Constructing Optimal Noise Channels for Enhanced Robustness in Quantum Machine Learning

David Winderl†, Nicola Franco†, Jeanette Miriam Lorenz†
†Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany
{david.winderl, nicola.franco, jeanette.miriam.lorenz}@iks.fraunhofer.de

***Abstract*—With the rapid advancement of Quantum Machine Learning (QML), the critical need to enhance security measures against adversarial attacks and protect QML models becomes increasingly evident. In this work, we outline the connection between quantum noise channels and differential privacy (DP), by constructing a family of noise channels which are inherently ϵ-DP: (α, γ)-channels. Through this approach, we successfully replicate the ϵ-DP bounds observed for depolarizing and random rotation channels, thereby affirming the broad generality of our framework. Additionally, we use a semi-definite program to construct an optimally robust channel. In a small-scale experimental evaluation, we demonstrate the benefits of using our optimal noise channel over depolarizing noise, particularly in enhancing adversarial accuracy. Moreover, we assess how the variables α and γ affect the certifiable robustness and investigate how different encoding methods impact the classifier's robustness.**

***Index Terms*—Differential Privacy, Quantum Machine Learning, Adversarial Robustness, Quantum Computing.**

I. Introduction & Related Work

Quantum Machine Learning (QML) emerges as a prominent example of the potential applications for Noisy Intermediate-Scale Quantum (NISQ) devices [24]. Research into QML is motivated by the anticipation that it could significantly improve certain computational tasks, surpassing the performance of conventional algorithms [30, 25, 5, 2, 6]. Despite these potential benefits, QML faces its own set of challenges, especially concerning susceptibility to adversarial attacks [21, 28, 12]. Within classical machine learning, Differential Privacy (DP) has played a crucial role in striking a balance between data utility and the need for privacy protection [11, 1]. Specifically, DP has been utilized to enhance the reliability of model predictions for specific inputs [8, 20]. This principle extends naturally into Quantum Computing (QC) and introduces a novel approach to safeguard the integrity and privacy of data processed by QML models [33].

Quantum noise channels, such as depolarizing and phase damping noise in NISQ devices, are used as sources of stochastic noise. This noise helps achieve DP by exploiting the natural error processes of these devices [4, 15, 33, 27, 10]. With this argumentation, Weber et al. [27] have provided a relationship among quantum hypothesis testing and adversarial robustness. In addition, Hirche et al. [15] has provided a relationship between quantum DP and the quantum hockestick divergence. Angrisani et al. [4] is possibly one of the most noteworthy recent publications; they have provided a more general framework for the robustness upper bound of quantum noise channels by defining the neighbourhood concept in terms of the Schatten-norm and providing a more general upper bound for noise channels in terms of a depolarizing channel as well as a single qubit Pauli channel. This broader framework was not included in our work as we initially focused on defining such a family of noise channels. The linear nature of quantum channels and their relationship to semi-definite programming has been already considered, e.g. Guan et al. [14], to attempt to construct an optimal noise channel that achieves epsilon-DP using semi-definite programming. Nonetheless, the extent to which these techniques can defend classifiers against practical adversarial input manipulations remains uncertain.

The project/research is supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy with funds from the Hightech Agenda Bayern.

As indicated in our earlier work [29], the effectiveness of these approaches in providing robustness varies with the factor of depolarization noise. This raises the question which noise channel is the most suitable for a given classification problem. On a pathway to answer this question, we develop a family of noise channels, so called (α, γ)-channels, which describe noise channels naturally facilitating DP. In addition, we offer theoretical derivations for the bounds of *depolarizing noise* [33] and *random rotations* [16] from our general framework of (α, γ)-channels. This indicates that a broader set of channels can be described in our framework. Next, we add a construction of an *optimal* quantum channel as a positive semi-definite constrained optimization problem. This construction allows the experimental evaluation of a depolarizing channel against its optimally constructed counterpart, providing a more definite statement about the possible utility of quantum noise channels for DP and adversarial robustness against evasion attacks.

In summary, the contributions of our work are:

- The introduction of (α, γ)-channels as a family of noise channels which suffice ϵ-DP, along with formal bound derivations for depolarizing noise and random rotations.
- Construction of an optimal (α, γ)-channel trough a semi definite program. The maximum robustness ratio is obtained for a quantum classifier and a given sample.
- Experimental evaluation of depolarizing noise channels compared to optimal noise channels and evaluation between amplitude and angle encoding on three datasets.

II. Preliminaries

A. Quantum Adversarial Robustness

Quantum adversarial robustness examines how specially crafted attacks can weaken a quantum-based machine learning model, just as traditional neural networks can be fooled by adversarial inputs [13, 19, 22]. While there are other means of attacking a machine learning based system, we focus on evasion attacks on the QML model. In the context of classical machine learning, adversarial robustness refers to the ability of a machine learning to maintain a stable prediction when confronted with adversarial examples. Formally, let $\mathcal{Y}$ denote the set of classes and $\|\cdot\|$ denote the euclidean norm on $\mathbb{R}^n$. Let $f : \mathbb{R}^n \to [0,1]^{|\mathcal{Y}|}$ be a *classifier*, $x \in \mathbb{R}^n$ and $\epsilon > 0$.

Definition 1 (Adversarial Robustness)**.** *Consider a classifier f that maps an input x to an output y. The adversarial robustness at input x can be defined as the smallest amount of perturbation δ needed to change the model's prediction:*

$$\text{argmax}_{c\in\mathcal{Y}}\, f_c(x+\delta) \neq \text{argmax}_{c\in\mathcal{Y}}\, f_c(x),$$

where $\|\delta\| \leq \epsilon$ is bounded by a maximum ϵ budget and $x_{adv} = x + \delta$ denotes an adversarial example.

In the context of QC, a study by Lu et al. [21] found that Quantum Neural Networks (QNN) can fall prey to these targeted perturbations. As part of their work, the quantum version of the Fast Gradient Sign Method (FGSM) [13] was introduced. Thus, let us denote $x \in \mathbb{R}^n$ as (classical) features, $\mathcal{L}$ as loss function incorporating the QNN weights and ∇_x as gradient. In the context of FGSM a adversarial budget of ϵ is assumed, which then allows to obtain an adversarial example as follows:

$$x_{adv} = \Pi_{x+\mathcal{S}}(x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L})),$$

where $\Pi_{x+\mathcal{S}}(\cdot)$ clips the value towards the space $\mathcal{S}$ of adversarial perturbations around x. Typically, the space of adversarial perturbations is defined by a norm around the input x. The fact that QNNs are vulnerable to adversarial perturbations motivates the discussion about specific measures to counteract adversarial examples. In this work, we consider *robustness accuracy* as the network's ability to main stable predictions against adversarial attacks.

Definition 2 (Robustness Accuracy)**.** *Let $\mathcal{D} = \{x_i, y_i\}_{i=1,\dots,N}$ be a dataset and $a(\cdot, f, \epsilon) : \mathbb{R}^n \to \mathbb{R}^n$ be an adversarial model that tries to find an adversary within the perturbation budget ϵ. Then, the robustness accuracy of the classifier f on $\mathcal{D}$ under attack $a(\cdot, f, \epsilon)$ is defined as:*

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y_i = f(a(x_i, f, \epsilon))) \tag{1}$$

Where $\mathbb{1}$ denotes the indicator function which corresponds to 1 if the argument is true, else is returning 0.

B. Quantum DP

DP, with parameters (ϵ, δ), enhances data analysis and machine learning protection against adversaries. Lower ϵ values increase privacy by limiting what an algorithm reveals about individual data points, while δ represents the small chance that this privacy might fail. Together, (ϵ, δ)-DP makes it hard for adversaries to exploit data, safeguarding data integrity and privacy against sophisticated attacks. In the following, we focus entirely on ϵ-DP, which implies that $\delta = 0$ throughout this work. [1]
Formally for QC, two neighbouring Hilbert states ρ and σ are assumed, which are bound by a distance measure. Typically, the trace distance:

$$\tau(\sigma,\rho) = \frac{1}{2}\text{Tr}\,|\sigma-\rho| = \frac{1}{2}\text{Tr}\left[\sqrt{(\sigma-\rho)^\dagger(\sigma-\rho)}\right] \tag{2}$$

A quantum algorithm A (for instance, a classifier) is assumed to be ϵ-DP if, for a given set of outcomes S of $A(x)$, the following equation holds:

$$\Pr(A(\rho) \in S) \leq e^\epsilon \Pr(A(\sigma) \in S). \tag{3}$$

Since computations on quantum devices are dictated by a CPTP map (a quantum channel $\mathcal{E}$) and outcomes can be characterized by a set of Positive Operator-Valued Measurements (POVMs) $\{\Pi_k\}$[2], as alternative, one can abstract the effect of a quantum classifier as follows:

$$\Pr(A(\sigma) \in S) = y_k(\sigma) = \text{Tr}\left[\Pi_k \mathcal{E}(\sigma)\right]. \tag{4}$$

This reformulation leads to the definition of quantum ϵ-DP as:

$$y_k(\rho) \leq e^\epsilon y_k(\sigma), \quad \text{or} \quad \frac{y_k(\rho)}{y_k(\sigma)} \leq e^\epsilon. \tag{5}$$

Eq. 5 is of particular interest, since it indicates, that ϵ controls the change of labels for two neighbouring states.

C. Bound of ϵ-DP

Recent research on quantum DP incorporates a specific noise channel alongside the quantum channel that models the quantum classifier to establish concrete upper bounds for ϵ. Here, we aim to outline two particular bounds relevant to our study: (i) depolarizing noise and (ii) random rotations.

1) Bound on Depolarizing Noise: A depolarizing noise channel is characterized by the parameter p, defined as follows:

$$\mathcal{E}(\rho) = \frac{p}{N} I_N + (1-p)\rho. \tag{6}$$

Here $N = 2^n$ is the dimensionality of the density matrix and I_N the identity matrix of dimensionality N. Zhou and Ying [33], initially showed that a depolarizing channel suffices ϵ-DP. So given two input states ρ and σ, bound by a trace distance: $\tau(\sigma,\rho) \leq \tau_D$ and a quantum classification algorithm with a

[1] This is justified, since typically ϵ-DP is used for classification problems. Nevertheless as pointed out by Angrisani et al. [4] ϵ-DP cannot describe depolarizing noise entirely.

[2] Set of orthogonal projectors.

POVM, perturbed by a depolarizing noise channel, the output probabilities of ρ and σ are ϵ-DP:

$$\epsilon = \ln\left(\frac{D(1-p)\tau_D}{p} + 1\right), \tag{7}$$

where D is the dimension of the measurement operator used in the POVM. Du et al. [10] expanded on this concept by noting that the placement of the depolarizing channel within the quantum circuit does not affect the robustness guarantee. Additionally, they pointed out that ϵ-DP can be utilized to enhance adversarial robustness.

2) Bound on random rotations: Huang et al. [16] provided a bound for ϵ-DP, based on a process, where at first a layer of random rotations parametrized by θ_i, with: $h_1 < \tan(\theta_i) < h_2$ are placed, further they show that for *binary classification*, the following bound can be derived:

$$\epsilon = \ln\left[\frac{\tau_D}{t^n} + 1\right], \tag{8}$$

where $t = (h+1)\cdot 2^{(-n)}$ is the noise level. Their bound is of particular interest, since it is independent of the underlying device.

D. Choi representation of noise channels

As modern quantum devices are in the NISQ era [24], they are commonly affected by noise. This noise is often characterized using Kraus operators:

$$\mathcal{E}(\rho) = \sum_k K_k \rho K_k^\dagger \quad \text{where} \quad \sum_k K_k K_k^\dagger = I. \tag{9}$$

In constrast, we utilize the Choi–Jamiołkowski isomorphism [7]. This formalism is able to represent an n-qubit noise channel of dimensionality $d = 2^n$ as a complex, completely positive, and trace preserving matrix:

$$\mathcal{J}_\mathcal{E} = (\mathcal{E} \otimes I_A)(|\phi^+\rangle\langle\phi^+|) \in \mathbb{C}^{d^2 \times d^2}, \tag{10}$$

where $|\phi^+\rangle$ represents the fully entangled state and A the auxiliary system which we assume to always have the same dimensionality as our n-qubit system. On a further note, we refer to the Choi matrix that describes the identity channel as identity Choi matrix: $\mathcal{J}_\mathcal{I} = |\phi^+\rangle\langle\phi^+|$.

Next, it is necessary to ensure that $\mathcal{J}_\mathcal{E}$ is Completly Positive (CP) and Trace Preserving (TP) for our resulting quantum channel to be Completly Positive and Trace Preserving (CPTP). This can be achieved by asserting that the Choi matrix is positive semi-definite ($\mathcal{J}_\mathcal{E} \succeq 0$) and that tracing out the system describing the noise channel lead to an identity matrix of the size of the subsystem ($\mathrm{Tr}_2[\mathcal{J}_\mathcal{E}] = I_A$). Thus, we can construct the application of a noise channel as [3] [18]:

$$\mathcal{E}(\rho) = \mathrm{Tr}_1\left[(\rho \otimes I_A)\mathcal{J}_\mathcal{E}\right]. \tag{11}$$

Consequently, a POVM is constructed as follows [18]:

$$\tilde{y}_k(\rho) = \mathrm{Tr}\left[(\rho \otimes \Pi_k)\mathcal{J}_\mathcal{E}\right], \tag{12}$$

[3]See [32] for a visualization using tensor networks.

where Π_k is a measurement operator.

Further, a Kraus-channel can be reconstructed from a Choi matrix by looking at its spectral decomposition: $\mathcal{J}(\mathcal{E}) = \sum_i \lambda_i |\Psi_i\rangle\langle\Psi_i|$ [32]. One can then construct a matrix A_i, s.t $|A_j\rangle\rangle = \sum_{i,j} A_{ij}|i\rangle \otimes |j\rangle = |\Psi_i\rangle$. [4].
The individual Kraus operators can then be recovered as:

$$K_i = \sqrt{\lambda_i} A_i \tag{13}$$

III. Introduction of (α, γ)-Channels

In this section, we introduce a family of quantum noise channels, which allow us to derive ϵ-DP bounds and robustness guarantees. In reviewing the proofs concerning ϵ-DP conducted by Du et al. [10], Huang et al. [16] and Zhou and Ying [33], we observe that their bounds share similarities, particularly in their use of the property that trace preserving operations are *contractive* with respect to the trace distance, formally:

$$\tau(\mathcal{E}(\sigma), \mathcal{E}(\rho)) \leq \tau(\sigma, \rho). \tag{14}$$

In addition, they typically manage to establish bounds for the minimum probability of the noise channel, i.e:

$$y_k(\sigma) > 0. \tag{15}$$

Essentially, their proofs rely on the fact that quantum noise channels are contractive, and some may offer a lower bound on the smallest possible POVM. This reasoning inspires the definition of a family of noise channels that can be demonstrated to meet the requirements of ϵ-DP.

Building on this idea, we formalize the concept by introducing (α, γ)-channels:

Definition 3 ((α, γ)-channels)**.** *We define an (α, γ)-channel as an arbitrary quantum channel $\mathcal{E}_{(\alpha,\gamma)}(\rho) : \mathbb{C}^N \to \mathbb{C}^N$, with the following properties:*

1) $\forall \sigma, \rho. \quad \tau(\mathcal{E}_{(\alpha,\gamma)}(\rho), \mathcal{E}_{(\alpha,\gamma)}(\sigma)) \leq \alpha\tau(\sigma, \rho)$
2) $\forall \Pi_k, \sigma. \quad Tr\left[\Pi_k \mathcal{E}_{(\alpha,\gamma)}(\sigma)\right] \geq \gamma$

with $\alpha, \gamma \in [0; 1]$.

Intuitively, any (α, γ)-channel has a contraction factor of at most α and any input σ result in at most γ after a POVM.

A. ϵ-DP of (α, γ)-channels

As already hinted in our introduction, the definition is inspired by reviewing the process of thought used in Du et al. [10], Zhou and Ying [33] and Huang et al. [16]. Hence, we use the underlying assumptions we have provided for our (α, γ)-Channels to show ϵ-DP for any channels that follow Definition 3:

Theorem 1. *For all input states ρ and σ, which are bounded by a trace distance $\tau(\rho, \sigma) \leq \tau_D$, any (α, γ)-channel, $\mathcal{E}_{(\alpha,\gamma)}(\rho)$ suffices ϵ-DP, with:*

$$\epsilon = \ln\left[1 + \frac{\alpha\tau_D}{\gamma}\right]$$

[4]Since we are assuming the standard basis for the construction of superoperators, this can simply be done by reshaping the matrix.

Proof. We assume the noisy probabilities as $\tilde{y}_k(\sigma) = \text{Tr}\left[\Pi_k \tilde{\mathcal{E}}(\sigma)\right]$ and the clean as $y_k(\sigma) = \text{Tr}\left[\Pi_k \mathcal{E}(\sigma)\right]$. We can note that the absolute difference between the two probabilities is bound by their sensitivity (See Angrisani et al. [4, Section 8.1]):

$$\begin{aligned} |\tilde{y}_k(\sigma) - y_k(\rho)| = \\ |\text{Tr}\left[\Pi_k(\tilde{\mathcal{E}}(\sigma) - \tilde{\mathcal{E}}(\rho))\right]| \leq \\ \frac{1}{2}\|\Pi_k\|_\infty \|(\tilde{\mathcal{E}}(\sigma) - \tilde{\mathcal{E}}(\rho))\|_1 = \\ \|\Pi_k\|_\infty \tau(\tilde{\mathcal{E}}(\sigma), \tilde{\mathcal{E}}(\rho)) \overset{\text{POVM norm is one}}{=} \\ \tau(\tilde{\mathcal{E}}(\sigma), \tilde{\mathcal{E}}(\rho)) \overset{\text{Definition 3}}{\leq} \\ \alpha\tau(\sigma, \rho) \end{aligned}$$

Secondly, we can note that for any POVM Measurement:

$$y_k(\sigma) = \text{Tr}\left[\Pi_k \mathcal{E}_{(\alpha,\gamma)}(\sigma)\right] \geq \gamma \tag{16}$$

With those two relations, we can derive an upper bound the DP relation:

$$\begin{aligned} \frac{y_k(\rho)}{y_k(\sigma)} - 1 = \frac{y_k(\rho) - y_k(\sigma)}{y_k(\sigma)} \qquad \leq \frac{|y_k(\rho) - y_k(\sigma)|}{y_k(\sigma)} \\ \leq \frac{\alpha\tau(\sigma,\rho)}{\gamma} \leq \frac{\alpha\tau_D}{\gamma} \end{aligned}$$

Given now:

$$\epsilon = \ln\left[1 + \frac{\alpha\tau_D}{\gamma}\right] \tag{17}$$

We can resubstitute Eq. 17 in the last relation above, which equates to the same term:

$$e^\epsilon - 1 = e^{\ln\left[1+\frac{\alpha\tau_D}{\gamma}\right]} - 1 = \frac{\alpha\tau_D}{\gamma}$$

□

B. Relation of α, γ-channels to depolarizing channels

Now that we have introduced an (α, γ)-channel, it is necessary to consider whether the bound provided for ϵ-DP is tight. Specifically, given that the depolarizing channel is one of the most significant sources of noise, it is logical to demonstrate a bound that it incorporates the results of Du et al. [10]. In addition, to outline the extendibility of our framework, we exploit our concept of (α, γ)-channels to include the same bound as Huang et al. [16].

1) Depolarizing Noise: One can find that the contraction of a depolarizing noise channel with value p is defined as $\kappa = (1-p)$ Nielsen and Chuang [23, Chapter 9], from which we directly conclude that $\alpha = (1-p)$. To upper bound the smallest eigenvalue of a noise channel, we follow the line of argument of Du et al. [10]. Essentially they argue, that the smallest possible value of the noisy channel is given as $\text{Tr}\left[\Pi_k \mathcal{E}(\sigma)\right] = {}^{p}/_{D}$. Substituting this into our bound of ϵ-DP, it is apparent that we arrive at the same bound as in Eq. 7:

$$\epsilon = \ln\left[1 + \frac{(1-p)\tau_D}{{}^{p}/_{D}}\right] = \ln\left[1 + D\frac{(1-p)\tau_D}{p}\right] \tag{18}$$

2) Random Rotation Bound: Huang et al. [16] described their noise process as incorporating random rotations *before* the operation of the quantum classifier. Given that this process still constitutes a quantum channel, we can safely set the trivial upper bound for α at 1. To establish a bound on γ, we consider the scenario of a binary classifier [5]. This assumption allows us to assume that the majority class must be at least greater or equal than $\frac{(1+h)^n}{2}$, with h the lower bound on the random rotations in the R_x-Gates:

$$y_C(\sigma) > \frac{(1+h)^n}{2}. \tag{19}$$

Hence, we assume w.l.o.g.:

$$y_C = \frac{(1+h)^n}{2} \geq 0.5.$$

From this, we deduce the following for the subsequent class:

$$y_{k\neq C} = 1 - \frac{(1+h)^n}{2} \leq \frac{(1+h)^n}{2},$$

and for all labels $y_C, y_{k\neq C} \geq \frac{(1+h)^n}{2}$.

This provides us with $\gamma = \frac{(1+h)^n}{2}$ and by substituting $t = (h+1)\cdot 2^{1-n}$, we can find $\gamma = t^n$.
In the end, setting $\alpha = 1$ and $\gamma = t^n$ in Eq. 7 yield the same bound as Huang et al. [16]:

$$\epsilon = \ln\left(\frac{\tau_D}{t^n} + 1\right)$$

C. Certifiable robustness on (α, γ)-channels

This section builds on the robustness bound from Du et al. [10] and generalizes it to an (α, γ)-channel. We focus on the infinite sampling case for a hypothetical noise model that can only be simulated, not physically implemented. Refer to Du et al. [10] for details on the finite sampling case.

Theorem 2 (Infinite sampling case)**.** *Given a binary classifier, altered by a (α, γ)-channel: $y_k(\sigma) = Tr\left[\Pi_k \mathcal{E}_{\alpha,\gamma}(\sigma)\right]$. Suppose that for a predicted class C, the following holds:*

$$y_C(\sigma) > e^{2\epsilon} y_{k\neq C}(\sigma) \qquad \epsilon = \ln\left[1 + \frac{\alpha\tau_D}{\gamma}\right],$$

than for a benign state ρ with $\tau(\sigma,\rho) \leq \tau_D$, the predicted class label C, will not change:

$$\arg\max_k y_k(\sigma) = C = \arg\max_k y_k(\rho)$$

Proof. We can construct the proof similarly as in Du et al. [10], by simply replacing the definition of ϵ. We will first note that due to ϵ-DP:

$$y_C(\rho) \geq e^{-\epsilon} y_C(\sigma) \tag{20}$$

Hence using $y_C(\sigma) > e^{2\epsilon} y_{k\neq C}(\sigma)$:

$$y_C(\rho) \geq e^{-\epsilon} y_C(\sigma) > e^{\epsilon} y_{k\neq C}(\sigma) \tag{21}$$

Next, we can note that:

$$y_{k\neq C}(\rho) \leq e^{\epsilon} y_{k\neq C}(\sigma) \tag{22}$$

[5] A base assumption on the robustness in Huang et al. [16]

Which completes the relation:

$$y_C(\rho) \geq e^{-\epsilon} y_C(\sigma) > e^{\epsilon} y_{k\neq C}(\sigma) \leq y_{k\neq C}(\rho) \tag{23}$$

Since $y_C(\rho) > y_{k\neq C}(\rho)$, we can conclude that:

$$\arg\max_k y_k(\sigma) = C = \arg\max_k y_k(\rho) \tag{24}$$

□

Based on Theorem 2, if the ratio $\frac{\tilde{y}_C(\sigma)}{\tilde{y}_{C\neq k}(\sigma)}$ exceeds $e^{2\epsilon}$, we can certify the data point against any benign sample within radius τ_D. This leads to the following corollary:

Corollary 1. *Given a binary quantum classifier perturbed by a (α, γ)-channel, with:*

$$B \equiv \frac{\tilde{y}_C(\sigma)}{\tilde{y}_{C\neq k}(\sigma)}$$

This classifier is robust against any perturbation: $\sigma \to \rho$, with $\tau(\sigma, \rho) \leq \tau_D$, if:

$$\frac{\tilde{y}_C(\sigma)}{\tilde{y}_{C\neq k}(\sigma)} = B > e^{2\epsilon}$$

With: $\epsilon = \ln\left[1 + \frac{\alpha\tau_D}{\gamma}\right]$

Building on our prior analysis of depolarizing noise (III-B1), where we established a certifiable distance for a single data point based on network output probabilities, we can now determine the minimum value of ϵ that bounds the fraction B. This involves calculating:

$$\epsilon_{min} = \frac{1}{2} \ln\left({}^{y_C(\sigma)}/_{y_{C\neq k}(\sigma)}\right) \tag{25}$$

Next, one can simply use $\epsilon = \ln\left[1 + \frac{\alpha\tau_D}{\gamma}\right]$, which will lead to:

$$\tau_D = (e^{\epsilon_{min}} - 1) \cdot \frac{\gamma}{\alpha} \tag{26}$$

Note that for $\alpha = 0$, we would get a $\tau_D = \infty$, which highlights the limitations of our approach and the necessity of incorporating (ϵ, δ)-DP into the analysis of (α, γ)-channels.

IV. Construction of optimal (α, γ)-Channels

In this section, we investigate optimal noise channels for adversarial robustness. We formulate a semidefinite program (SDP) to find the best channel based on the certification distance from Corollary 1.

We consider a single data point (later expanded to a set) represented by a density matrix $\sigma = \sigma(x)$. A pretrained quantum classifier described by a unitary U acting on n-qubits and measured on d qubits by 2^d POVM-measurements: $\{\Pi_k\}$. We assume a true label C (w.l.o.g.). For brevity, we denote $\sigma_D = U\sigma U^\dagger$ and abbreviate this by σ. Refer to Eq. 4 for the POVM measurement definition: $y_k(\sigma) = \text{Tr}\left[\Pi_k \mathcal{E}_{(\alpha,\gamma)}(\sigma)\right]$, where $\mathcal{E}_{(\alpha,\gamma)}$ represents the noise channel with parameters α and γ. Given a threshold $B = \frac{y_C(\sigma)}{y_{k\neq C}(\sigma)} > e^{2\epsilon}$, we recognize that ϵ is fixed by the chosen noise channel. Hence, to maximize classifier robustness, a natural approach is to maximize the fraction B. More specifically we can formulate this as the following optimization problem:

$$\begin{aligned} \max \quad & {}^{y_C(\sigma)}/_{y_{k\neq C}(\sigma)}, \\ \text{s.t:} \quad & y_k(\sigma) = \text{Tr}\left[\Pi_k \mathcal{E}(\sigma)\right], \\ & \mathcal{E}(.) \text{ is a quantum channel}, \\ & \forall \sigma, \rho \quad \tau(\mathcal{E}(\sigma), \mathcal{E}(\rho)) \leq \alpha\tau(\sigma, \rho) + \beta, \\ & \forall k \quad y_k(\sigma) \geq \gamma. \end{aligned} \tag{27}$$

Unfortunately, the current problem formulation violates the rules of convexity and might not be directly solvable using SDPs. To address this, in the following subsections, we introduce reformulations to make it suitable for SDPs.

1) CPTP Properties: An underlying assumption of this work is that quantum channels can be described by Choi matrices, which must be CPTP. As pointed out earlier, we use the Choi–Jamiołkowski isomorphism [7], since kraus operators do not necessarily provide a convex or linear representation of our variables, which is a necessity in SDPs. We follow Knee et al. [17] and encode the two properties (CP and TP) individually as follows:

$$\mathcal{J}_\mathcal{E} \succeq 0 \tag{28}$$

$$\text{Tr}_{\text{out}}\left[\mathcal{J}_\mathcal{E}\right] = I_A \tag{29}$$

2) α-Bound for the SDP: To constrain the α parameter, we leverage Wolf [31, Theorem 8.17]. While the original source includes a disclaimer about potential inaccuracies, we provide a complete proof in Appendix A for clarity. Here, we directly utilize Wolf [31, Theorem 8.17] to bound the contraction $\alpha = (1 - \kappa)$ within our SDP formulation, as shown below:

$$\mathcal{E} - \kappa\mathcal{E}' \succeq 0. \tag{30}$$

Here $\mathcal{E}$ is our noise channel and $\mathcal{E}'$ describes a reference channel. Wolf [31] suggests to set: $\mathcal{E}' = {}^I/_d$ the fully mixed state as a reference, which will yield the following constraint on the Choi matrix of our noise channel:

$$\mathcal{J}_\mathcal{E} - \frac{\kappa}{d} I \succeq 0 \tag{31}$$

While the current constraint provides a valid upper bound, it has limitations. Given that the smallest eigenvalue of $\mathcal{J}_\mathcal{E}$ is zero, this bound offers no advantage over the trivial case $\kappa = 0 \implies (1 - \kappa) = 1$. Nonetheless, due to the lack of better methods for encoding the contraction factor within SDPs, we employ this approach, acknowledging that it restricts the representable noise channels.

A. γ-Bound for the SDP

To enforce the constraint on γ, we refrained from solving the SDP for all possible density matrices due to the clear in applicability of encoding a continuous set in a discrete setting. Henceforth, we achieve a similar behavior by encoding the following constraint in our PSD:

$$\forall \Pi_k \quad \text{Tr}_2\left[(\Pi_k \otimes I)\mathcal{J}_\mathcal{E}\right] - \gamma I \succeq 0 \tag{32}$$

Since any POVM measurement in the context of Choi matrices is defined as in Eq. 12, we can follow from the constraint, that: $\text{Tr}_2\left[(I \otimes \rho^T)\mathcal{J}_\mathcal{E}\right] \geq \gamma I$, which we can use to verify that all POVM measurements are bounded by γ:

$$\begin{aligned} y_k(\sigma) &= \text{Tr}\left[(\Pi_k \otimes \rho^T)\mathcal{J}_\mathcal{E}\right] \\ &= \text{Tr}\left[\Pi_k \text{Tr}_2\left[(I \otimes \sigma^T)\mathcal{J}_\mathcal{E}\right]\right] \\ &\geq \text{Tr}\left[\Pi_k \gamma I\right] \\ &= \gamma \end{aligned}$$

B. Combination of the SDP formulation

Given the constraints above, we can formulate an SDP to optimize for the best possible noise channel. Note that the function ${}^{y_C(\rho)}/_{y_{C\neq k}(\rho)}$ is non-convex and therefore not applicable in the context of SDP formulations. Hence, we simply maximize $y_C(\rho)$, since for the binary case $y_{C\neq k}(\rho) = 1 - y_C(\rho)$, hence the target optimization goal should still be met. With the constraints established, we can now formulate an SDP to find the optimal noise channel. However, the original objective function involving the ratio $\frac{y_C(\rho)}{y_{C\neq k}(\rho)}$ is non-convex and unsuitable for SDPs. To address this, we propose a simpler approach: maximize $y_C(\rho)$. Since in binary classification, the probabilities for the two classes sum to 1 ($y_{C\neq k}(\rho) = 1 - y_C(\rho)$), than, maximizing the probability $y_C(\rho)$ for the correct class inherently minimizes the probability for the incorrect class.

So far we have only assumed one data sample in our line of argument, since solving the SDP for every possible input will correspond to label leaking, we decided to construct the noise channel over the sum of all possible probabilities $y_{C_i}(\sigma_i)$ in the test set. We have chosen the test set to ensure that the noise channel is optimal for the samples attacked in our evaluation and no assumption about the generality have to be made. Next, we consider the class imbalance present on the test set by computing the weight vector $w = \frac{\sum_k C_k}{|C|}$, which leads to the overall optimization target:$\sum_i w_i \cdot y_{C_i}(\sigma_i)$.
During our evaluation, we found overall two scenarios that can be solved using our SDP: At first, one can apply the noise channel *after* the quantum classifier (post-order SDP) and *infront of* the quantum classifier (pre-order SDP). Both scenarious are outlined in the following.

1) Post order SDP: The post order SDP is rather a straight forward application of the previously discussed concepts.

$$\begin{aligned} \max \quad & \sum_i w_i \cdot y_{C_i}(\sigma_i) \\ \text{s.t:} \quad & \forall i \quad y_{C_i}(\sigma_i) = \text{Tr}\left[(U\sigma_k U^\dagger \otimes \Pi_{C_k})\mathcal{J}_\mathcal{E}\right] \\ & \mathcal{J}_\mathcal{E} \succeq 0 \\ & \text{Tr}_2\left[\mathcal{J}_\mathcal{E}\right] = I_A \\ & \forall \Pi_k \quad \text{Tr}_2\left[(\Pi_k \otimes I)\mathcal{J}_\mathcal{E}\right] - \gamma I \succeq 0 \\ & \mathcal{J}_\mathcal{E} - {}^{\alpha}/_{D} I \succeq 0 \end{aligned} \tag{33}$$

2) Pre-order SDP: With the pre-order SDP, we can note that the unitary will "act" on the Choi matrix upon measurements, which will alter our measurements in the following way:

$$y_k(\sigma) = \text{Tr}\left[(\sigma \otimes \Pi_k)(I \otimes U)\mathcal{J}_\mathcal{E}(I \otimes U^\dagger)\right] \tag{34}$$

Interestingly, this also implies for our γ constraint that:

$$\text{Tr}_2\left[(\Pi_k \otimes I)(I \otimes U)\mathcal{J}_\mathcal{E}(I \otimes U^\dagger)\right] - \gamma I \succeq 0 \tag{35}$$

Aside from this change, the SDP formulation will be unchanged leading us to the formulation:

$$\begin{aligned} \max \quad & \sum_i w_i \cdot y_{C_i}(\sigma_i) \\ \text{s.t:} \quad & \forall i \quad y_{C_i}(\sigma_i) = \text{Tr}\left[[(\sigma_i \otimes \Pi_{C_i})(I \otimes U)\mathcal{J}_\mathcal{E}(I \otimes U^\dagger)\right] \\ & \mathcal{J}_\mathcal{E} \succeq 0 \\ & \text{Tr}_2\left[\mathcal{J}_\mathcal{E}\right] = I_A \\ & \forall \Pi_k \quad \text{Tr}_2\left[(\Pi_k \otimes I)(I \otimes U)\mathcal{J}_\mathcal{E}(I \otimes U^\dagger)\right] - \gamma I \succeq 0 \\ & \mathcal{J}_\mathcal{E} - {}^{\alpha}/_{D} I \succeq 0 \end{aligned} \tag{36}$$

V. Experiments

In our experiments, we wanted to highlight the utility and limitations of utilizing a SDP Formulation as a bechmark for optimally robust quantum channels. See Appendix B for a detailed description of our data preprocessing and SDP implementation steps. We have considered the Iris dataset (**Iris**) with *Angle* and *Amplitude* embedding, the Pima Indians Diabetes dataset (**PID**) and the breast cancer dataset (**BC**). The latter two where only utilized using Amplitude embedding.

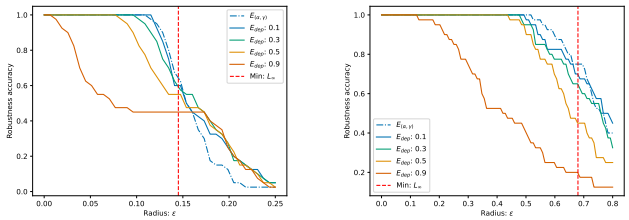
In our experiments, we want to highlight the performance of depolarizing noise against our optimal (α, γ) counter part. Further we want to analyze the effect of α and γ on the certifiablility and at last we want to compare the effectiveness of different encodings.

A. Classification of Depolarizing Noise against optimal noise

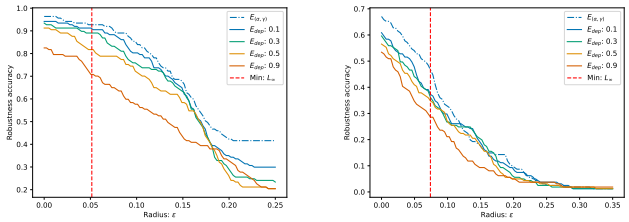
Providing an optimal noise channel for various values of depolarizing noise serves as an experimental benchmark for the optimality of specific noise channels. Here, we outline the process. As a first step, one needs to reduce α and γ to the parameters of the specific noise channel. Next, one can iterate over a set of parameters and compare the adversarial accuracy to that of the optimal noise channel, as illustrated in Fig. 1. For depolarizing noise, we observe that in most datasets, a small amount of depolarizing noise provides a good upper bound, but typically the effect decays with an increase in depolarizing noise.

B. Influence on α and γ on the certifiable robustness

Next, we want to outline the effect of α and γ on the certifiable accuracy. Therefore, we consider Eq. 26 to compute the certifiable distance for each datasample in terms of the trace distance and then compute the portion of certifiable samples on the test set given a set of distances $\tau_D \in \{0.05, 0.10, 0.15\}$. We first note that the effect of α and γ on the difference

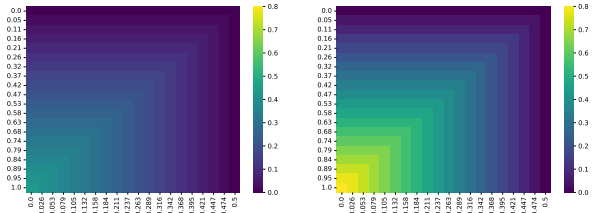


(a) Iris (Amplitude) (b) Iris (Angle)

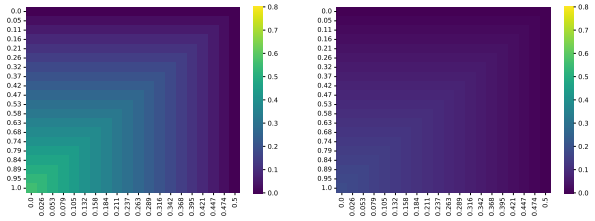


(c) BC (Amplitude) (d) PID (Amplitude)

Fig. 1: Adversarial accuracy given a FGSM attack for the various datasets used in this study. The dashed line indicates the adversarial accuracy's of the best case noise channel's. The red dashed vertical line corresponds to the smallest L_∞ distance between two data samples of the opposite class. Note that we used a single (α, γ)-channel, since they showed the tendency to overlap (See section C).



(a) Iris (Amplitude) (b) Iris (Angle)



(c) BC (Amplitude) (d) PID (Amplitude)

Fig. 2: Difference $y_C - y_{k\neq C}$ for the various different datasets used in this study.

between the class label and the follow up class (which is later on described in the fraction B in Corollary 1) scale symetrically with α and γ. See Fig. 2 for a visualization. Additionally one can observe that the difference always converge to zero for all datasets given: $\alpha = 0$ and $\gamma = 0.5$. For these values the quantum classifer is perturbed by a noise channel describing random guessing, which yields naturally a certifiable distance of $\tau_D = \infty$. Given the limited usefulness of a random guessing network, we decide to constrain the certifiable distance shown by the plots in Fig. 4 by a δ on the difference among y labels. To this end, we choose a δ of 0.05 for the Iris dataset (Angle and Amplitude Embedding) and the BC dataset. For the PID dataset, we pick a δ of 0.01, since the class probabilities are closed together due to the lower accuracy.

Concerning the certifiable samples in Fig. 4, one can see that the influence of α and γ have an effect on the amount of certifiable distances given by the classifier. The area where the maximum of samples can be certified clearly shrinks in case the cut-off distance is increased from 0.05 to 0.15 and typically converges to a line like structure which is the maximum for $(\alpha = 0.11, \gamma = 0.447)$ (PID, Iris Angle and BC), $(\alpha = 0.16, \gamma = 0.421)$ (Iris Amplitude). Which additionally describes the smallest allowed difference considered in our experiments.

Interestingly, across various datasets, lower values of γ tend to extend the range of certifiable samples. These samples maintain their high certification level for a longer period before experiencing a characteristic sharp drop. This drop point depends on the chosen certifiable distance, the specific dataset, and the embedding used. In contrast, the parameter α seems to induce a smoother transition in the certification level. It remains an open question whether this observed behavior is a consequence of the SDP formulation itself or an inherent property of (α, γ)-channels.

C. Comparison of amplitude and angle embedding

In Definition 3, it is apparent, that the ϵ of the ϵ-DP mechanism is proportional to α, β and the trace distance of the two input states, which highlights that different encodings, might additionally provide further robustness bounds on quantum classifiers. To provide further insights, we have normalized the ϵ values used in this study from the respective intervals of $[0; 0.25]$ (Amplitude embedding) and $[0; 0.9]$ (Angle Embedding) of the Iris dataset to the interval $[0; 1]$. The results can be observed in Fig. 3. Overall it is apparent, that a better encoding shows a higher degree of adversarial accuracy, which decays at a similar rate with an increase in depolarizing noise and is still bounded by the optimally constructed (α, γ)-Channel.

VI. Conclusion

In this work, we have defined a family of noise channels, which are certifiable robust. We have hereby related robustness of QNNs to the contraction of the algorithm, to the smallest possible label assignment as well as to the encoding strategy used by the quantum classifier. Our SDP formulation, while theoretically useful for constructing optimal noise channels (maximizing the fraction B), suffers from numerical scaling limitations in practical applications. Nevertheless, the study reveals several interesting implications and future research directions.

Firstly, the framework could be extended to evaluate other noise sources beyond depolarizing noise, potentially leading

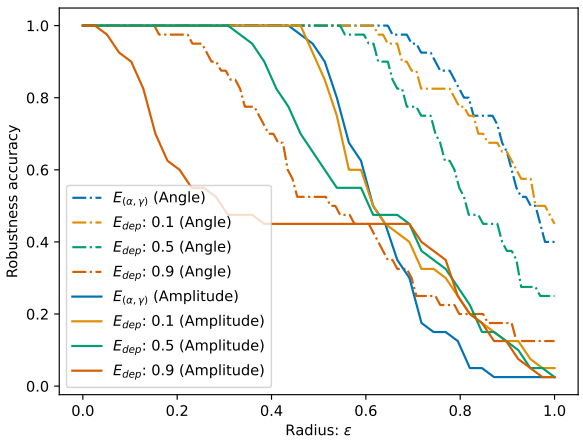


Fig. 3: Normalized adversarial accuracy's between amplitude and angle encoding. The adversarial attacks where done using FGSM with a L_∞-norm.

to a broader understanding of quantum noise and its impact on ϵ-DP.

Secondly, while a small amount of depolarizing noise shows promise as a defense against evasion attacks, it remains unclear if noise inherent to NISQ devices can truly provide robustness. Our findings, along with previous work, suggest that increasing depolarization leads to decreased certifiability.

Thirdly, the concept of noise channels as a general model for quantum computation, coupled with the existence of certifiable robust channels, opens exciting avenues. Constructing a variational ansatz with inherently certifiable behavior would be a significant advance. Such an ansatz could be easier to optimize compared to finding the *optimal* noise model (a convex optimization problem), despite the computational challenges arising from exponentially growing Hilbert spaces. This could pave the way for theoretically grounded adversarial robustness in quantum classifiers. Additionally, the observed improvement in robustness due to encoding suggests that quantum classifiers might be inherently robust when data point separation is effective. This combination of factors could lead to a paradigm shift, where quantum classifiers are recognized for their inherent robustness that scales with data separability.

Appendix A
Proof of Wolf [31, Theorem 8.17]

In the following Wolf [31, Theorem 8.17] is outlined combined with its proof and the accompanying references.

Theorem 3. *Given $\mathcal{E}, \mathcal{E}'$ be two trace preserving and Hermitian preserving linear maps, where we define $\mathcal{E}'$ as:*

$$\mathcal{E}'(\rho) = Tr[\rho]\, Y; \quad Tr[Y] = 1$$

If $\mathcal{E} - \kappa\mathcal{E}'$ is positive for some $\kappa \geq 0$, then for all density matrices ρ, σ:

$$Tr|\mathcal{E}(\sigma - \rho)| \leq (1-\kappa) Tr|\sigma - \rho|$$

Proof. We start off by noting that [6] $\mathcal{E}'(\sigma - \rho) = \mathrm{Tr}[\sigma - \rho]\, Y = 0$. We can hence reformulate the trace distance as follows:

$$\begin{aligned}
\frac{1}{2}\mathrm{Tr}|\mathcal{E}(\sigma-\rho)| = \\
\sup_{0\leq P\leq I}\{\mathrm{Tr}[\mathcal{E}(\sigma-\rho)P]\} = \\
\sup_{0\leq P\leq I}\{\mathrm{Tr}[(\mathcal{E}(\sigma-\rho) - \kappa\mathcal{E}'(\sigma-\rho))P]\} = \\
\sup_{0\leq P\leq I}\{\mathrm{Tr}[((\mathcal{E}-\kappa\mathcal{E}')(\sigma-\rho))P]\}
\end{aligned}$$

Next, we will define the dual map, such that:

$$\mathrm{Tr}[((\mathcal{E}-\kappa\mathcal{E}')(\sigma-\rho))P] = \mathrm{Tr}[((\mathcal{E}-\kappa\mathcal{E}')^*(P))(\sigma-\rho)] \tag{37}$$

Since $\mathcal{E} - \kappa\mathcal{E}'$ is defined as positive, we can conclude that $(\mathcal{E}-\kappa\mathcal{E}')^*$ is positive too [26, Proposition 2.18]. we can follow that since P is positive and bounded by I, that $P' = (\mathcal{E} - \kappa\mathcal{E}')^*(P)$ is positive. Since $\mathcal{E}$ and $\mathcal{E}'$ are trace preserving, the application of identity to their dual channel scales I as: $(\mathcal{E}-\kappa\mathcal{E}')^*(I) = (1-\kappa)I$. Given now that by definition $P \leq I$, we can follow that

$$(\mathcal{E}-\kappa\mathcal{E}')^*(I) = (1-\kappa)I \geq P' = (\mathcal{E}-\kappa\mathcal{E}')^*(P) \tag{38}$$

We can follow, that the resulting projection is bounded by: $0 \leq P' \leq (1-\kappa)I$, which allows the following relation:

$$\begin{aligned}
\sup_{0\leq P\leq I}\{\mathrm{Tr}[((\mathcal{E}-\kappa\mathcal{E}')(\sigma-\rho))P]\} = \\
\sup_{0\leq P\leq I}\{\mathrm{Tr}[((\mathcal{E}-\kappa\mathcal{E}')^*(P))(\sigma-\rho)]\} \leq \\
\sup_{0\leq P'\leq (1-\kappa)I}\{\mathrm{Tr}[(\sigma-\rho)P']\} = \\
(1-\kappa)\sup_{0\leq P'\leq I}\{\mathrm{Tr}[(\sigma-\rho)P']\} = \\
\frac{1-\kappa}{2}\mathrm{Tr}|\sigma-\rho|
\end{aligned}$$

From which we can follow, that:

$$\frac{1}{2}\mathrm{Tr}|\mathcal{E}(\sigma-\rho)| \leq \frac{1-\kappa}{2}\mathrm{Tr}|\sigma-\rho| \tag{39}$$

□

Appendix B
Experimental Setup

A. Training of Neural networks

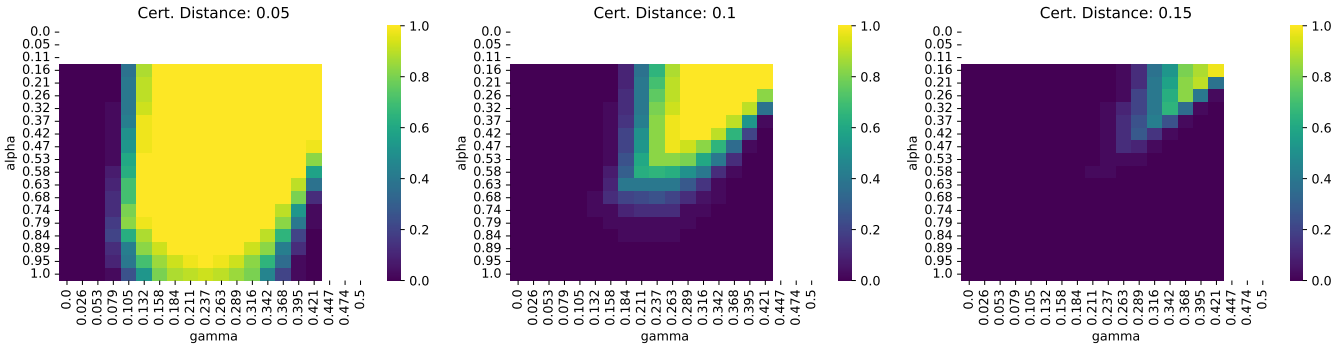
We trained four different QNNs, using a similar scheme as in our previous study. Overall we used the Pima Indians Diabetes Dataset, the Breast Cancer and the Iris dataset. We used CategoricalCrossEntropy [7] as a loss function for all datasets and considered the class imblanance by adding weights to the loss. As a optimizer we used Adam.

B. Data preprocessing

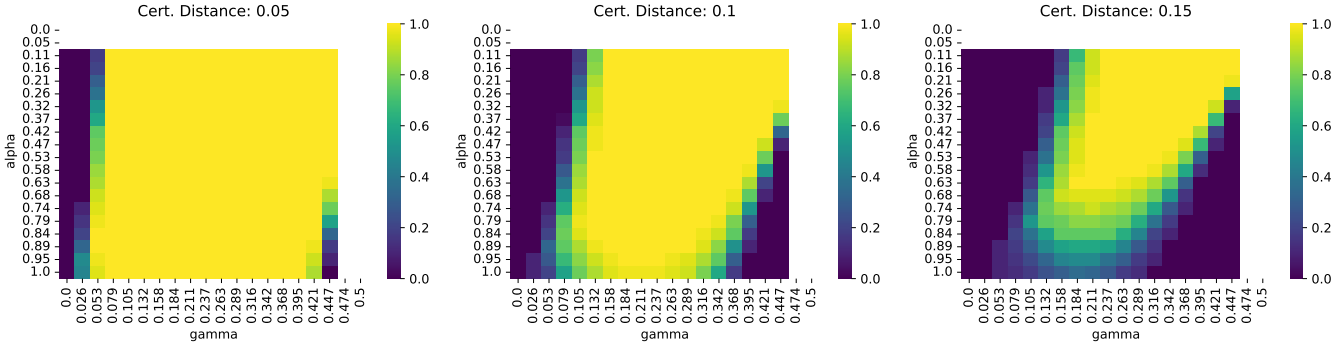
As already pointed out we used overall three datasets `Iris`, `Pima Indians Diabetes` and `Breast Cancer`.

[6] Since $\mathrm{Tr}[\sigma-\rho] = \mathrm{Tr}[\sigma] - \mathrm{Tr}[\rho] = 1 - 1 = 0$

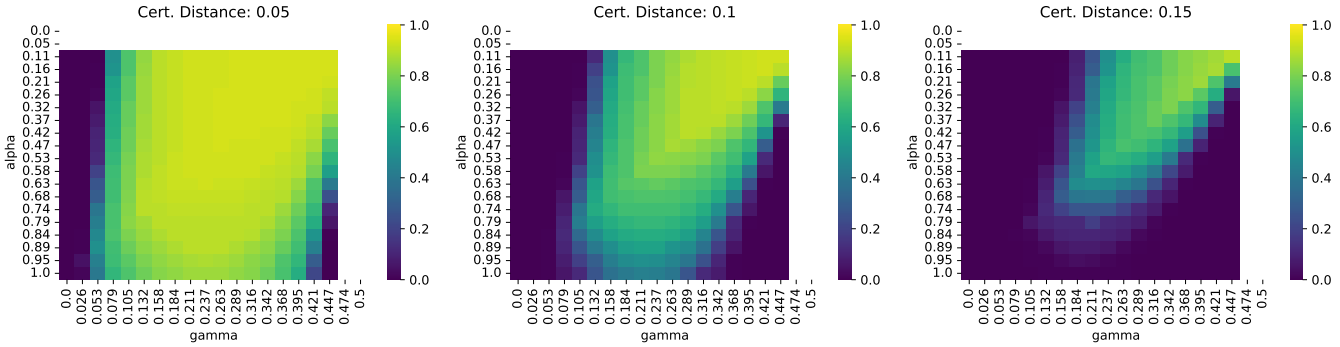
[7] The standardxx



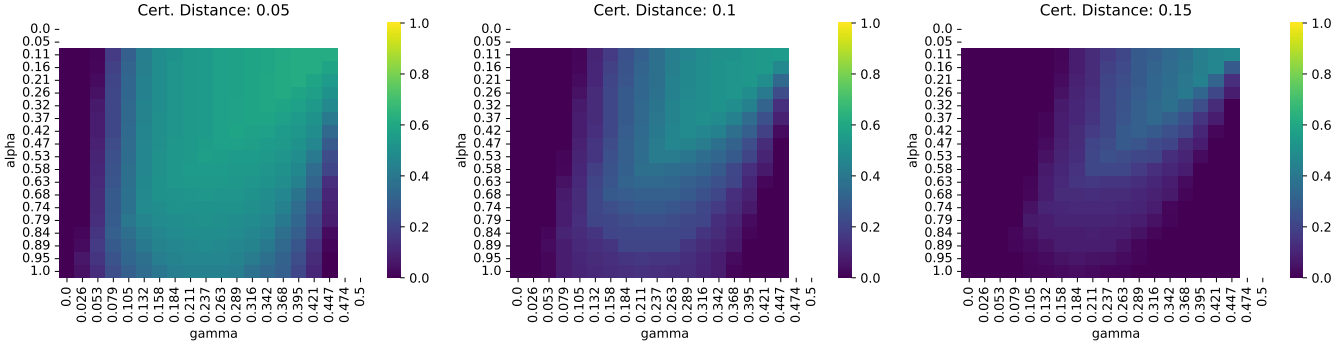
(a) Iris (Amplitude)



(b) Iris (Angle)



(c) BC (Amplitude)



(d) PID (Amplitude)

Fig. 4: Portion of certifieable distances among the test set for different values of α and γ. The portion is computed by grouping among the levels: $\{0.05, 0.10, 0.15\}$.

1) Iris: For Iris we followed the approach by Du et al. [10], dropping the `petal_width` feature as well as the `versicolor` class. We then normalized each datapoint by its respective $||.||_2$-Norm and performed a train/test split of $40:60$. Regarding the Angle Embedding, the datapoints where min max normalized to the interval $[0;\pi]$.

2) Pima Indians Diabetes (PID): For the Pima Indians Diabetes dataset, we first dropped all nan or duplicate features from the dataset, next to overcome the present class imbalance, we removed overall 232 datapoints from the majority class at random. We used the remaining dataset with 536 datapoints and split them into a train set consisting of 375 datapoints and a test set of 161 datapoints. To provide a comparabilty with the Iris dataset, we additionally normalized each datapoint with it's respective $||.||_2$-Norm.

3) Breast Cancer Wisconsin (BC): For the Breast Cancer Wisconsin dataset we found overall 683 features, with a dimension of 9, since this would require four qubits in the angle embedding, we performed a PCA with eight components, effectively rescaling the dimensionality to eight. We rescaled the data into the interval $(0;1]$ and normalized each feature vector by it's respective $||.||_2$-Norm. No further preprocessing was conducted.

C. QNN Architecture

As pointed out earlier, we used a similar scheme of dataembedding plus strongly entangling layers. For embedding we trained Iris once using angle embedding and once amplitude embedding. Since we where only able to solve up to 3 qubits, we refrained from using qngle embedding for the PID and BC datasets and applied hereby amplitude embedding. For an overview of the various hyperparameters used in this study, refer to Table I.

TABLE I: Hyperparameters used for the various types of QNNs for the datasets Iris, Pima Indians Diabetes (PID) and Breast Cancer (BC).

Datset	Embedding	Qubits	Layers	Batch	Learn. Rate	Epochs
Iris	Amplitude	2	2	30	0.05	100
Iris	Angle	3	2	30	0.01	100
BC	Amplitude	3	40	16	0.0005	10
PID	Amplitude	3	16	16	0.005	10

At last we want to point out how the depolarizing mechanism was constructed in our study

D. Implementation of the SDP

1) Complexity analysis: In this study, we use MOSEK in combination with CVXpy [9] with an interior-point method, scaling its runtime with $\mathcal{O}(v^{3.5})$ [3] to solve the SDP, where v represents the number of variables in the SDP. In our work, we embed a Choi matrix of size $d^2 \times d^2$, where $d = 2^n$ describes the dimensionality of the Hilbert space. This indicates that the number of variables scales biquadratically with the dimensionality of the Hilbert space of the used quantum classifier, and hence exponentially in terms of the number of qubits:

$$v = d^4 = (2^n)^4 \tag{40}$$

With d the dimensionality of the hilberspace and n the number of qubits. The biquadratic scaling also indicates why, in contrast to other methods like Guan et al. [14], we can only exhibit our study on maximally three qubits.

TABLE II: Adversarial Accuracies for Iris Amplitude, among various values of α and γ parameterized by different values of depolarizing noise. Produced by the optimal noise channel

		Adversarial Budget (ϵ)									
α	γ	0.0	0.03	0.05	0.08	0.1	0.13	0.15	0.18	0.21	0.23
1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.9	0.5	0.17	0.05	0.03
0.9	0.05	1.0	1.0	1.0	1.0	1.0	0.9	0.5	0.17	0.05	0.03
0.7	0.15	1.0	1.0	1.0	1.0	1.0	0.9	0.5	0.17	0.05	0.03
0.5	0.25	1.0	1.0	1.0	1.0	1.0	0.9	0.5	0.17	0.05	0.03
0.1	0.45	1.0	1.0	1.0	1.0	1.0	0.9	0.5	0.17	0.05	0.03
0.05	0.48	1.0	1.0	1.0	1.0	1.0	0.9	0.5	0.17	0.05	0.03
0.0	0.5	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43

2) Numerical Considerations: As already pointed out, we used CVXpy [9] to implement the SDP in Eq. 36 combined with MOSEK as a solver. During optimization, we resolved multiple numerical issues. Initially, we approximated the constraint: $\mathrm{Tr}_2[\mathcal{J}] = I_A$, by replacing it with: $\|\mathrm{Tr}_2[\mathcal{J}] - I_A\|_{\mathrm{F}} \leq \delta$, setting $\delta = 1 \times 10^{-10}$. Secondly, instead of explicitly stating the constraint that $\mathcal{J} \succeq 0$, we asserted on the variable itself that $\mathcal{J}$ is PSD, which typically guided the optimizer to feasible solutions. Lastly, we emphasize that our optimal noise channel was constructed on the test set, on which we further applied the attack. The main reasoning behind this was to provide the actual best possible noise channel when averaging over the test set.

APPENDIX C
OVERLAPPING ROBUSTNESS IN OPTIMAL NOISE CHANNELS

During our work, we found that typically the optimally constructed noise channels coincide in terms of their adversarial accuracy. This provides a challenging perspective as α and γ seemingly control the contractiveness and the minimal possible measurement, respectively. Nevertheless, we want to emphasize that α provides an upper bound on the contractiveness of the noise channel by Definition 3, and γ *a lower bound* on the minimal measurements. Those bounds suffice to control ϵ-DP. Nevertheless, the optimizer might choose a different specific values describing the contraction and smallest possible measurement. We hence saw as outlined for Iris Amplitude in Table II, that the adversarial accuracies typically where overlapping. We hence simply used one value for the effect of α and γ in the following when describing adversarial accuracy's.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

[3] Erling D. Andersen. The homogeneous and self-dual model and algorithm for linear optimization. mosek technical report: Tr-1-2009. 2009. URL https://api.semanticscholar.org/CorpusID:16771233.

[4] Armando Angrisani, Mina Doosti, and Elham Kashefi. A unifying framework for differentially private quantum algorithms. *arXiv preprint arXiv:2307.04733*, 2023.

[5] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

[6] M Cerezo, Guillaume Verdon, Hsin-Yuan Huang, Lukasz Cincio, and Patrick J Coles. Challenges and opportunities in quantum machine learning. *Nature Computational Science*, 2(9):567–576, 2022.

[7] Man-Duen Choi. Completely positive linear maps on complex matrices. *Linear Algebra and its Applications*, 10(3):285–290, 1975. ISSN 0024-3795. doi:https://doi.org/10.1016/0024-3795(75)90075-0. URL https://www.sciencedirect.com/science/article/pii/0024379575900750.

[8] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[9] Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17:1–5, 2016. URL http://www.cvxpy.org/.

[10] Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Dacheng Tao, and Nana Liu. Quantum noise protects quantum classifiers against adversaries. *PHYSICAL REVIEW RESEARCH*, 3: 23153, 2021. doi:10.1103/PhysRevResearch.3.023153.

[11] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[12] Nicola Franco, Alona Sakhnenko, Leon Stolpmann, Daniel Thuerck, Fabian Petsch, Annika Rüll, and Jeanette Miriam Lorenz. Predominant aspects on security for quantum machine learning: Literature review. *arXiv preprint arXiv:2401.07774*, 2024.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[14] Ji Guan, Wang Fang, and Mingsheng Ying. *Robustness Verification of Quantum Classifiers*, page 151–174. Springer International Publishing, 2021. ISBN 9783030816858. doi:10.1007/978-3-030-81685-8_7. URL http://dx.doi.org/10.1007/978-3-030-81685-8_7.

[15] Christoph Hirche, Cambyse Rouzé, and Daniel Stilck França. Quantum differential privacy: An information theory perspective. *IEEE Transactions on Information Theory*, 2023.

[16] Jhih-Cing Huang, Yu-Lin Tsai, Chao-Han Huck Yang, Cheng-Fang Su, Chia-Mu Yu, Pin-Yu Chen, and Sy-Yen Kuo. Certified robustness of quantum classifiers against adversarial examples through quantum noise. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[17] George C. Knee, Eliot Bolduc, Jonathan Leach, and Erik M. Gauger. Quantum process tomography via completely positive and trace-preserving projection. *Physical Review A*, 98(6), December 2018. ISSN 2469-9934. doi:10.1103/physreva.98.062336. URL http://dx.doi.org/10.1103/PhysRevA.98.062336.

[18] George C Knee, Eliot Bolduc, Jonathan Leach, and Erik M Gauger. Quantum process tomography via completely positive and trace-preserving projection. *Physical Review A*, 98(6):062336, 2018.

[19] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2017.

[20] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pages 656–672. IEEE, 2019.

[21] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. Quantum adversarial machine learning. *Physical Review Research*, 2(3):033212, 2020.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[23] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2011. ISBN 9781107002173.

[24] John Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, 2:79, August 2018. ISSN 2521-327X. doi:10.22331/q-2018-08-06-79. URL https://doi.org/10.22331/q-2018-08-06-79.

[25] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015.

[26] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018.

[27] Maurice Weber, Nana Liu, Bo Li, Ce Zhang, and Zhikuan Zhao. Optimal provable robustness of quantum classification via quantum hypothesis testing. *npj Quantum Information 2021 7:1*, 7:1–12, 5 2021. ISSN 2056-6387. doi:10.1038/s41534-021-00410-5. URL https:

//www.nature.com/articles/s41534-021-00410-5.

[28] Maxwell T West, Sarah M Erfani, Christopher Leckie, Martin Sevior, Lloyd CL Hollenberg, and Muhammad Usman. Benchmarking adversarially robust quantum machine learning at scale. *Physical Review Research*, 5(2):023186, 2023.

[29] David Winderl, Nicola Franco, and Jeanette Miriam Lorenz. Quantum neural networks under depolarization noise: Exploring white-box attacks and defenses. *arXiv preprint arXiv:2311.17458*, 2023.

[30] Peter Wittek. *Quantum machine learning: what quantum computing means to data mining*. Academic Press, 2014.

[31] Michael M Wolf. Quantum channels and operations-guided tour. 2012.

[32] Christopher J Wood, Jacob D Biamonte, and David G Cory. Tensor networks and graphical calculus for open quantum systems. *Quantum Information & Computation*, 15(9-10):759–811, 2015.

[33] Li Zhou and Mingsheng Ying. Differential privacy in quantum computation. *Proceedings - IEEE Computer Security Foundations Symposium*, pages 249–262, 9 2017. ISSN 19401434. doi:10.1109/CSF.2017.23.