n}_k N^{-1}$.
Extract the most likely class $k_A \leftarrow \arg\max_k \hat{\mathbf{y}}_k^{(N)}$.
Set $p_A \leftarrow \hat{\mathbf{y}}_{k_A}^{(N)}(\sigma) - \sqrt{\frac{-\log(\alpha)}{2N}}$.
if $p_A > 1/2$ **then**
    Calculate robust radius $r_F \leftarrow \frac{1}{2} + \sqrt{p_A(1-p_A)}$.
    return prediction k_A, class score p_A, robust radius r_F.
else
    return ABSTAIN
end if

G

ADDITIONAL RESULTS IN ROBUSTNESS INTERVALS FOR QUANTUM EXPECTATION VALUES

G.1 FIDELITY ESTIMATION

Here we give proofs for the fidelity lower bounds reported in Section 10.2.4. In the sequel, let H be a Hamiltonian with spectral decomposition

$$H = \sum_{i=0}^{m} \lambda_i \Pi_i \tag{893}$$

where λ_i are the eigenvalues (in increasing order), Π_i is the projections onto the eigenspace associated with λ_i and m is the number of distinct eigenvalues. We write $\mathrm{Eig}_H(\lambda_i)$ for the space spanned by eigenvectors of H with eigenvalue λ_i. In the following we first consider the non-degenerate case, that is when $\mathrm{Eig}_H(\lambda_0)$ is of dimension 1 and treat the degenerate case separately.

G.1.1 *The non-degenerate Case*

We first consider the non-degenerate case, in which case $\Pi_0 = |\psi_0\rangle\langle\psi_0|$.

ECKART'S CRITERION. Eckart's criterion [57] is a method to lower bound the fidelity of an approximate state σ with one of the ground states of the Hamiltonian H. We include the proof here for completeness. For general H and σ, note that

$$\begin{aligned}\langle H - \lambda_0 \mathbb{1}_d \rangle_\sigma &= \sum_{n=1}^{m} (\lambda_i - \lambda_0)\mathrm{Tr}\left[\Pi_i \sigma\right] && (894)\\ &\geqslant (\lambda_1 - \lambda_0)(1 - \langle\psi_0|\sigma|\psi_0\rangle) && (895)\end{aligned}$$

and thus

$$\langle\psi_0|\sigma|\psi_0\rangle \geqslant \frac{\lambda_1 - \langle H\rangle_\sigma}{\lambda_1 - \lambda_0}. \tag{896}$$

BOUNDS FROM (295) & (296). The fidelity bound from (295) has been shown in [151] for pure states. Here, we extend this to mixed states and will discuss the degenerate case in the next section. Recall that δ is a lower bound on the spectral gap, $\lambda_1 - \lambda_0 \geqslant \delta$. Note that

$$\begin{aligned}\langle H\rangle_\sigma &= \lambda_0\langle\psi_0|\sigma|\psi_0\rangle + \sum_{i=1}^{m} \lambda_i \mathrm{Tr}\left[\Pi_i\sigma\right] && (897)\\ &\geqslant \lambda_0\langle\psi_0|\sigma|\psi_0\rangle + \sum_{i=1}^{m} (\lambda_0 + \delta)\mathrm{Tr}\left[\Pi_i\sigma\right] && (898)\\ &= \lambda_0\langle\psi_0|\sigma|\psi_0\rangle + (\lambda_0 + \delta)(1 - \langle\psi_0|\sigma|\psi_0\rangle) && (899)\\ &= \lambda_0 + \delta(1 - \langle\psi_0|\sigma|\psi_0\rangle). && (900)\end{aligned}$$

Since by assumption λ_0 is non-degenerate and $\langle H\rangle_\sigma \leqslant \frac{1}{2}(\lambda_0+\lambda_1)$ it follows from Eckart's condition that $\langle\psi_0|\sigma|\psi_0\rangle \geqslant \frac{1}{2}$. By plugging this lower bound into the Gramian eigenvalue bound (Theorem 12), we recover Weinstein's lower bound [245] for mixed states

$$\lambda_0 \geqslant \langle H\rangle_\sigma - \Delta H_\sigma \tag{901}$$

where $(\Delta H_\sigma)^2$ is the variance of H. Using this to lower bound λ_0 in (900) and rearranging terms leads to the bound in (295)

$$\langle\psi_0|\sigma|\psi_0\rangle \geqslant 1 - \frac{\Delta H_\sigma}{\delta}. \tag{902}$$

If, on the other hand, we lower bound λ_0 in (295) by the Gramian eigenvalue lower bound (Theorem 12), we obtain the inequality

$$\langle\psi_0|\sigma|\psi_0\rangle - 1 + \frac{\Delta H_\sigma}{\delta}\sqrt{\frac{1}{\langle\psi_0|\sigma|\psi_0\rangle} - 1} \geqslant 0. \tag{903}$$

The left hand side can be rewritten as a cubic polynomial in $\langle\psi_0|\sigma|\psi_0\rangle$. Under the assumption that $\langle H\rangle_\sigma \leqslant \frac{1}{2}(\lambda_0+\lambda_1)$ we again use Eckart's condition to find that $\langle\psi_0|\sigma|\psi_0\rangle \geqslant \frac{1}{2}$. It then follows that the inequality is satisfied if

$$\langle\psi_0|\sigma|\psi_0\rangle \geqslant \frac{1}{2}\left(1 + \sqrt{1 - \left(\frac{\Delta H_\sigma}{\delta/2}\right)^2}\right) \tag{904}$$

which is the bound given in (296).

G.1.2 *The degenerate Case*

If λ_0 is degenerate, then $\Pi_0 = \sum_{j=0}^{d_0}|\psi_{0,j}\rangle\langle\psi_{0,j}|$ where d_0 denotes the dimensionality of the eigenspace associated with λ_0. In the following, we first show that if σ is a pure state, then there exists an element $|\psi\rangle \in \mathrm{Eig}_H(\lambda_0)$ for which each of the fidelity bounds holds. If, on the other hand, σ is allowed to be mixed, we construct a simple counterexample for which the fidelity bounds are violated.

PURE STATES Suppose that σ is a pure state $\sigma = |\phi\rangle\langle\phi|$. For Eckart's criterion, an analogous calculation leads to

$$\sum_{j=0}^{d_0}\left|\langle\psi_{0,j}|\phi\rangle\right|^2 \geqslant \frac{\lambda_1 - \langle H\rangle_\sigma}{\lambda_1 - \lambda_0}. \tag{905}$$

Consider the state

$$|\psi\rangle = \Gamma^{-1/2}\sum_i \langle\psi_{0,i}|\phi\rangle|\psi_{0,i}\rangle, \quad \Gamma = \sum_i |\langle\psi_{0,i}|\phi\rangle|^2 \tag{906}$$

and note that $\langle\psi|\psi\rangle = 1$ and $|\psi\rangle \in \mathrm{Eig}_H(\lambda_0)$. Furthermore, we have

$$|\langle\psi|\phi\rangle|^2 = \sum_{j=0}^{d_0}\left|\langle\psi_{0,j}|\phi\rangle\right|^2 \tag{907}$$

and hence

$$|\langle\psi|\phi\rangle|^2 \geqslant \frac{\lambda_1 - \langle H\rangle_\sigma}{\lambda_1 - \lambda_0}. \tag{908}$$

so that Eckart's criterion holds in the degenerate case for this particular choice of eigenstate $|\psi\rangle$ and for pure approximation states $|\phi\rangle$. Using again analogous calculations, we also obtain the extensions of the bounds (295) and (296) for pure approximation state $|\phi\rangle$ in the degenerate case and for the same choice of eigenstate $|\psi\rangle$.

COUNTEREXAMPLE FOR MIXED STATES If the approximation state σ is allowed to be arbitrarily mixed, the above fidelity bounds do not hold in general. Indeed, consider the Hamiltonian

$$H = U \cdot \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \mu \end{pmatrix} \cdot U^\dagger \tag{909}$$

for arbitrary $\lambda, \mu \in \mathbb{R}$ with $\lambda < \mu$ and some arbitrary unitary U. Furthermore, let σ be the maximally mixed state $\sigma = \frac{1}{3}\mathbb{1}_3$ and note that $\langle H\rangle_\sigma = \frac{2\lambda+\mu}{3}$. Thus, for any $|\psi\rangle$ we find that

$$\langle\psi|\sigma|\psi\rangle = \frac{1}{3} < \frac{2}{3} = \frac{\mu - \langle H\rangle_\sigma}{\mu - \lambda} \tag{910}$$

in violation of Eckart's criterion. To see that we can also construct a counterexample for the other two bounds, we calculate the variance

$$(\Delta H_\sigma)^2 = \langle H^2\rangle_\sigma - \langle H\rangle_\sigma^2 = \frac{2(\mu-\lambda)^2}{9} \tag{911}$$

and notice that

$$\langle\psi|\sigma|\psi\rangle = \frac{1}{3} < 1 - \frac{\sqrt{2}}{3} = 1 - \frac{\Delta H_\sigma}{\delta} \tag{912}$$

and similarly

$$\langle\psi|\sigma|\psi\rangle = \frac{1}{3} < \frac{2}{3} = \frac{1}{2}\left(1 + \sqrt{1 - \left(\frac{\Delta H_\sigma}{\delta/2}\right)^2}\right). \tag{913}$$

G.2 ADDITIONAL SIMULATIONS

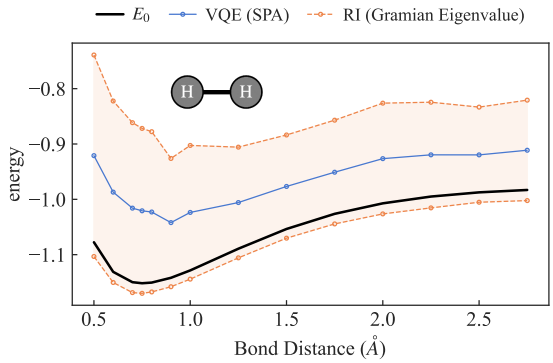


Figure 41: Bond dissociation curves and robustness interval (RI) for H_2 in a basis-set-free approach [117, 118], using the noise model of the 5-Qubit `ibmq_vigo` processor, one of the IBM Quantum Canary processors [104].

Table 35: Noisy simulations of VQE for ground state energies of LiH(2, 4) with an SPA Ansatz. The noise model consists of bitflip on single qubit gates and depolarization error on two qubit gates. The error probability for both noise channels is set to 10%.

Bond Distance (Å)	E_0	VQE (SPA)	Fidelity	Gramian Eigenvalue		Gramian Expectation	SDP	
				lower bound	upper bound	lower bound	lower bound	upper bound
0.50	−7.21863	−6.88805	0.642	−7.25193	−6.52475	−7.32715	−7.34604	−5.78115
0.75	−7.70845	−7.29597	0.643	−7.74638	−6.84531	−7.83840	−7.86084	−5.93280
1.00	−7.90403	−7.49060	0.645	−7.94329	−7.03703	−8.03507	−8.05745	−6.10197
1.25	−7.97808	−7.57789	0.642	−8.01822	−7.13739	−8.10760	−8.12808	−6.23202
1.40	−7.99541	−7.61078	0.645	−8.03193	−7.18843	−8.12397	−8.14492	−6.30208
1.50	−8.00062	−7.62723	0.645	−8.03743	−7.21626	−8.12843	−8.14912	−6.34578
1.60	−8.00251	−7.63936	0.646	−8.03613	−7.24368	−8.12772	−8.14878	−6.40143
1.70	−8.00213	−7.65290	0.644	−8.03798	−7.26681	−8.12450	−8.14473	−6.45257
1.75	−8.00127	−7.65684	0.646	−8.03181	−7.28154	−8.12101	−8.14171	−6.49063
2.00	−7.99319	−7.68221	0.644	−8.02412	−7.34029	−8.10203	−8.12208	−6.63209
2.25	−7.98161	−7.69596	0.639	−8.01110	−7.38116	−8.08242	−8.10091	−6.75887
2.50	−7.96941	−7.70682	0.634	−7.99792	−7.41511	−8.06577	−8.08292	−6.85437
2.75	−7.95765	−7.71482	0.632	−7.98661	−7.44292	−8.04982	−8.06660	−6.94268
3.00	−7.94669	−7.71586	0.624	−7.97520	−7.45615	−8.06343	−8.07915	−6.96150
3.25	−7.93706	−7.72211	0.623	−7.96702	−7.47666	−8.09013	−8.11182	−6.95623
3.50	−7.92842	−7.72265	0.613	−7.96128	−7.48390	−8.10708	−8.13833	−6.94904
3.75	−7.92096	−7.72526	0.606	−7.95586	−7.49476	−8.11762	−8.16116	−6.94122
4.00	−7.91490	−7.72740	0.597	−7.95494	−7.50039	−8.13010	−8.18053	−6.93847
4.25	−7.90968	−7.72717	0.583	−7.95372	−7.50055	−8.14249	−8.19643	−6.93280
4.50	−7.90590	−7.72953	0.568	−7.95324	−7.50594	−8.15344	−8.21024	−6.92201
4.75	−7.90306	−7.73045	0.551	−7.95891	−7.50178	−8.16757	−8.22021	−6.92056
5.00	−7.90106	−7.72925	0.527	−7.96652	−7.49266	−8.18161	−8.22696	−6.91288
5.25	−7.89982	−7.73207	0.516	−7.96729	−7.49630	−8.18806	−8.23073	−6.91086

Table 36: Noisy simulations of VQE for ground state energies of LiH(2, 4) with an UpCCGSD Ansatz. The noise model consists of bitflip on single qubit gates and depolarization error on two qubit gates. The error probability for both noise channels is set to 10%.

Bond Distance (Å)	E_0	VQE (UpCCGSD)	Fidelity	Gramian Eigenvalue		Gramian Expectation	SDP	
				lower bound	upper bound	lower bound	lower bound	upper bound
0.50	−7.21863	−6.66418	0.118	−7.71356	−5.61455	−7.34604	−7.34604	−5.48348
0.75	−7.70845	−6.99642	0.120	−8.30129	−5.69055	−7.86084	−7.86084	−5.55101
1.00	−7.90403	−7.17549	0.122	−8.49462	−5.85839	−8.05745	−8.05745	−5.71484
1.25	−7.97808	−7.27489	0.121	−8.56334	−5.98558	−8.12808	−8.12808	−5.85695
1.40	−7.99541	−7.30592	0.107	−8.61254	−6.00004	−8.14492	−8.14492	−5.93376
1.50	−8.00062	−7.34297	0.117	−8.56754	−6.12004	−8.14912	−8.14912	−5.99011
1.60	−8.00251	−7.37233	0.120	−8.53701	−6.20478	−8.14878	−8.14878	−6.05390
1.70	−8.00213	−7.39912	0.120	−8.52712	−6.26860	−8.14473	−8.14473	−6.12434
1.75	−8.00127	−7.41376	0.121	−8.51673	−6.31010	−8.14171	−8.14171	−6.16141
2.00	−7.99319	−7.47053	0.120	−8.46034	−6.48104	−8.12208	−8.12208	−6.34924
2.25	−7.98161	−7.51937	0.120	−8.41918	−6.62024	−8.10091	−8.10091	−6.51810
2.50	−7.96941	−7.55501	0.118	−8.38562	−6.72552	−8.08292	−8.08292	−6.65920
2.75	−7.95765	−7.58323	0.116	−8.36295	−6.80337	−8.06660	−8.06660	−6.77282
3.00	−7.94669	−7.60655	0.116	−8.33990	−6.87168	−8.07915	−8.07915	−6.81481
3.25	−7.93706	−7.62472	0.115	−8.32053	−6.92821	−8.11182	−8.11182	−6.82651
3.50	−7.92842	−7.64013	0.115	−8.31261	−6.96878	−8.13833	−8.13833	−6.83591
3.75	−7.92096	−7.65034	0.103	−8.33392	−6.96726	−8.16116	−8.16116	−6.84353
4.00	−7.91490	−7.66562	0.111	−8.31035	−7.02064	−8.18053	−8.18053	−6.85062
4.25	−7.90968	−7.68425	0.089	−8.34869	−7.01927	−8.19643	−8.19643	−6.85685
4.50	−7.90590	−7.69349	0.089	−8.35523	−7.03175	−8.21024	−8.21024	−6.86264
4.75	−7.90306	−7.70511	0.090	−8.35773	−7.05194	−8.22021	−8.22021	−6.86835
5.00	−7.90106	−7.70031	0.098	−8.36531	−7.03541	−8.22696	−8.22696	−6.87387
5.25	−7.89982	−7.70455	0.077	−8.45740	−6.95222	−8.23073	−8.23073	−6.87932

BIBLIOGRAPHY

[1] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. "The power of quantum neural networks." In: *Nature Computational Science* 1.6 (2021), pp. 403–409.

[2] M. Abramowitz and I. A. (Eds.) Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*. New York: New York: Dover, 1972.

[3] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing." In: *Medical physics* 45.3 (2018), pp. 1150–1158.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. "Invariant risk minimization." In: *arXiv preprint arXiv:1907.02893* (2019).

[5] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Sergio Boixo, Michael Broughton, Bob B Buckley, David A Buell, et al. "Hartree-Fock on a superconducting qubit quantum computer." In: *Science* 369.6507 (2020), pp. 1084–1089. URL: https://doi.org/10.1126/science.abb9811.

[6] Anish Athalye, Nicholas Carlini, and David Wagner. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples." In: *Proceedings of the 35th International Conference on Machine Learning*. 2018, pp. 274–283.

[7] Anish Athalye, Nicholas Carlini, and David Wagner. "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples." In: *International conference on machine learning*. PMLR. 2018, pp. 274–283.

[8] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. "Certifying Geometric Robustness of Neural Networks." In: *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019, pp. 15287–15297.

[9] Sara Beery, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 456–473.

[10] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. "Robust solutions of optimization problems affected by uncertain probabilities." In: *Management Science* 59.2 (2013), pp. 341–357.

[11] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. "Parameterized quantum circuits as machine learning models." In: *Quantum Sci. Technol.* 4.4 (2019), p. 043001.

[12] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, et al. "Noisy intermediate-scale quantum algorithms." In: *Rev. Mod. Phys.* 94 (1 2022), p. 015004. URL: https://link.aps.org/doi/10.1103/RevModPhys.94.015004.

[13] Rajendra Bhatia. *Matrix analysis*. Vol. 169. Springer, 2013.

[14] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. "Quantum machine learning." In: *Nature* 549 (2017), pp. 195–202.

[15] Battista Biggio, Blaine Nelson, and Pavel Laskov. "Poisoning Attacks against Support Vector Machines." In: *Proceedings of the 29th International Coference on Machine Learning*. 2012, 1467–1474.

[16] Jose Blanchet and Karthyek Murthy. "Quantifying distributional model risk via optimal transport." In: *Mathematics of Operations Research* 44.2 (2019), pp. 565–600.

[17] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. "Random Smoothing Might be Unable to Certify ℓ_∞ Robustness for High-Dimensional Images." In: *Journal of Machine Learning Research (JMLR)* 21.211 (2020), pp. 1–21.

[18] Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping, Micah Goldblum, Tom Goldstein, and Arjun Gupta. "Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff." In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 3855–3859.

[19] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. "A large annotated corpus for learning natural language inference." In: *EMNLP*. 2015.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[21] Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. "Anomalous Example Detection in Deep Learning: A Survey." In: *IEEE Access* 8 (2020), pp. 132330–132347.

[22] Zhenyu Cai, Ryan Babbush, Simon C Benjamin, Suguru Endo, William J Huggins, Ying Li, Jarrod R McClean, and Thomas E O'Brien. "Quantum error mitigation." In: *arXiv preprint arXiv:2210.00921* (2022).

[23] A Robert Calderbank and Peter W Shor. "Good quantum error-correcting codes exist." In: *Physical Review A* 54.2 (1996), p. 1098.

[24] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. "Poisoning web-scale training datasets is practical." In: *arXiv preprint arXiv:2302.10149* (2023).

[25] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. "Hidden voice commands." In: *25th USENIX security symposium (USENIX security 16)*. 2016, pp. 513–530.

[26] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. "Are aligned neural networks adversarially aligned?" In: *arXiv preprint arXiv:2306.15447* (2023).

[27] Nicholas Carlini and David Wagner. "Adversarial examples are not easily detected: Bypassing ten detection methods." In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM. 2017, pp. 3–14.

[28] Nicholas Carlini and David Wagner. "Adversarial examples are not easily detected: Bypassing ten detection methods." In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 2017, pp. 3–14.

[29] Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks." In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 39–57.

[30] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. "Unlabeled Data Improves Adversarial Robustness." In: *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019, pp. 11192–11203.

[31] Yelp Dataset Challenge. data retrieved from Yelp Dataset Challenge, https://www.yelp.com/dataset/challenge.

[32] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. "Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering." In: *SafeAI@ AAAI*. 2019.

[33] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. "Deepinspect: a black-box Trojan detection and mitigation framework for deep neural networks." In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press. 2019, pp. 4658–4664.

[34] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. "Targeted backdoor attacks on deep learning systems using data poisoning." In: *arXiv:1712.05526* (2017).

[35] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. "Maximum resilience of artificial neural networks." In: *Automated Technology for Verification and Analysis: 15th International Symposium, ATVA 2017, Pune, India, October 3–6, 2017, Proceedings 15*. Springer. 2017, pp. 251–268.

[36] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. "Sentinet: Detecting localized universal attacks against deep learning systems." In: *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE. 2020, pp. 48–54.

[37] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. "Parseval networks: Improving robustness to adversarial examples." In: *International Conference on Machine Learning*. PMLR. 2017, pp. 854–863.

[38] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. "Ranking and empirical minimization of U-statistics." In: *The Annals of Statistics* 36.2 (2008), pp. 844–874.

[39] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. "Certified Adversarial Robustness via Randomized Smoothing." In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. 2019, pp. 1310–1320.

[40] Maurice Cohen and Tova Feldmann. "Lower bounds to eigenvalues." In: *Canadian Journal of Physics* 47.17 (1969), pp. 1877–1879.

[41] Iris Cong, Soonwon Choi, and Mikhail D Lukin. "Quantum convolutional neural networks." In: *Nature Physics* 15.12 (2019), pp. 1273–1278.

[42] Zac Cranko, Zhan Shi, Xinhua Zhang, Richard Nock, and Simon Kornblith. "Generalised lipschitz regularisation equals distributional robustness." In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2178–2188.

[43] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. "RobustBench: A Standardized Adversarial Robustness Benchmark." In: *arXiv preprint arXiv:2010.09670* 10 (2020).

[44] Dengxin Dai and Luc Van Gool. "Dark model adaptation: Semantic image segmentation from daytime to nighttime." In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 3819–3824.

[45] Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy Liang, et al. "Enabling Certification of Verification-Agnostic Networks via Memory-Efficient Semidefinite Programming." In: *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 5318–5331.

[46] Chandler Davis and William Morton Kahan. "The rotation of eigenvectors by a perturbation. III." In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pp. 1–46. URL: https://doi.org/10.1137/0707001.

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[48] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018).

[49] Samuel Drews, Aws Albarghouthi, and Loris D'Antoni. "Proving Data-Poisoning Robustness in Decision Trees." In: *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation*. PLDI 2020. London, UK: Association for Computing Machinery, 2020, 1083–1097.

[50] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Dacheng Tao, and Nana Liu. "Quantum noise protects quantum classifiers against adversaries." In: Preprint at https://arxiv.org/abs/2003.09416 (2020).

[51] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[52] John C Duchi, Peter W Glynn, and Hongseok Namkoong. "Statistics of robust optimization: A generalized empirical likelihood approach." In: *Mathematics of Operations Research* (2021).

[53] John C Duchi and Hongseok Namkoong. "Learning models with uniform performance via distributionally robust optimization." In: *The Annals of Statistics* 49.3 (2021), pp. 1378–1406.

[54] John Duchi and Hongseok Namkoong. "Variance-based regularization with convex objectives." In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 2450–2504.

[55] Vedran Dunjko and Hans J Briegel. "Machine learning & artificial intelligence in the quantum domain: a review of recent progress." In: *Rep. Prog. Phys.* 81.7 (2018), p. 074001.

[56] Krishnamurthy Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. "A Framework for Robustness Certification of Smoothed Classifiers using f-Divergences." In: *2020 International Conference on Learning Representations (ICLR)*. 2020.

[57] Carl Eckart. "The Theory and Calculation of Screening Constants." In: *Phys. Rev.* 36 (5 1930), pp. 878–892. DOI: 10.1103/PhysRev.36.878. URL: https://link.aps.org/doi/10.1103/PhysRev.36.878.

[58] Erik Englesson and Hossein Azizpour. "Generalized Jensen-Shannon Divergence Loss for Learning with Noisy Labels." In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. 2021. URL: https://openreview.net/forum?id=TiwPYwg3IRf.

[59] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. "Exploring the Landscape of Spatial Robustness." In: *2019 International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 1802–1811.

[60] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust Physical-World Attacks on Deep Learning Visual Classification." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 1625–1634.

[61] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Robust physical-world attacks on deep learning visual classification." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1625–1634.

[62] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A quantum approximate optimization algorithm." In: *arXiv:1411.4028* (2014). URL: https://arxiv.org/abs/1411.4028.

[63] Edward Farhi, Hartmut Neven, et al. "Classification with Quantum Neural Networks on Near Term Processors." In: *Quantum Rev. Lett.* 1.2 (2020) (2020), pp. 129–153.

[64] Claudio Ferrari, Mark Niklas Müller, Nikola Jovanović, and Martin Vechev. "Complete Verification via Multi-Neuron Relaxation Guided Branch-and-Bound." In: *International Conference on Learning Representations*. 2022.

[65] Marc Fischer, Maximilian Baader, and Martin Vechev. "Certified Defense to Image Transformations via Randomized Smoothing." In: *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 8404–8417.

[66] Rui Gao and Anton J Kleywegt. "Distributionally robust stochastic optimization with Wasserstein distance." In: *arXiv preprint arXiv:1604.02199* (2016).

[67] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. "STRIP: A Defence against Trojan Attacks on Deep Neural Networks." In: *Proceedings of the 35th Annual Computer Security Applications Conference*. New York, NY, USA: Association for Computing Machinery, 2019, 113–125.

[68] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. "Ai2: Safety and robustness certification of neural networks with abstract interpretation." In: *2018 IEEE symposium on security and privacy (SP)*. IEEE. 2018, pp. 3–18.

[69] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. "Breaking Certified Defenses: Semantic Adversarial Examples with Spoofed Robustness Certificates." In: *2020 International Conference on Learning Representations (ICLR)*. OpenReview, 2020.

[70] Tejas Gokhale, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J Thiagarajan, Chitta Baral, and Yezhou Yang. "Attribute-Guided Adversarial Training for Robustness to Natural Perturbations." In: *2021 AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 35. Association for the Advancement of Artificial Intelligence Press, 2021, pp. 7574–7582.

[71] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." In: *International Conference on Learning Representations (ICLR)* (2015).

[72] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. "Robust empirical optimization is almost the same as mean–variance optimization." In: *Operations research letters* 46.4 (2018), pp. 448–452.

[73] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. "On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models." In: Preprint at https://arxiv.org/abs/1810.12715 (2018).

[74] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. "Scalable Verified Training for Provably Robust Image Classification." In: *2019 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 4842–4851.

[75] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks." In: *IEEE Access* 7 (2019), pp. 47230–47244.

[76] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." In: *arXiv:1708.06733* (2017).

[77] Ji Guan, Wang Fang, and Mingsheng Ying. "Robustness Verification of Quantum Machine Learning." In: Preprint at https://arxiv.org/abs/2008.07230 (2020).

[78] Ishaan Gulrajani and David Lopez-Paz. "In Search of Lost Domain Generalization." In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=lQdXeXDoWtI.

[79] James A Hanley and Barbara J McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1 (1982), pp. 29–36.

[80] Robert J Harrison, Gregory Beylkin, Florian A Bischoff, Justus A Calvin, George I Fann, Jacob Fosso-Tande, Diego Galindo, Jeff R Hammond, Rebecca Hartman-Baker, Judith C Hill, et al. "MADNESS: A multiresolution, adaptive numerical environment for scientific simulation." In: *SIAM J. Sci. Comput.* 38.5 (2016), S123–S142. URL: https://doi.org/10.1137/15M1026171.

[81] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. "Supervised learning with quantum-enhanced feature spaces." In: *Nature* 567.7747 (2019), pp. 209–212.

[82] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. "SPECTRE: defending against backdoor attacks using robust statistics." In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 4129–4139. URL: https://proceedings.mlr.press/v139/hayase21a.html.

[83] Jamie Hayes. "Extensions and Limitations of Randomized Smoothing for Robustness Guarantees." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2020, pp. 3413–3421.

[84] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1026–1034.

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[86] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "Deberta: Decoding-enhanced bert with disentangled attention." In: *arXiv preprint arXiv:2006.03654* (2020).

[87] T. Helgaker, P. Jorgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. Wiley, 2014. ISBN: 9781119019558. URL: https://books.google.ca/books?id=lNVLBAAAQBAJ.

[88] Carl W Helstrom. "Detection theory and quantum mechanics." In: *Inform. Control* 10.3 (1967), pp. 254–291.

[89] Carl W. Helstrom. "Detection theory and quantum mechanics (II)." In: *Information and Control* 13.2 (1968), pp. 156–171. URL: https://www.sciencedirect.com/science/article/pii/S0019995868907468.

[90] Carl Wilhelm Helstrom. *Quantum detection and estimation theory*. New York, NY: Academic Press, 1976.

[91] Dan Hendrycks and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations." In: *2018 International Conference on Learning Representations (ICLR)*. OpenReview, 2018.

[92] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty." In: *2020 International Conference on Learning Representations (ICLR)*. OpenReview, 2020.

[93] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." In: *arXiv preprint arXiv:1503.02531* (2015).

[94] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables." In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30.

[95] Alexander S Holevo. "Statistical decision theory for quantum systems." In: *J. Multivar. Anal.* 3.4 (1973), pp. 337–394.

[96] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. DOI: 10.1017/CBO9780511810817.

[97] Hossein Hosseini, Yize Chen, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. "Blocking transferability of adversarial examples in black-box learning systems." In: *arXiv preprint arXiv:1703.04318* (2017).

[98] Hossein Hosseini and Radha Poovendran. "Semantic Adversarial Examples." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2018, pp. 1614–1619.

[99] Jeremy Howard. *Imagenette*. URL: https://github.com/fastai/imagenette/.

[100] Weiwei Hu and Ying Tan. "Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN." In: *arXiv preprint arXiv:1702.05983* 02 (2017).

[101] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. "Densely connected convolutional networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[102] William J Huggins, Jarrod R McClean, Nicholas C Rubin, Zhang Jiang, Nathan Wiebe, K Birgitta Whaley, and Ryan Babbush. "Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers." In: *npj Quantum Information* 7.1 (2021), pp. 1–9. URL: https://doi.org/10.1038/s41534-020-00341-7.

[103] Hisham Husain. "Distributional Robustness with IPMs and links to Regularization and GANs." In: *arXiv preprint arXiv:2006.04349* (2020).

[104] *IBM Quantum*. https://quantum-computing.ibm.com/. 2021.

[105] Jean Jacod and Philip Protter. *Probability essentials*. Berlin: Springer Science & Business Media, 2000.

[106] Jongheon Jeong and Jinwoo Shin. "Consistency Regularization for Certified Robustness of Smoothed Classifiers." In: *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 10558–10570.

[107] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. "Intrinsic Certified Robustness of Bagging against Data Poisoning Attacks." In: *AAAI*. 2021.

[108] Jinyuan Jia, Yupei Liu, Xiaoyu Cao, and Neil Zhenqiang Gong. "Certified Robustness of Nearest Neighbors against Data Poisoning and Backdoor Attacks." In: *AAAI*. 2022.

[109] Bojan Karlaš, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. "Nearest Neighbor Classifiers over Incomplete Information: From Certain Answers to Certain Predictions." In: *Proc. VLDB Endow.* 14.3 (2020), 255–267.

[110] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. "Reluplex: An efficient SMT solver for verifying deep neural networks." In: *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*. Springer. 2017, pp. 97–117.

[111] Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, et al. "The marabou framework for verification and analysis of deep neural networks." In: *Computer Aided Verification: 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I 31*. Springer. 2019, pp. 443–452.

[112] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. "Learning not to learn: Training deep neural networks with biased data." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9012–9020.

[113] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. "Wilds: A benchmark of in-the-wild distribution shifts." In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5637–5664.

[114] Jakob S Kottmann, Sumner Alperin-Lea, Teresa Tamayo-Mendoza, Alba Cervera-Lierta, Cyrille Lavigne, Tzu-Ching Yen, Vladyslav Verteletskyi, Philipp Schleich, Abhinav Anand, Matthias Degroote, et al. "Tequila: A platform for rapid development of quantum algorithms." In: *Quantum Science and Technology* 6.2 (2021), p. 024009. URL: https://doi.org/10.1088/2058-9565/abe567.

[115] Jakob S. Kottmann, Abhinav Anand, and Alán Aspuru-Guzik. "A feasible approach for automatically differentiable unitary coupled-cluster on quantum computers." In: *Chem. Sci.* (2021), pp. –. DOI: 10.1039/D0SC06627C. URL: http://dx.doi.org/10.1039/D0SC06627C.

[116] Jakob S. Kottmann and Alán Aspuru-Guzik. "Optimized low-depth quantum circuits for molecular electronic structure using a separable-pair approximation." In: *Phys. Rev. A* 105 (3 2022), p. 032449. URL: https://link.aps.org/doi/10.1103/PhysRevA.105.032449.

[117] Jakob S. Kottmann, Florian A. Bischoff, and Edward F. Valeev. "Direct determination of optimal pair-natural orbitals in a real-space representation: The second-order Moller–Plesset energy." In: *The Journal of Chemical Physics* 152.7 (2020), p. 074105. URL: https://doi.org/10.1063/1.5141880.

[118] Jakob S. Kottmann, Philipp Schleich, Teresa Tamayo-Mendoza, and Alán Aspuru-Guzik. "Reducing Qubit Requirements while Maintaining Numerical Precision for the Variational Quantum Eigensolver: A Basis-Set-Free Approach." In: *J. Phys. Chem. Lett.* 12.1 (2021), pp. 663–673. URL: https://doi.org/10.1021/acs.jpclett.0c03410.

[119] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. 2009.

[120] Alex Krizhevsky, Geoffrey Hinton, et al. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009.

[121] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. "Wasserstein distributionally robust optimization: Theory and applications in machine learning." In: *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019, pp. 130–166.

[122] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. "Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness." In: *2020 International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 5458–5467.

[123] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[124] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." In: *arXiv preprint arXiv:1611.01236* (2016).

[125] Ryan LaRose and Brian Coyle. "Robust data encodings for quantum classifiers." In: *Phys. Rev. A* 102 (3 2020), p. 032420.

[126] Henry Lam. "Robust sensitivity analysis for stochastic systems." In: *Mathematics of Operations Research* 41.4 (2016), pp. 1248–1275.

[127] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[128] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. "Certified Robustness to Adversarial Examples with Differential Privacy." In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 656–672.

[129] Jaeho Lee and Maxim Raginsky. "Minimax statistical learning with wasserstein distances." In: *arXiv preprint arXiv:1705.07815* (2017).

[130] Joonho Lee, William J. Huggins, Martin Head-Gordon, and K. Birgitta Whaley. "Generalized Unitary Coupled Cluster Wave functions for Quantum Computation." In: *Journal of Chemical Theory and Computation* 15.1 (2019), pp. 311–324. URL: https://doi.org/10.1021/acs.jctc.8b01004.

[131] Alexander Levine and Soheil Feizi. "Deep Partition Aggregation: Provable Defenses against General Poisoning Attacks." In: *9th International Conference on Learning Representations*. 2021.

[132] Alexander Levine, Aounon Kumar, Thomas Goldstein, and Soheil Feizi. "Tight second-order certificates for randomized smoothing." In: *arXiv preprint arXiv:2010.10549* (2020).

[133] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. "Certified Adversarial Robustness with Additive Noise." In: *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019, pp. 9459–9469.

[134] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. "Data poisoning attacks on factorization-based collaborative filtering." In: *Advances in neural information processing systems*. 2016, pp. 1885–1893.

[135] Linyi Li, Tao Xie, and Bo Li. "Sok: Certified robustness for deep neural networks." In: *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2023, pp. 1289–1310.

[136] Linyi Li, Zexuan Zhong, Bo Li, and Tao Xie. "Robustra: Training Provable Robust Neural Networks over Reference Adversarial Space." In: *2019 International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 4711–4717.

[137] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. "A structured self-attentive sentence embedding." In: *arXiv preprint arXiv:1703.03130* (2017).

[138] Nana Liu and Peter Wittek. "Vulnerability of quantum classification to adversarial perturbations." In: *Phys. Rev. A* 101 (6 2020), p. 062331.

[139] Seth Lloyd. "Enhanced Sensitivity of Photodetection via Quantum Illumination." In: *Science* 321.5895 (2008), pp. 1463–1465.

[140] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. "Quantum embeddings for machine learning." In: Preprint at https://arxiv.org/abs/2001.03622 (2020).

[141] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. "Quantum adversarial machine learning." In: *Phys. Rev. Research* 2 (3 2020), p. 033212.

[142] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Michael E. Houle, Dawn Song, and James Bailey. "Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality." In: *International Conference on Learning Representations*. 2018.

[143] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. "Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality." In: *2018 International Conference on Learning Representations (ICLR)*. OpenReview, 2018.

[144] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. "Data Poisoning against Differentially-Private Learners: Attacks and Defenses." In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 4732–4738.

[145] J. K. L. MacDonald. "On the Modified Ritz Variation Method." In: *Phys. Rev.* 46 (9 1934), pp. 828–828. DOI: 10.1103/PhysRev.46.828. URL: https://link.aps.org/doi/10.1103/PhysRev.46.828.

[146] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks." In: *International Conference on Learning Representations (ICLR)*. 2018.

[147] William Matthews and Stephanie Wehner. "Finite blocklength converse bounds for quantum channels." In: *IEEE Trans. Inf. Theory* 60.11 (2014), pp. 7317–7329.

[148] Andreas Maurer and Massimiliano Pontil. "Empirical Bernstein bounds and sample variance penalization." In: *Proceedings of the Twenty Second Annual Conference on Computational Learning Theory*. 2009.

[149] Sam McArdle, Tyson Jones, Suguru Endo, Ying Li, Simon C. Benjamin, and Xiao Yuan. "Variational ansatz-based quantum simulation of imaginary time evolution." In: *npj Quantum Information* 5.1 (2019), p. 75. URL: https://doi.org/10.1038/s41534-019-0187-2.

[150] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. "Barren plateaus in quantum neural network training landscapes." In: *Nature Communications* 9.1 (2018), p. 4812. URL: https://doi.org/10.1038/s41467-018-07090-4.

[151] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. "The theory of variational hybrid quantum-classical algorithms." In: *New Journal of Physics* 18.2 (2016), p. 023023. URL: https://doi.org/10.1088/1367-2630/18/2/023023.

[152] Jarrod R McClean, Nicholas C Rubin, Kevin J Sung, Ian D Kivlichan, Xavier Bonet-Monroig, Yudong Cao, Chengyu Dai, E Schuyler Fried, Craig Gidney, Brendan Gimby, et al. "OpenFermion: the electronic structure package for quantum computers." In: *Quantum Science and Technology* 5.3 (2020), p. 034014. URL: https://doi.org/10.1088/2058-9565/ab8ebc.

[153] Matthew Mirman, Timon Gehr, and Martin Vechev. "Differentiable Abstract Interpretation for Provably Robust Neural Networks." In: *2018 International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 3575–3583.

[154] Jeet Mohapatra, Ching-Yun Ko, Lily Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. "Hidden Cost of Randomized Smoothing." In: *2021 International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2021, pp. 4033–4041.

[155] Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. "Higher-order certification for randomized smoothing." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4501–4511.

[156] Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. "Towards Verifying Robustness of Neural Networks Against A Family of Semantic Perturbations." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 244–252.

[157] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. "DeepStack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker." In: *Science* 356.6337 (2017), pp. 508–513.

[158] Mario Motta, Chong Sun, Adrian T. K. Tan, Matthew J. O'Rourke, Erika Ye, Austin J. Minnich, Fernando G. S. L. Brandão, and Garnet Kin-Lic Chan. "Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution." In: *Nature Physics* 16.2 (2020), pp. 205–210. URL: https://doi.org/10.1038/s41567-019-0704-4.

[159] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli. "Towards poisoning of deep learning algorithms with back-gradient optimization." In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM. 2017, pp. 27–38.

[160] Kouhei Nakaji and Naoki Yamamoto. "Expressibility of the alternating layered ansatz for quantum computation." In: *Quantum* 5 (2021), p. 434. URL: https://doi.org/10.22331/q-2021-04-19-434.

[161] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D Joseph, Benjamin IP Rubinstein, Udam Saini, Charles Sutton, J Doug Tygar, and Kai Xia. "Exploiting machine learning to subvert your spam filter." In: *LEET* 8.1-9 (2008), pp. 16–17.

[162] Jerzy Neyman and Egon Sharpe Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses." In: *Phil. Trans. Roy. Statistical Soc. A* 231.694-706 (1933), pp. 289–337.

[163] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. "Nonparametric estimation of the likelihood ratio and divergence functionals." In: *2007 IEEE International Symposium on Information Theory*. IEEE. 2007, pp. 2016–2020.

[164] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. "Estimating divergence functionals and the likelihood ratio by convex risk minimization." In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5847–5861.

[165] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. "Adversarial Robustness Toolbox v1. 0.0." In: *arXiv:1807.01069* (2018).

[166] OpenCV. *OpenCV: Transformations of Images*. https://docs.opencv.org/master/dd/d52/tutorial_js_geometric_transformations.html. 2020. (Visited on 12/01/2020).

[167] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. "Entanglement-Induced Barren Plateaus." In: *PRX Quantum* 2 (4 2021), p. 040316. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.040316.

[168] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. "Distillation as a defense to adversarial perturbations against deep neural networks." In: *2016 IEEE symposium on security and privacy (SP)*. IEEE. 2016, pp. 582–597.

[169] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[170] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. "DeepXplore: Automated Whitebox Testing of Deep Learning Systems." In: *2017 Symposium on Operating Systems Principles (SOSP)*. Association for Computing Machinery, 2017, pp. 1–18.

[171] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. "Towards Practical Verification of Machine Learning: The Case of Computer Vision Systems." In: *arXiv preprint: arXiv:1712.01785* 12 (2017).

[172] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only." In: *arXiv preprint arXiv:2306.01116* (2023).

[173] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. "A variational eigenvalue solver on a photonic quantum processor." In: *Nature Communications* 5.1 (2014), p. 4213. URL: https://doi.org/10.1038/ncomms5213.

[174] John Preskill. "Quantum computing in the NISQ era and beyond." In: *Quantum* 2 (2018), p. 79. URL: https://doi.org/10.22331/q-2018-08-06-79.

[175] Luca Pulina and Armando Tacchella. "An abstraction-refinement approach to verification of artificial neural networks." In: *Computer Aided Verification: 22nd International Conference, CAV 2010, Edinburgh, UK, July 15-19, 2010. Proceedings 22*. Springer. 2010, pp. 243–257.

[176] PyTorch. *torchvision.models — Torchvison 0.10.0 documentation*. https://pytorch.org/vision/stable/models.html. 2021. (Visited on 09/15/2021).

[177] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. "Semidefinite Relaxations for Certifying Robustness to Adversarial Examples." In: *2018 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2018, pp. 10877–10887.

[178] J.W. Rayleigh. "In finding the correction for the open end of an organ-pipe." In: *Phil. Trans.* 161 (1870), p. 77.

[179] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. "Quantum Support Vector Machine for Big Data Classification." In: *Phys. Rev. Lett.* 113 (13 2014), p. 130503.

[180] Yankun Ren, Longfei Li, and Jun Zhou. "Simtrojan: Stealthy Backdoor Attack." In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pp. 819–823.

[181] Robert D. Richtmeyer. *Principles of Advanced Mathematical Physics*. Vol. 1. Springer, 1978.

[182] Walter Ritz. "Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik." In: *Journal für die reine und angewandte Mathematik* 135 (1909), pp. 1–61. URL: http://eudml.org/doc/149295.

[183] Joseph D Romano, Trang T Le, William La Cava, John T Gregg, Daniel J Goldberg, Praneel Chakraborty, Natasha L Ray, Daniel Himmelstein, Weixuan Fu, and Jason H Moore. "PMLB v1.0: an open-source dataset collection for benchmarking machine learning methods." In: *Bioinformatics* 38.3 (2021), pp. 878–880.

[184] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. "Certified robustness to label-flipping attacks via randomized smoothing." In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8230–8241.

[185] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge." In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[186] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. "Hidden trigger backdoor attacks." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 11957–11965.

[187] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." In: *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019, pp. 11289–11300.

[188] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. "A Convex Relaxation Barrier to Tight Robustness Verification of Neural Networks." In: *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019, pp. 9835–9846.

[189] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. "Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models." In: *2018 International Conference on Learning Representations (ICLR)*. OpenReview, 2018.

[190] Herbert Scarf. "A min-max solution of an inventory problem." In: *Studies in the mathematical theory of inventory and production* (1958).

[191] Maria Schuld and Nathan Killoran. "Quantum machine learning in feature Hilbert spaces." In: *Phys. Rev. Lett.* 122.4 (2019), p. 040504.

[192] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. "An introduction to quantum machine learning." In: *Contemp. Phys.* 56 (2015), pp. 172–185.

[193] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. "Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks." In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 9389–9398.

[194] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. "Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks." In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9389–9398.

[195] Gael Sentís, Alex Monras, Ramon Muñoz-Tapia, John Calsamiglia, and Emilio Bagan. "Unsupervised classification of quantum data." In: *Phys. Rev. X* 9.4 (2019), p. 041029.

[196] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks." In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, 6106–6116.

[197] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. "Adversarial Training for Free!" In: *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019, pp. 3358–3369.

[198] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. "Regularization via mass transportation." In: *Journal of Machine Learning Research* 20.103 (2019), pp. 1–68.

[199] Peter W Shor. "Scheme for reducing decoherence in quantum computer memory." In: *Physical review A* 52.4 (1995), R2493.

[200] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. "Mastering the Game of Go without Human Knowledge." In: *Nature* 550.7676 (2017), pp. 354–359.

[201] Sukin Sim, Peter D Johnson, and Alán Aspuru-Guzik. "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms." In: *Advanced Quantum Technologies* 2.12 (2019), p. 1900070. URL: https://doi.org/10.1002/qute.201900070.

[202] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. "An abstract domain for certifying neural networks." In: *Proceedings of the ACM on Programming Languages* 3.POPL (2019), pp. 1–30.

[203] Aman Sinha, Hongseok Namkoong, and John Duchi. "Certifiable Distributional Robustness with Principled Adversarial Training." In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=Hk6kPgZA-.

[204] Sreejith Sreekumar, Zhengxin Zhang, and Ziv Goldfeld. "Non-asymptotic Performance Guarantees for Neural Estimation of f-Divergences." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3322–3330.

[205] Matthew Staib and Stefanie Jegelka. "Distributionally robust optimization and generalization in kernel methods." In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 9134–9144.

[206] Andrew M Steane. "Error correcting codes in quantum theory." In: *Physical Review Letters* 77.5 (1996), p. 793.

[207] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. "Certified defenses for data poisoning attacks." In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 3520–3532.

[208] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. "Evaluating model robustness and stability to dataset shift." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2611–2619.

[209] Mingjie Sun, Siddhant Agarwal, and J. Zico Kolter. "Poisoned classifiers are not only backdoored, they are fundamentally broken." In: *arXiv:2010.09080*. 2020.

[210] Yasunari Suzuki, Yoshiaki Kawase, Yuya Masumura, Yuria Hiraga, Masahiro Nakadai, Jiabao Chen, Ken M. Nakanishi, Kosuke Mitarai, Ryosuke Imai, Shiro Tamiya, et al. "Qulacs: a fast and versatile quantum circuit simulator for research purpose." In: *Quantum* 5 (Oct. 2021), p. 559. ISSN: 2521-327X. URL: https://doi.org/10.22331/q-2021-10-06-559.

[211] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." In: *2nd International Conference on Learning Representations*. 2014.

[212] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing Properties of Neural Networks." In: *2014 International Conference on Learning Representations (ICLR)*. OpenReview, 2014.

[213] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." In: *International Conference on Learning Representations*. 2014. URL: http://arxiv.org/abs/1312.6199.

[214] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[215] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. "Demon in the Variant: Statistical Analysis of {DNNs} for Robust Backdoor Contamination Detection." In: *30th USENIX Security Symposium*. 2021, pp. 1541–1558.

[216] Jiaye Teng, Guang-He Lee, and Yang Yuan. ℓ_1 *Adversarial Robustness Certificates: A Randomized Smoothing Approach*. 2020. URL: https://openreview.net/forum?id=H1lQIgrFDS.

[217] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. "DeepTest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars." In: *2018 International Conference on Software Engineering (ICSE)*. Association for Computing Machinery, 2018, pp. 303–314.

[218] Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. "Evaluating Robustness of Neural Networks with Mixed Integer Programming." In: *2018 International Conference on Learning Representations (ICLR)*. OpenReview, 2018.

[219] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, Ning Zhang, and Yevgeniy Vorobeychik. "Improving Robustness of ML Classifiers against Realizable Evasion Attacks Using Conserved Features." In: *2019 USENIX Conference on Security Symposium (USENIX Security)*. USA: USENIX Association, 2019, 285–302. ISBN: 9781939133069.

[220] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. "Llama: Open and efficient foundation language models." In: *arXiv preprint arXiv:2302.13971* (2023).

[221] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. "On Adaptive Attacks to Adversarial Example Defenses." In: *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 1633–1645.

[222] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. "Ensemble Adversarial Training: Attacks and Defenses." In: *2018 International Conference on Learning Representations (ICLR)*. OpenReview, 2018.

[223] Brandon Tran, Jerry Li, and Aleksander Madry. "Spectral signatures in backdoor attacks." In: *Advances in Neural Information Processing Systems*. 2018, pp. 8000–8010.

[224] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness May Be at Odds with Accuracy." In: *7th International Conference on Learning Representations*. 2019.

[225] Armin Uhlmann. "The transition probability in the state space of a $\star$-algebra." In: *Reports on Mathematical Physics* 9.2 (1976), pp. 273–279. URL: https://doi.org/10.1016/0034-4877(76)90060-4.

[226] Vladyslav Verteletskyi, Tzu-Ching Yen, and Artur F. Izmaylov. "Measurement optimization in the variational quantum eigensolver using a minimum clique cover." In: *J. Chem. Phys.* 152.12 (2020), p. 124114. DOI: 10.1063/1.5141458. URL: https://doi.org/10.1063/1.5141458.

[227] Aladin Virmaux and Kevin Scaman. "Lipschitz regularity of deep neural networks: analysis and efficient estimation." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

[228] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." In: *Nature Methods* 17 (2020), pp. 261–272. URL: https://doi.org/10.1038/s41592-019-0686-2.

[229] Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. "Towards robust CNN-based object detection through augmentation with synthetic rain variations." In: *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE. 2019, pp. 285–292.

[230] John Von Neumann and Herman H Goldstine. "Numerical inverting of matrices of high order." In: (1947).

[231] Jane Wakefield. "Microsoft chatbot is taught to swear on Twitter." In: *BBC News* (2016). URL: https://www.bbc.com/news/technology-35890188.

[232] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. "Poisoning Language Models During Instruction Tuning." In: *arXiv preprint arXiv:2305.00944* (2023).

[233] Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong, et al. "On Certifying Robustness against Backdoor Attacks via Randomized Smoothing." In: *CVPR 2020 Workshop on Adversarial Machine Learning in Computer Vision* (2020).

[234] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks." In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 707–723.

[235] Ligong Wang and Renato Renner. "One-Shot Classical-Quantum Capacity and Hypothesis Testing." In: *Phys. Rev. Lett.* 108 (20 2012), p. 200501.

[236] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J. Coles. "Noise-induced barren plateaus in variational quantum algorithms." In: *Nature Communications* 12.1 (2021), p. 6961. URL: https://doi.org/10.1038/s41467-021-27045-6.

[237] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. "Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification." In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 29909–29921.

[238] Siyue Wang, Xiao Wang, Shaokai Ye, Pu Zhao, and Xue Lin. "Defending dnn adversarial attacks with pruning and logits augmentation." In: *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2018, pp. 1144–1148.

[239] Yaxiao Wang, Yuanzhang Li, Quanxin Zhang, Jingjing Hu, and Xiaohui Kuang. "Evading PDF Malware Classifiers with Generative Adversarial Network." In: *2019 International Symposium on Cyberspace Safety and Security (CSS)*. Springer International Publishing, 2019, pp. 374–387.

[240] Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. "RAB: Provable Robustness Against Backdoor Attacks." In: Preprint at https://arxiv.org/abs/2003.08904 (2020).

[241] Dave Wecker, Matthew B Hastings, and Matthias Troyer. "Progress towards practical quantum variational algorithms." In: *Physical Review A* 92.4 (2015), p. 042303. URL: https://doi.org/10.1103/PhysRevA.92.042303.

[242] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. "Jailbroken: How does llm safety training fail?" In: *arXiv preprint arXiv:2307.02483* (2023).

[243] Frank Weinhold. "Lower bounds to expectation values." In: *Journal of Physics A: General Physics* 1.3 (1968), p. 305. URL: https://doi.org/10.1088/0305-4470/1/3/301.

[244] Frank Weinhold. "Criteria of Accuracy of Approximate Wavefunctions." In: *Journal of Mathematical Physics* 11.7 (1970), pp. 2127–2138. URL: https://doi.org/10.1063/1.1665372.

[245] D. H. Weinstein. "Modified Ritz Method." In: *Proceedings of the National Academy of Sciences* 20.9 (1934), pp. 529–532. URL: https://www.pnas.org/content/20/9/529.

[246] Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. "Towards Fast Computation of Certified Robustness for ReLU Networks." In: *2018 International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 5276–5285.

[247] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data." In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, pp. 4003–4012. URL: https://aclanthology.org/2020.lrec-1.494.

[248] Nathan Wiebe and Ram Shankar Siva Kumar. "Hardening quantum machine learning against adversaries." In: *New J. Phys.* 20.12 (2018), p. 123019.

[249] Wikipedia contributors. *SHA-2 — Wikipedia, The Free Encyclopedia*. [Online; accessed 18-March-2020]. 2020. URL: https://en.wikipedia.org/w/index.php?title=SHA-2\&oldid=944705336.

[250] Mark M. Wilde, Marco Tomamichel, Seth Lloyd, and Mario Berta. "Gaussian Hypothesis Testing and Quantum Illumination." In: *Phys. Rev. Lett.* 119 (2017), p. 120501.

[251] James H Wilkinson. "Modern error analysis." In: *SIAM review* 13.4 (1971), pp. 548–568.

[252] Eric Wong and Zico Kolter. "Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope." In: *2018 International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 5286–5295.

[253] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. "Scaling Provable Adversarial Defenses." In: *2018 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2018, pp. 8400–8409.

[254] Chaowei Xiao, Bo Li, Jun Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. "Generating Adversarial Examples with Adversarial Networks." In: *2018 International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 3905–3911.

[255] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. "Generating Adversarial Examples with Adversarial Networks." In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, 3905–3911.

[256] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. "Spatially Transformed Adversarial Examples." In: *2018 International Conference on Learning Representations (ICLR)*. OpenReview, 2018.

[257] Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. "Crfl: Certifiably robust federated learning against backdoor attacks." In: *International Conference on Machine Learning*. PMLR. 2021, pp. 11372–11382.

[258] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. "Empirical evaluation of rectified activations in convolutional network." In: *arXiv preprint arXiv:1505.00853* (2015).

[259] Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. "Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond." In: *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 1129–1141.

[260] Weilin Xu, David Evans, and Yanjun Qi. "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks." In: *25th Annual Network and Distributed System Security Symposium*. The Internet Society, 2018.

[261] Weilin Xu, Yanjun Qi, and David Evans. "Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers." In: *2016 Network and Distributed Systems Symposium (NDSS)*. Vol. 10. Internet Society, 2016.

[262] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. "Detecting ai trojans using meta neural analysis." In: *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2021, pp. 103–120.

[263] Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. "Generative poisoning attack method against neural networks." In: *arXiv:1703.01340* (2017).

[264] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. "Randomized Smoothing of All Shapes and Sizes." In: *2020 International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 10693–10705.

[265] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. "Randomized smoothing of all shapes and sizes." In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10693–10705.

[266] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. "Characterizing Audio Adversarial Examples Using Temporal Dependency." In: *International Conference on Learning Representations*. 2019.

[267] Zhuolin Yang, Zhikuan Zhao, Hengzhi Pei, Boxin Wang, Bojan Karlas, Ji Liu, Heng Guo, Bo Li, and Ce Zhang. "End-to-end Robustness for Sensing-Reasoning Machine Learning Pipelines." In: *arXiv:2003.00120* (2020).

[268] Zhuolin Yang, Zhikuan Zhao, Boxin Wang, Jiawei Zhang, Linyi Li, Hengzhi Pei, Bojan Karlaš, Ji Liu, Heng Guo, Ce Zhang, et al. "Improving certified robustness via statistical learning with logical reasoning." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 34859–34873.

[269] Tzu-Ching Yen, Vladyslav Verteletskyi, and Artur F. Izmaylov. "Measuring All Compatible Operators in One Series of Single-Qubit Measurements Using Unitary Transformations." In: *Journal of Chemical Theory and Computation* 16.4 (Apr. 2020), pp. 2400–2409. URL: https://doi.org/10.1021/acs.jctc.0c00008.

[270] Xiao Yuan, Suguru Endo, Qi Zhao, Ying Li, and Simon C. Benjamin. "Theory of variational quantum simulation." In: *Quantum* 3 (Oct. 2019), p. 191. URL: https://doi.org/10.22331/q-2019-10-07-191.

[271] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. "MACER: Attack-Free and Scalable Robust Training via Maximizing Certified Radius." In: *2020 International Conference on Learning Representations (ICLR)*. OpenReview, 2020.

[272] Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, and Qiang Liu. "Black-Box Certification with Randomized Smoothing: A Functional Optimization Based Framework." In: *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2020, pp. 2316–2326.

[273] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. "Dolphinattack: Inaudible voice commands." In: *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 2017, pp. 103–117.

[274] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. "Towards Stable and Efficient Training of Verifiably Robust Neural Networks." In: *2020 International Conference on Learning Representations (ICLR)*. OpenReview, 2020.

[275] Huan Zhang, Shiqi Wang, Kaidi Xu, Linyi Li, Bo Li, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. "General cutting planes for bound-propagation-based neural network verification." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1656–1670.

[276] Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. "BagFlip: A Certified Defense Against Data Poisoning." In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022. URL: https://openreview.net/forum?id=ZidkM5b92G.

[277] Zhikuan Zhao, Jack K Fitzsimons, Patrick Rebentrost, Vedran Dunjko, and Joseph F Fitzsimons. "Smooth input preparation for quantum and quantum-inspired machine learning." In: *Quantum Mach. Intell.* 3.1 (2021), pp. 1–6.

[278] Zhikuan Zhao, Alejandro Pozas-Kerstjens, Patrick Rebentrost, and Peter Wittek. "Bayesian deep learning on a quantum computer." In: *Quantum Mach. Intell.* 1.1 (2019), pp. 41–51.

[279] Haoti Zhong, Cong Liao, Anna Cinzia Squicciarini, Sencun Zhu, and David Miller. "Backdoor embedding in convolutional neural network models via invisible perturbation." In: *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*. 2020, pp. 97–108.

[280] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. "Transferable Clean-Label Poisoning Attacks on Deep Neural Nets." In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 7614–7623.

[281] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. "Universal and transferable adversarial attacks on aligned language models." In: *arXiv preprint arXiv:2307.15043* (2023).

[282] Christa Zoufal, David Sutter, and Stefan Woerner. "Error Bounds for Variational Quantum Time Evolution." In: *arXiv:2108.00022* (2021). URL: https://arxiv.org/abs/2108.00022.

[283] Abraham et al. *Qiskit: An Open-source Framework for Quantum Computing*. 2021. DOI: 10.5281/zenodo.2573505.