npj | Quantum Information

www.nature.com/npjqi

ARTICLE OPEN

Check for updates

Optimal provable robustness of quantum classification via quantum hypothesis testing

Maurice Weber[1], Nana Liu[2,3,4], Bo Li[5], Ce Zhang[1]✉ and Zhikuan Zhao[1]✉

Quantum machine learning models have the potential to offer speedups and better predictive accuracy compared to their classical counterparts. However, these quantum algorithms, like their classical counterparts, have been shown to also be vulnerable to input perturbations, in particular for classification problems. These can arise either from noisy implementations or, as a worst-case type of noise, adversarial attacks. In order to develop defense mechanisms and to better understand the reliability of these algorithms, it is crucial to understand their robustness properties in the presence of natural noise sources or adversarial manipulation. From the observation that measurements involved in quantum classification algorithms are naturally probabilistic, we uncover and formalize a fundamental link between binary quantum hypothesis testing and provably robust quantum classification. This link leads to a tight robustness condition that puts constraints on the amount of noise a classifier can tolerate, independent of whether the noise source is natural or adversarial. Based on this result, we develop practical protocols to optimally certify robustness. Finally, since this is a robustness condition against worst-case types of noise, our result naturally extends to scenarios where the noise source is known. Thus, we also provide a framework to study the reliability of quantum classification protocols beyond the adversarial, worst-case noise scenarios.

npj Quantum Information (2021)7:76 ; https://doi.org/10.1038/s41534-021-00410-5

INTRODUCTION

The flourishing interplay between quantum computation and machine learning has inspired a wealth of algorithmic invention in recent years[1–3]. Among the most promising proposals are quantum classification algorithms that aspire to leverage the exponentially large Hilbert space uniquely accessible to quantum algorithms to either drastically speed up computational bottlenecks in classical protocols[4–7], or to construct quantum-enhanced kernels that are practically prohibitive to compute classically[8–10]. Although these quantum classifiers are recognized 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[11–14]. These perturbations can occur either due to imperfect implementation that is prevalent in the noisy, intermediate-scale quantum (NISQ) era[15], or, more menacingly, due to adversarial attacks where a malicious party aims to fool a classifier by carefully crafting practically undetectable noise patterns that trick a model into misclassifying a given input.

In order to address these short-comings in reliability and security of quantum machine learning, several protocols in the setting of adversarial quantum learning, i.e., learning under the worst-case noise scenario, have been developed[11,12,16–18]. More recently, data encoding schemes are linked to robustness properties of classifiers with respect to different noise models in ref. [19]. The connection between provable robustness and quantum differential privacy is investigated in ref. [17], where naturally occurring noise in quantum systems is leveraged to increase robustness against adversaries. A further step toward robustness guarantees is made in ref. [18] 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 quantum machine learning algorithms against adversarial 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 that 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 that is both sufficient and necessary. Here we set out to confront both aspects of this open problem by generalizing 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[13,14,20]. Many heuristic defense strategies have been proposed but have subsequently been shown to fail against suitably powerful adversaries[21,22]. In response, provable defense mechanisms that provide robustness guarantees have been developed. One line of work, interval bound propagation, uses interval arithmetic[23,24] to certify neural networks. Another approach makes use of randomizing inputs and adopts techniques from differential privacy[25] and, to our particular interest, statistical hypothesis testing[26,27] that has a natural counter-part in the quantum domain. Since the pioneering works by Helstrom[28]

[1]Department of Computer Science, ETH Zürich, Zürich, Switzerland. [2]Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, China. [3]Ministry of Education, Key Laboratory in Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai, China. [4]University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai, China. [5]Department of Computer Science, University of Illinois, Urbana, IL, USA. ✉email: ce.zhang@inf.ethz.ch; zhikuan.zhao@inf.ethz.ch

and Holevo[29], the task of quantum hypothesis testing (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[30,31], estimation theory[32], to quantum illumination[33,34].

In this work, we lay bare a fundamental connection between QHT and the robustness of quantum classifiers against unknown noise sources. The methods of QHT enable us to derive a robustness condition that, 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 generalization to randomized smoothing, a technique that has recently been applied to certify the robustness of classical machine learning models[26]. 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 how to optimally discriminate between two given states, certifying adversarial robustness aims at giving a guarantee for which two states cannot 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 quantum machine learning, 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 paper is organized as follows. We first introduce the notations and terminologies and review results from QHT essential for our purpose. We then proceed to formally define quantum classifiers and the assumptions on the threat model. In "Results", we present our main findings on provable robustness from QHT. In addition, these results are demonstrated and visualized with a simple toy example for which we also consider the randomized input setting and analyze specifically randomization with depolarization channel. In "Discussion" we conclude with a higher-level view on our findings and layout several related open problems with an outlook for future research. Finally, in "Methods", we give proofs for central results: the robustness condition in terms of type-II error probabilities of QHT, the tightness of this result and, finally, the method used to derive robustness conditions in terms of fidelity.

RESULTS

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 the context, the dimensionality is explicitly indicated through the notation $\mathbb{1}_d$. The set of density operators (i.e., positive semidefinite 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 $A > B$ ($A \geq B$) if $A - B$ is positive (semi-)definite and $A < B$ ($A \leq B$) 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 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 $F(\rho,\ \sigma) := \mathrm{Tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]^2$ that 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 hypothesis testing. Typically, QHT is formulated in terms of state discrimination where several quantum states have to be discriminated through a measurement[28]. 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 $0 \leq M \leq \mathbb{1}_d$, which corresponds to rejecting the null in favor of the alternative. The two central quantities of interest are the probabilities of making a type-I or type-II error. The former corresponds to rejecting the null when it is true, while the latter occurs if the null is accepted when the alternative 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 $\alpha(M)$ and the type-II error probability as $\beta(M)$, so that

$$\alpha(M;\ \sigma) := \mathrm{Tr}[\sigma M] \qquad \text{(type-I error)} \tag{1}$$

$$\beta(M;\ \rho) := \mathrm{Tr}[\rho(\mathbb{1} - M)] \qquad \text{(type-II error)} \tag{2}$$

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

In this paper, we consider asymmetric hypothesis testing (Neyman–Pearson approach)[32], where the two types of errors are associated with a different cost. 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 program (SDP)

$$\begin{aligned} \beta^*_{\alpha_0}(\sigma,\ \rho) := \text{minimize} \quad & \beta(M;\ \rho) \\ \text{s.t.} \quad & \alpha(M;\ \sigma) \leq \alpha_0, \\ & 0 \leq M \leq \mathbb{1}_d \end{aligned} \tag{3}$$

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 $t \geq 0$ let $P_{t,+} := \{\rho - t\sigma > 0\}$, $P_{t,-} := \{\rho - t\sigma < 0\}$ 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 analog to the Neyman–Pearson Lemma[35] shows optimality of operators of the form

$$M_t := P_{t,+} + X_t, \quad 0 \leq X_t \leq P_{t,0}. \tag{4}$$

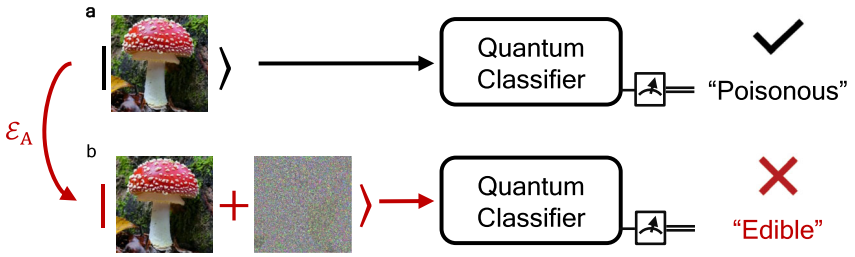

Fig. 1 Adversarial attack. a A quantum classifier correctly classifies the (toxic) mushroom as "poisonous". **b** An adversary perturbs the image to fool the classifier into believing that the mushroom is "edible".

The choice of the scalar $t \geq 0$ and the operator X_t is such that the preassigned type-I error probability α_0 is attained. An explicit construction for these operators is based on the inequalities

$$\alpha(P_{\tau(\alpha_0),+}) \leq \alpha_0 \leq \alpha(P_{\tau(\alpha_0),+} + P_{\tau(\alpha_0),0}) \tag{5}$$

where $\alpha_0 \in (0,1)$ and $\tau(\alpha_0)$ is the smallest non-negative number such that $\alpha(P_{\tau(\alpha_0),+}) \leq \alpha_0$, i.e., $\tau(\alpha_0) := \inf\{t \geq 0 : \alpha(P_{t,+}) \leq \alpha_0\}$. These inequalities can be seen from the observation that the function $t \mapsto \alpha(P_{t,+})$ is non-increasing and right-continuous while $t \mapsto \alpha(P_{t,+} + P_{t,0})$ is non-increasing and left-continuous. A detailed proof for this is given in Supplementary Notes 1 and 2. We will henceforth refer to operators of the form (4) as Helstrom operators[32].

Quantum classifiers. We define a K-class quantum classifier of states of the quantum system $\mathcal{H}$, described by density operators, as a map $\mathcal{A} : \mathcal{S}(\mathcal{H}) \to \mathcal{C}$ that maps states $\sigma \in \mathcal{S}(\mathcal{H})$ to class labels $k \in \mathcal{C} = \{1, \ldots, K\}$. Any such classifier is described by a completely positive and trace-preserving (CPTP) map $\mathcal{E}$ and a positive-operator valued measure (POVM) $\{\Pi_k\}_k$. Formally, a quantum state σ is passed through the quantum channel $\mathcal{E}$ and then the measurement $\{\Pi_k\}_k$ is performed. Finally, the probability of measuring outcome k is identified with the class probability $\mathbf{y}_k(\sigma)$, i.e.,

$$\sigma \mapsto \mathbf{y}_k(\sigma) := \mathrm{Tr}[\Pi_k \mathcal{E}(\sigma)]. \tag{6}$$

We treat the POVM element Π_k as a projector $\Pi_k = |k\rangle\langle k| \otimes \mathbb{1}_{d/K}$ that determines whether the output is classified into class k. 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. The final prediction is given by the most likely class

$$\mathcal{A}(\sigma) \equiv \arg\max_k \mathbf{y}_k(\sigma). \tag{7}$$

Throughout this paper, we refer to $\mathcal{A}$ as the classifier and to $\mathbf{y}$ as the score function. In the context of quantum machine learning, the input state σ can be an encoding of classical data by means of, e.g., amplitude encoding or otherwise[19,36], or inherently quantum input data, while $\mathcal{E}$ can be realized, e.g., by a trained parametrized quantum circuit potentially involving ancillary registers[37]. However, it is worth noting that the above-defined notion of quantum classifier more generally describes the procedure of a broader class of quantum algorithms whose output is obtained by repeated sampling of measurement outcomes.

Quantum adversarial robustness. Adversarial examples are attacks on classification models where an adversary aims to induce a misclassification using typically imperceptible modifications of a benign input example. Specifically, given a classifier $\mathcal{A}$ and a benign input state σ, an adversary can craft a small perturbation $\sigma \to \rho$ that results in a misclassification, i.e., $\mathcal{A}(\rho) \neq \mathcal{A}(\sigma)$. An illustration for this threat scenario is given in Fig. 1. In this paper, 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, then it is guaranteed 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 so as to change the most likely class by applying an arbitrary quantum operation $\mathcal{E}_A : \mathcal{S}(\mathcal{H}) \to \mathcal{S}(\mathcal{H})$ to σ resulting 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 that satisfies $\mathcal{R}$, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$.

In the following, we will derive a robustness condition for quantum classifiers with the QHT formalism, which provides a provable 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 an SDP formulation in terms of the optimal type-II error probabilities. We then simplify this condition and derive closed form solutions in terms of Uhlmann fidelity, Bures metric, and trace distance between benign and adversarial inputs. The closed form solutions in terms of fidelity and Bures metric are shown to be sufficient and necessary for general states and in the same regime where the SDP 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 stemming from QHT considerations are then contrasted and compared with an alternative approach that directly applies Hölder duality to trace distances to obtain a sufficient robustness condition. The different robustness bounds and robustness conditions are summarized in Table 1.

Robustness condition from quantum hypothesis testing

Recall that QHT 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 provable robustness, on the other hand, one aims to find a neighborhood around a benign state σ where the class that is most likely to be measured is constant or, expressed differently, where the classifier cannot discriminate between states. It becomes thus clear that QHT and classification robustness aim to achieve a similar goal,

Table 1. Summary of results.

	Input states	Quantum differential privacy	Hölder duality	Quantum hypothesis testing			
				SDP[a]	Fidelity	Bures metric	Trace distance
No smoothing	Pure	–	Lemma 2[b]	**Theorem 1**	**Theorem 3**	Eq. (20)	Eq. (16)
	Mixed	–					Lemma 2
Depolarization smoothing	Pure	Lemma 2 in ref. [17]	Eq. (45)	**Theorem 1**	–	–	Eq. (43) **(single-qubit)**
	Mixed				–	–	–

In this work, we establish a fundamental connection between QHT and the robustness of quantum classification algorithms against adversarial input perturbations. This connection naturally leads to a robustness condition formulated as a semidefinite program in terms of optimal type-II error probabilities of distinguishing between benign and adversarial states (QHT condition: Theorem 1). Under certain practical assumptions about the class probabilities on benign input, we prove that the QHT condition is optimal (Theorem 2). We then show that the QHT condition implies closed form solutions in terms of explicit robustness bound on the fidelity, Bures metric, and trace distance. We numerically compare an alternative robustness bound directly implied by the definition of trace distance and application of Hölder duality (Lemma 2 and ref. [18]) with the explicit forms of the robustness bounds arising from QHT (Fig. 2). Based on these technical findings, we provide a practical protocol to assess the resilience of a classifier against adversarial perturbations, a protocol to certify whether a given noisy input has been classified the same as the noiseless input, without requiring access to the latter, and we derive robustness bounds on noise parameters in amplitude and phase damping. Finally, we instantiate our results with a single-qubit pure state example both in the noiseless and depolarization smoothing input scenarios, which allows for numerical comparison of all the known robustness bounds, arising from Hölder duality, differential privacy[17], and QHT (Fig. 5). Tight robustness conditions are indicated in bold font.
[a]Robustness condition expressed in terms of type-II error probabilities β^* associated with an optimal quantum hypothesis test.
[b]Independently discovered in ref. [18].

although viewed from different angles. 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 classifier $\mathcal{A}$) can be viewed as a hypothesis test discriminating between σ and ρ. Notice that, for $p_A \in [0,1]$ with $\mathbf{y}_{k_A}(\sigma) = \mathrm{Tr}[\Pi_{k_A}\sigma] \geq p_A$, the operator $\mathbb{1}_d - \Pi_{k_A}$ is feasible for the SDP $\beta^*_{1-p_A}(\sigma,\ \rho)$ in (3) and hence

$$\mathbf{y}_{k_A}(\rho) = \beta(\mathbb{1}_d - \Pi_{k_A};\ \rho) \geq \beta^*_{1-p_A}(\sigma,\ \rho). \tag{8}$$

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 1 (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*

$$\mathbf{y}_{k_A}(\sigma) \geq p_A > p_B \geq \max_{k \neq k_A} \mathbf{y}_k(\sigma). \tag{9}$$

Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with

$$\beta^*_{1-p_A}(\sigma,\ \rho) + \beta^*_{p_B}(\sigma,\ \rho) > 1 \tag{10}$$

To get some more intuition of Theorem 1, we first note that for $p_B = 1 - p_A$, the robustness condition (10) simplifies to

$$\beta^*_{1-p_A}(\sigma,\ \rho) > 1/2 \tag{11}$$

with this, the relation between QHT 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 QHT 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.

Optimality. The robustness condition (10) from QHT is provably optimal in the regime of $p_A + p_B = 1$, which covers binary classifications in full generality and multiclass 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 (10) is violated, then there exists a classifier $\mathcal{A}^\star$ that is consistent with the class probabilities (9) on the benign input but that 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 2 (Tightness) *Suppose that $p_A + p_B = 1$. Then, if the adversarial state ρ violates condition (10), there exists a quantum classifier $\mathcal{A}^\star$ that is consistent with the class probabilities (9) 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 that classifies ρ differently from σ while still being consistent with the class probabilities (9). We refer the reader to "Methods" for a detailed proof. 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 that is consistent with p_A and p_B for benign input but leads to a different classification on adversarial input upon violating condition (10), 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.

Closed form robustness conditions

Although Theorem 1 provides a general condition for robustness with provable tightness, it is formulated as an SDP 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 1, here we derive robustness conditions expressed in terms of Uhlmann's fidelity F, Bures distance d_B, and in terms of the trace distance T. To that end, we first concentrate on pure state inputs and will then leverage 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.

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 (10), Theorem 2 implies tightness of these bounds.

Lemma 1 *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* **y** *satisfies (9). Then, it is guaranteed that $\mathcal{A}(\psi_\rho) = \mathcal{A}(\psi_\sigma)$ for any ψ_ρ with*

$$\left|\langle\psi_\sigma|\psi_\rho\rangle\right|^2 > \frac{1}{2}\left(1+\sqrt{g(p_A,\ p_B)}\right), \tag{12}$$

where the function g is given by

$$g(p_A,\ p_B) = 1 - p_B - p_A(1-2p_B) + 2\sqrt{p_A p_B (1-p_A)(1-p_B)}. \tag{13}$$

This condition is equivalent to (10) and is hence both sufficient and necessary whenever $p_A + p_B = 1$.

This result thus provides a closed form robustness bound that is equivalent to the SDP formulation in condition (10) and is hence sufficient and necessary in the regime $p_A + p_B = 1$. We remark that, under this assumption, the robustness bound (12) has the compact form

$$\left|\langle\psi_\sigma|\psi_\rho\rangle\right|^2 > \frac{1}{2} + \sqrt{p_A(1-p_A)}. \tag{14}$$

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

$$d_B(|\psi_\rho\rangle,\ |\psi_\sigma\rangle) < \left[2 - \sqrt{2(1+\sqrt{g(p_A,p_B)})}\right]^{\frac{1}{2}} \tag{15}$$

is equivalent to (10). Furthermore, since the states are pure, we can directly link (12) to a bound in terms of the trace distance via the relation $T(|\psi_\rho\rangle,|\psi_\sigma\rangle)^2 = 1 - \left|\langle\psi_\sigma|\psi_\rho\rangle\right|^2$, so that

$$T(|\psi_\rho\rangle,\ |\psi_\sigma\rangle) < \left[\frac{1}{2}\left(1-\sqrt{g(p_A,p_B)}\right)\right]^{\frac{1}{2}} \tag{16}$$

is equivalent to (10). Due to the equivalence of these bounds to (10), Theorem 2 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.

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. First, in a realistic scenario, the assumption that an adversary can only produce pure states is too restrictive and gives an incomplete picture. Second, 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 that 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 (12) for mixed states. As for pure states, it is then straight forward to obtain a bound in terms of the Bures metric.

Theorem 3 *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* **y** *satisfies (9). Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with*

$$F(\rho,\ \sigma) > \frac{1}{2}\left(1+\sqrt{g(p_A,\ p_B)}\right) =: r_F \tag{17}$$

where g is defined as in (13). This condition is both sufficient and necessary if $p_A + p_B = 1$.

Proof To show sufficiency of (17), we notice that **y** can be rewritten as

$$\mathbf{y}_k(\sigma) = \mathrm{Tr}[\Pi_k \mathcal{E}(\sigma)] \tag{18}$$

$$= \mathrm{Tr}[\Pi_k(\mathcal{E}\circ \mathrm{Tr}_E)(|\psi_\sigma\rangle\langle\psi_\sigma|)] \tag{19}$$

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 **y** as a score function on the larger Hilbert space that 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 $F(\rho,\ \sigma) = \left|\langle\psi_\sigma|\psi_\rho\rangle\right|^2$. Robustness at ρ then follows from (17) by (18) and Lemma 1. 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 $\tilde{r}_F < \left|\langle\psi_\rho|\psi_\sigma\rangle\right|^2 \le r_F$. However, this is a contradiction, since $\left|\langle\psi_\rho|\psi_\sigma\rangle\right|^2 \ge r_F$ is both sufficient and necessary in the given regime, i.e., by Theorem 2, there exists a classifier $\mathcal{A}^\star$ whose score function satisfies (9) and for which $\mathcal{A}^\star(\psi_\sigma) \neq \mathcal{A}^\star(\psi_\rho)$. It follows that $\tilde{r}_F \ge r_F$ 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

$$d_B(\rho,\ \sigma) < \left[2 - \sqrt{2(1+\sqrt{g(p_A,p_B)})}\right]^{\frac{1}{2}} \tag{20}$$

that inherits the tightness properties of the fidelity bound (17). 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 1 and the trace distance bound for pure states (16) 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 1 (Pure benign and 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* **y** *satisfies (9). Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with*

$$T(\rho,\ \sigma) < \delta(p_A,\ p_B)\left(1-\sqrt{1-\delta(p_A,p_B)^2}\right) \tag{21}$$

where $\delta(p_A,\ p_B) = [\frac{1}{2}(1-g(p_A,p_B))]^{\frac{1}{2}}$.

We refer the reader to Supplementary Note 4 for a detailed proof of this result. Intuitively, condition (21) 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 Eq. (16)) is a natural sufficient condition for robustness. As such, the corresponding robustness radius in condition (21) 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 1 and Eq. (16) to Corollary 1, 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

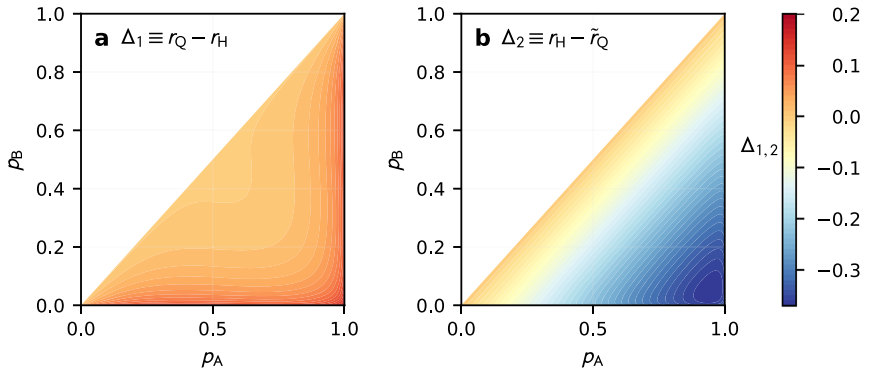

Fig. 2 Comparison between robustness bounds in terms of trace distance. a Difference $r_Q - r_H$ between the pure state bound derived from QHT r_Q, given in Eq. (16) and the Hölder duality bound r_H from Lemma 2. **b** Difference $r_H - \tilde{r}_Q$ between the Hölder duality bound r_H and the bound $\tilde{r}_Q$ derived from the convex hull approximation to the QHT robustness condition from Theorem 1 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 $\tilde{r}_Q$.

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 somewhat straightforward application of Hölder duality for trace norms:

Lemma 2 (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 (9). Then, it is guaranteed that $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$ for any ρ with*

$$\frac{1}{2}||\rho - \sigma||_1 < \frac{p_A - p_B}{2}. \tag{22}$$

Proof Let $\delta := \frac{1}{2}||\rho - \sigma||_1 = \sup_{0 \le P \le \mathbb{I}} \mathrm{Tr}[P(\rho - \sigma)]$, which follows from Hölder duality. We have that $\mathbf{y}_{k_A}(\sigma) - \mathbf{y}_{k_A}(\rho) \le \delta$ and that $\mathbf{y}_{k_A}(\sigma) \ge p_A$, hence $\mathbf{y}_{k_A}(\rho) \ge p_A - \delta$. We also have, for k' such that $\mathbf{y}_{k'}(\rho) = \max_{k \ne k_A} \mathbf{y}_k(\rho)$, that $\mathbf{y}_{k'}(\rho) - \mathbf{y}_{k'}(\sigma) \le \delta$, and that $\mathbf{y}_{k'}(\sigma) \le p_B$, hence $\max_{k \ne k_A} \mathbf{y}_k(\rho) \le p_B + \delta$. Thus, $\frac{1}{2}||\rho - \sigma||_1 < \frac{p_A - p_B}{2} \iff p_A - \delta > p_B + \delta \Rightarrow \mathbf{y}_{k_A}(\rho) > \max_{k \ne k_A} \mathbf{y}_k(\rho)$. □

We acknowledge that the above robustness bound from Hölder duality was independently discovered in Lemma 1 of ref. [18]. For intuitive insights, it is worth remarking that condition (22) 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, (22) is clearly a sufficient condition. However, the Hölder measurement on σ does not necessarily result in class probabilities consistent with Eq. (9). Without additional constraints on desired class probabilities on the benign input, the robustness condition (22) from Hölder duality is stronger than necessary. In contrast, the QHT bound from Theorem 1, 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 1, while for trace distance, tightness is inherited for pure states, but partially lost in Corollary 1 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 Fig. 2.

Toy example with single-qubit pure states

We now present a simple example to highlight the connection between QHT and classification robustness. We consider a single-qubit system that is prepared either in the state σ or ρ described by

$$|\sigma\rangle = |0\rangle, \tag{23}$$

$$|\rho\rangle = \cos(\theta_0/2)|0\rangle + \sin(\theta_0/2)e^{i\phi_0}|1\rangle \tag{24}$$

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 that are central to both QHT and robustness are the Helstrom operators (4) that 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 $t \ge 0$ and given by

$$\eta_1 = \frac{1}{2}(1 - t) + R > 0, \tag{25}$$

$$\eta_2 = \frac{1}{2}(1 - t) - R \le 0 \tag{26}$$

$$R = \frac{1}{2}\sqrt{(1 - t)^2 + 4t(1 - |\gamma|^2)} \tag{27}$$

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

$$|\eta_1\rangle = (1 - \eta_1)A_1|0\rangle - \gamma A_1|\rho\rangle \tag{28}$$

$$|A_1|^{-2} = 2R|\eta_1 - \sin^2(\theta_0/2)| \tag{29}$$

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 $\alpha(M_t) = \alpha_0$.

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 Fig. 3. We note that, for this choice of states, we obtain an expression for the eigenvector $|\eta_1\rangle$ given by

$$|\eta_1\rangle = \frac{9 - \sqrt{3}}{\sqrt{30}}|0\rangle - 3\sqrt{\frac{2}{5}}|\rho\rangle. \tag{30}$$

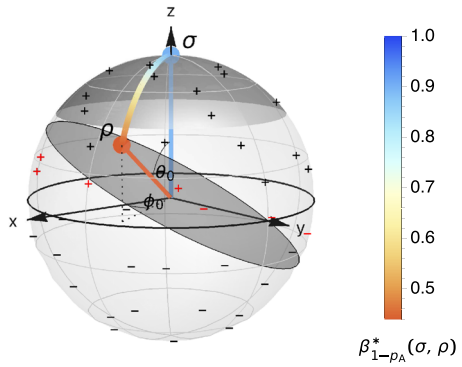

Fig. 3 Example classifier for single-qubit quantum states. The decision boundary is represented by the gray disk passing through the origin of the Bloch sphere. The robust region around σ is indicated by the dark spherical cap. States belonging to different classes are marked with + and – and are color 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 that admits 0.1 type-I error probability for testing σ against ρ has type-II error probability 0.44.

that yields the type-II error probability

$$\beta^*_{1-p_A}(\sigma,\ \rho) = \ \beta(M_t) = 1 - |\langle\eta_1|\rho\rangle|^2 \approx 0.44 < 1/2. \tag{31}$$

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. However, since $\beta(M_t) \not> 1/2$, Theorem 1 implies that ρ is not guaranteed to be classified equally as σ by a classifier that 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 1 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

$$\beta^*_{1-p_A}(\sigma,\ \rho) > 1/2. \tag{32}$$

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 (17), we find the robustness condition on θ_0

$$\begin{aligned} |\langle\rho|\sigma\rangle|^2 = \cos^2(\theta_0/2) > \tfrac{1}{2} + \sqrt{p_A(1-p_A)} \\ \iff\ \theta_0 < 2\cdot\arccos\sqrt{\tfrac{1}{2} + \sqrt{p_A(1-p_A)}}. \end{aligned} \tag{33}$$

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 3 illustrates this scenario: the dark region around σ contains all states ρ for which it 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 $\mathcal{A}$ that 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

$$|\psi_{\theta,\phi}\rangle := \cos(\theta/2)|0\rangle + \sin(\theta/2)e^{i\phi}|1\rangle \tag{34}$$

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 Fig. 3, where the decision boundary of $\mathcal{A}$ is represented by the gray disk crossing the origin of the Bloch sphere. The states marked with a black + correspond to + states that 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.

Robustness certification

The theoretical results in "Closed form robustness conditions" 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. Second, 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. Third, we analyze the robustness of quantum classifiers against known noise models, namely phase and amplitude damping.

Assessing resilience against adversaries. In security critical applications, such as 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 that has been classified correctly and with a robust radius of at least r (i.e., an adversary cannot 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 labeled samples

$\mathcal{T} = \{(\sigma_i,\ y_i)\}_{i=1}^{|\mathcal{T}|}$, the certified test set accuracy at fidelity $1-\varepsilon$ is given by

$$\frac{1}{|\mathcal{T}|}\sum_{(\sigma,y)\in\mathcal{T}} \mathbb{1}\{\mathcal{A}(\sigma) = y \ \wedge\ r_F(\sigma) \le 1-\varepsilon\} \tag{35}$$

where $r_F(\sigma)$ is the minimum robust fidelity (17) 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 (17) 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

$$\Pr\left[\frac{n_k}{N} - \langle\Lambda_k\rangle_\sigma \ge \delta\right] \le \exp\{-2N\delta^2\} \tag{36}$$

with $\Lambda_k = \mathcal{E}^\dagger(\Pi_k)$ and setting $\delta = \sqrt{-\ln(\alpha)/2N}$. In Supplementary Note 6, this algorithm is shown in detail in Protocol 1.

Certification 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 that 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 3 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 that 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 (17) 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 us 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.

Robustness for known noise models. Now, we analyze the robustness of a quantum classifier against known noise models that are parametrized by a noise parameter γ. Specifically, we investigate robustness against phase damping and amplitude damping. Using Theorem 3, we calculate the fidelity between the clean input σ and the noisy input $\mathcal{N}_\gamma(\sigma)$ and rearrange the robustness condition (17) such that it yields a bound on the maximal noise that the classifier tolerates.

Phase damping describes the loss of quantum information without losing 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 that are given by

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}, \quad K_1 = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{\gamma} \end{pmatrix} \tag{37}$$

where γ is the noise parameter. From this description alone, we can see that a system that 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 behavior 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 3 leads to the robustness condition $\gamma \le 1$ if $\alpha = 0$ or $\beta = 0$ and, for any $\alpha, \beta \ne 0$,

$$\gamma < 1 - \left(\max\left\{0,\ 1 + \frac{r_F - 1}{2|\alpha|^2|\beta|^2}\right\}\right)^2 \tag{38}$$

where $r_F = \frac{1}{2}(1 + \sqrt{g(p_A,\ p_B)})$ is the fidelity bound from Theorem 3 and p_A, p_B are the corresponding class probability bounds. This bound is illustrated in Fig. 4 as a function of $|\alpha|^2$ and p_A. The expected behavior toward 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 $\gamma \le 1$.

Amplitude damping models effects 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 that spontaneously emits a photon. The quantum channel corresponding to this noise model can be written in terms of Kraus operators

$$K_0 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}, \quad K_1 = \begin{pmatrix} 0 & \sqrt{\gamma} \\ 0 & 0 \end{pmatrix}, \tag{39}$$

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 3 leads to the robustness condition $\gamma \le 1$ if $|\alpha| = 1$ and, for any $\alpha, \beta \ne 0$,

$$\gamma < 1 - \left[\frac{|\alpha|^2}{|\alpha|^2 - |\beta|^2} \cdot \left(1 - \sqrt{1 - \frac{|\alpha|^2 - |\beta|^2}{|\alpha|^2|\beta|^2} \cdot \frac{\max\{0, r_F - |\alpha|^2\}}{|\alpha|^2}}\right)\right]^2 \tag{40}$$

where again $r_F = \frac{1}{2}(1 + \sqrt{g(p_A,\ p_B)})$ is the fidelity bound from Theorem 3. This bound is illustrated in Fig. 4 as a function of $|\alpha|^2$ and p_A. It can be seen again that the bound shows the expected behavior, namely that when $|\alpha|^2 \to 1$, then the classifier is robust under all noise parameters $\gamma \le 1$.

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.

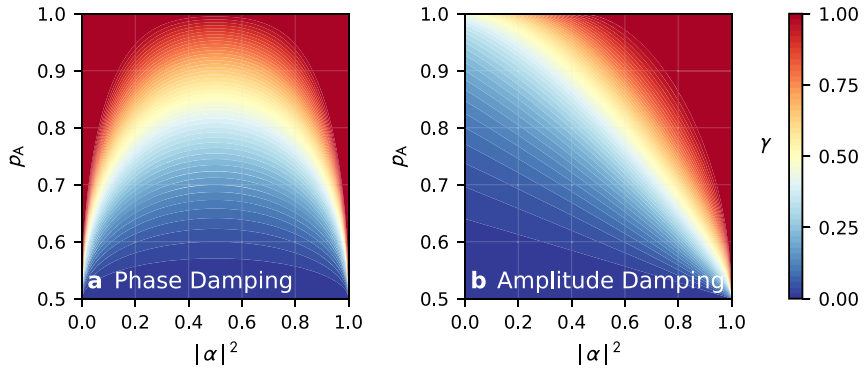

Fig. 4 Robustness against known noise models. Both plots show the maximal noise parameter γ for which the classifier $\mathcal{A}$ is still guaranteed to be robust, for (**a**) phase damping and (**b**) amplitude damping, when classifying a pure state input $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. In **a**, we can see that for states $|0\rangle$ and $|1\rangle$, the classifier is robust against any $\gamma \leq 1$, while for (**b**) the same holds if the input state is $|1\rangle$.

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 provide background on randomized smoothing, a technique for provable robustness from classical machine learning. We then proceed to present provable robustness in terms of trace distance that is equivalent to the robustness condition (10) from Theorem 1 but with depolarized inputs. The bound is then compared numerically with the Hölder duality bound from Lemma 2 and with a result obtained recently from quantum differential privacy[17].

Randomized smoothing. Randomized smoothing is a technique that has recently been proposed to certify the robustness and obtain tight provable robustness guarantees in the classical setting[26]. The key idea is to randomize inputs to classifiers by perturbing them with additive Gaussian noise. This results in smoother decision boundaries that in turn leads to improved robustness to adversarial attacks. In this section, we extend this concept to the quantum setting by interpreting quantum noise channels as "smoothing" channels. The idea of harnessing actively induced input noise in quantum classifiers to increase robustness has recently been proposed in ref. [17] where a robustness bound with techniques from quantum differential privacy has been derived. In the following, we take a similar path and consider a depolarization noise channel and analytically derive a larger robustness radius for pure single-qubit input states.

Quantum channel smoothing: depolarization. Consider depolarization noise that maps a state σ onto a linear combination of itself and the maximally mixed state

$$\sigma \mapsto \mathcal{E}_p^{\mathrm{dep}}(\sigma) := (1-p)\sigma + \frac{p}{d}\mathbb{1}_d \tag{41}$$

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 toward 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 1 in the following way. Let σ be a benign input state and suppose that the classifier $\mathcal{A}$ with score function $\mathbf{y}$ satisfies

$$\mathbf{y}_{k_A}(\mathcal{E}_p^{\mathrm{dep}}(\sigma)) \geq p_A > p_B \geq \max_{k \neq k_A} \mathbf{y}_k(\mathcal{E}_p^{\mathrm{dep}}(\sigma)). \tag{42}$$

Then $\mathcal{A}$ is robust at $\mathcal{E}_p^{\mathrm{dep}}(\rho)$ for any adversarial input state ρ that satisfies the robustness condition (10), where β^* is the optimal type-II error probability for testing $\mathcal{E}_p^{\mathrm{dep}}(\sigma)$ against $\mathcal{E}_p^{\mathrm{dep}}(\rho)$. 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 as $T(\rho,\sigma) < r_Q(p)$ with

$$r_Q(p) = \begin{cases} \sqrt{\frac{1}{2} - \frac{\sqrt{g(p,\,p_A)}}{1-p}}, & p_A < \frac{1+3(1-p)^2}{2+2(1-p)^2} \\ \sqrt{\frac{p\cdot(2-p)\cdot(1-2p_A)^2}{8(1-p)^2\cdot(1-p_A)}}, & p_A \geq \frac{1+3\cdot(1-p)^2}{2+2\cdot(1-p)^2} \end{cases} \tag{43}$$

where

$$g(p,\,p_A) = \frac{1}{2}\left(2p_A(1-p_A) - p\left(1-\frac{p}{2}\right)\right). \tag{44}$$

A detailed derivation of this bound is given in Supplementary Note 5.

The Hölder bound from Lemma 2 can also be adapted to the noisy setting. Specifically, since for two states σ and ρ, the trace distance obeys $T(\mathcal{E}_p^{\mathrm{dep}}(\rho),\ \mathcal{E}_p^{\mathrm{dep}}(\sigma)) = (1-p)\cdot T(\rho,\ \sigma)$, Lemma 2 implies robustness given that the trace distance is less than $T(\rho,\sigma) < r_H(p)$ where

$$r_H(p) = \frac{2p_A - 1}{2(1-p)}. \tag{45}$$

It has been shown in ref. [17] 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

$$r_{DP}(p) = \frac{p}{2(1-p)}\left(\sqrt{\frac{p_A}{1-p_A}} - 1\right) \tag{46}$$

and robustness is guaranteed for any adversarial state ρ with $T(\rho,\sigma) < r_{DP}(p)$. The three bounds are compared graphically in Fig. 5 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

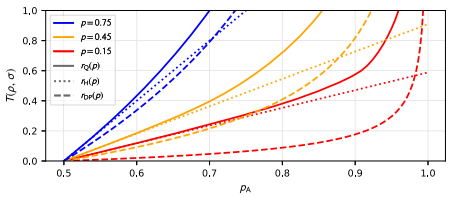

Fig. 5 Robustness bounds with depolarized input states. Comparison of robustness bounds for single-qubit pure states derived from quantum hypothesis testing $r_Q(p)$, Hölder duality $r_H(p)$, and quantum differential privacy $r_{DP}(p)$[17] with different levels of depolarization noise p.

presents a valid, non-trivial condition with noiseless (without smoothing) quantum input (Theorems 1, 3, Corollary 1, and Lemma 2). This contrasts 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[26]. 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.

DISCUSSION

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 optimal 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 randomized smoothing in robustness certification of classical machine learning.

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. We now know that both classical and quantum optimal robustness bounds for classification protocols depend on bounds provided by hypothesis testing. However, hypothesis testing involves the comparison of probability distributions, which can only be possible in the classical case with the addition of stochastic noise sources if the classical data are initially non-stochastic. This means that the optimal robustness bounds in the classical case only exist for noisy classifiers that also require training under the additional noise[26]. This is in contrast to the quantum scenario. Our quantum adversarial robustness bound can be proved independently of randomized input, even though it can be enhanced by it, like through a depolarization channel. Thus, in the quantum regime, unlike in the classical deterministic scenario, we are not forced to consider training under actively induced noise.

Our optimal provable robustness bound and the connection to QHT also provide a first step toward more rigorously identifying the limitations of quantum classifiers in its power of distinguishing between 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[38] and is an important direction of future research. It is also of independent interest to explore possible connections between tasks that use QHT, such as quantum illumination[33] and state discrimination[39], with accuracy and robustness in quantum classification.

METHODS

Proof of Theorem 1

The proof of this theorem is based on showing that the measurement operators of the classifier can be viewed as an operator that is feasible for the SDP (3). Specifically, note that in the Heisenberg picture we can write the score function $\mathbf{y}$ of the classifier $\mathcal{A}$ as

$$\mathbf{y}_k(\sigma) = \mathrm{Tr}\left[\mathcal{E}^\dagger(\Pi_k)\sigma\right] = \mathrm{Tr}[\Lambda_k\sigma] \tag{47}$$

where $\Lambda_k := \mathcal{E}^\dagger(\Pi_k)$. Since $\mathcal{E}$ is a CPTP map, its dual is completely positive and unital and thus $0 \le \Lambda_k \le \mathbb{1}$ and

$$\sum_k \Lambda_k = \sum_k \mathcal{E}^\dagger(\Pi_k) = \mathcal{E}^\dagger(\mathbb{1}) = \mathbb{1}. \tag{48}$$

Note that the operator $\mathbb{1} - \Lambda_{k_A}$ is feasible for the SDP $\beta^*_{1-p_A}(\sigma,\ \rho)$ since by assumption

$$\alpha(\mathbb{1} - \Lambda_{k_A};\ \sigma) = 1 - \mathbf{y}_{k_A}(\sigma) \le 1 - p_A. \tag{49}$$

It follows that

$$\mathbf{y}_{k_A}(\rho) = \beta(\mathbb{1} - \Lambda_{k_A};\ \rho) \ge \beta^*_{1-p_A}(\sigma,\ \rho). \tag{50}$$

Similarly, let $k \neq k_A$ be arbitrary. Then, the operator Λ_k is feasible for the SDP $\beta^*_{p_B}(\sigma,\ \rho)$ since

$$\alpha(\Lambda_k;\ \sigma) = \mathbf{y}_k(\sigma) \le p_B \tag{51}$$

and hence

$$1 - \mathbf{y}_k(\rho) = \beta(\Lambda_k;\ \rho) \ge \beta^*_{p_B}(\sigma,\ \rho) \tag{52}$$

Since $k \neq k_A$ is arbitrary, it follows that if ρ satisfies

$$\beta^*_{1-p_A}(\sigma,\ \rho) + \beta^*_{p_B}(\sigma,\ \rho) > 1 \tag{53}$$

then it is guaranteed that

$$\mathbf{y}_{k_A}(\rho) > \max_{k \neq k_A} \mathbf{y}_k(\rho) \tag{54}$$

and thus $\mathcal{A}(\rho) = \mathcal{A}(\sigma)$. □

Proof of Theorem 2

Note that, since $p_B = 1 - p_A$ by assumption, the robustness condition (10) reads

$$\beta^*_{1-p_A}(\sigma,\ \rho) > 1/2. \tag{55}$$

Let $M^\star_A$ be an optimizer of the corresponding SDP such that $\alpha(M^\star_A) = 1 - p_A$ and

$$\beta(M^\star_A;\ \rho) = \beta^*_{1-p_A}(\sigma,\ \rho). \tag{56}$$

Consider the classifier $\mathcal{A}^\star$ with score function $\mathbf{y}^\star$ defined by the POVM $\{\mathbb{1} - M^\star_A,\ M^\star_A,\ 0\}$ 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 (9) since

$$\mathbf{y}^\star_{k_A}(\sigma) = \alpha(\mathbb{1} - M^\star_A;\ \sigma) = p_A \tag{57}$$

$$\mathbf{y}^\star_{k_B}(\sigma) = \alpha(M^\star_A;\ \sigma) = 1 - p_A = p_B. \tag{58}$$

Furthermore, if ρ violates (55), then we have

$$\mathbf{y}_{k_A}(\rho) = \beta(M^\star_A;\ \rho) \le 1/2 \tag{59}$$

and thus, in particular $\mathcal{A}^\star(\rho) \neq k_A = \mathcal{A}^\star(\sigma)$. □

Fidelity robustness condition

Recall that the robustness condition in Theorem 1 is expressed in terms of the SDP from the Neyman–Pearson approach to QHT. Thus, in order to use Theorem 1 to obtain robustness bounds in terms of a meaningful distance between quantum states, we need to connect the optimal type-II error with this distance. Here, we look specifically at the fidelity between pure quantum states and sketch the proof for Lemma 1. We refer the reader to Supplementary Note 3 for details.

Proof. Proof of Lemma 1 (sketch). The key challenge to proving this result is connecting the robustness condition (10), written in terms of type-II error probabilities, to the fidelity F which, for pure states, is given by the squared overlap $\left|\langle\psi_\sigma|\psi_\rho\rangle\right|^2$. It is well known that optimizers to the SDP (3) are given by Helstrom operators, M_t, which can be expressed in terms of the projection onto the positive and null eigenspaces of the operator $\rho - t\sigma$. The first step is thus to solve the eigenvalue problem

$$(\rho - t\sigma)|\eta\rangle = \eta|\eta\rangle \tag{60}$$

which, for pure states, can be expressed in terms of the squared overlap $\left|\langle\psi_\sigma|\psi_\rho\rangle\right|^2$. Given these solutions, one then derives an expression for the Helstrom operators $M_{\mathrm{A}}^{\star}$ and $M_{\mathrm{B}}^{\star}$ with type-I error probabilities $1 - p_{\mathrm{A}}$ and p_{B}, respectively. This leads to the robustness condition $\beta(M_{\mathrm{A}}^{\star};\ \rho) + \beta(M_{\mathrm{B}}^{\star};\ \rho) > 1$ being an inequality that can be rewritten as a condition on the fidelity that takes the desired form (12). □

In a similar manner, one can derive the trace distance bound for depolarized input states presented in the "Results" section of this paper. The full proof for the robustness bound in Eq. (43) is given in Supplementary Note 5.

Received: 21 October 2020; Accepted: 14 April 2021;
Published online: 21 May 2021

REFERENCES

1. Dunjko, V. & Briegel, H. J. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Rep. Prog. Phys.* **81**, 074001 (2018).
2. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195–202 (2017).
3. Schuld, M., Sinayskiy, I. & Petruccione, F. An introduction to quantum machine learning. *Contemp. Phys.* **56**, 172–185 (2015).
4. Zhao, Z., Pozas-Kerstjens, A., Rebentrost, P. & Wittek, P. Bayesian deep learning on a quantum computer. *Quantum Mach. Intell.* **1**, 41–51 (2019).
5. Cong, I., Choi, S. & Lukin, M. D. Quantum convolutional neural networks. *Nat. Phys.* **15**, 1273–1278 (2019).
6. Rebentrost, P., Mohseni, M. & Lloyd, S. Quantum support vector machine for big data classification. *Phys. Rev. Lett.* **113**, 130503 (2014).
7. Farhi, E. & Neven, H. Classification with quantum neural networks on near term processors. *Quantum Rev. Lett.* **1**, 129–153 (2020).
8. Havlíček, V. et al. Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209–212 (2019).
9. Schuld, M. & Killoran, N. Quantum machine learning in feature hilbert spaces. *Phys. Rev. Lett.* **122**, 040504 (2019).
10. Lloyd, S., Schuld, M., Ijaz, A., Izaac, J. & Killoran, N. Quantum embeddings for machine learning. Preprint at https://arxiv.org/abs/2001.03622 (2020).
11. Lu, S., Duan, L.-M. & Deng, D.-L. Quantum adversarial machine learning. *Phys. Rev. Res.* **2**, 033212 (2020).
12. Liu, N. & Wittek, P. Vulnerability of quantum classification to adversarial perturbations. *Phys. Rev. A* **101**, 062331 (2020).
13. Szegedy, C. et al. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations* (2014).
14. Goodfellow, I., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations* (2015).
15. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
16. Wiebe, N. & Kumar, R. S. S. Hardening quantum machine learning against adversaries. *New J. Phys.* **20**, 123019 (2018).
17. Du, Y., Hsieh, M.-H., Liu, T., Tao, D. & Liu, N. Quantum noise protects quantum classifiers against adversaries. Preprint at https://arxiv.org/abs/2003.09416 (2020).
18. Guan, J., Fang, W. & Ying, M. Robustness verification of quantum machine learning. Preprint at https://arxiv.org/abs/2008.07230 (2020).
19. LaRose, R. & Coyle, B. Robust data encodings for quantum classifiers. *Phys. Rev. A* **102**, 032420 (2020).
20. Eykholt, K. et al. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1625–1634 (2018).
21. Carlini, N. & Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 3–14 (2017).
22. Athalye, A., Carlini, N. & Wagner, D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, 274–283 (2018).
23. Gowal, S. et al. On the effectiveness of interval bound propagation for training verifiably robust models. Preprint at https://arxiv.org/abs/1810.12715 (2018).
24. Mirman, M., Gehr, T. & Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, 3578–3586 (2018).
25. Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D. & Jana, S. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, 656–672 (IEEE, 2019).
26. Cohen, J., Rosenfeld, E. & Kolter, Z. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 1310–1320 (2019).
27. Weber, M., Xu, X., Karlaš, B., Zhang, C. & Li, B. Rab: provable robustness against backdoor attacks. Preprint at https://arxiv.org/abs/2003.08904 (2020).
28. Helstrom, C. W. Detection theory and quantum mechanics. *Inform. Control* **10**, 254–291 (1967).
29. Holevo, A. S. Statistical decision theory for quantum systems. *J. Multivar. Anal.* **3**, 337–394 (1973).
30. Wang, L. & Renner, R. One-shot classical-quantum capacity and hypothesis testing. *Phys. Rev. Lett.* **108**, 200501 (2012).
31. Matthews, W. & Wehner, S. Finite blocklength converse bounds for quantum channels. *IEEE Trans. Inf. Theory* **60**, 7317–7329 (2014).
32. Helstrom, C. W. *Quantum Detection and Estimation Theory*. (Academic Press, New York, NY, 1976).
33. Wilde, M. M., Tomamichel, M., Lloyd, S. & Berta, M. Gaussian hypothesis testing and quantum illumination. *Phys. Rev. Lett.* **119**, 120501 (2017).
34. Lloyd, S. Enhanced sensitivity of photodetection via quantum illumination. *Science* **321**, 1463–1465 (2008).
35. Neyman, J. & Pearson, E. S. Ix. on the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Statistical Soc. A* **231**, 289–337 (1933).
36. Zhao, Z., Fitzsimons, J. K., Rebentrost, P., Dunjko, V. & Fitzsimons, J. F. Smooth input preparation for quantum and quantum-inspired machine learning. *Quantum Mach. Intell.* **3**, 1–6 (2021).
37. Benedetti, M., Lloyd, E., Sack, S. & Fiorentini, M. Parameterized quantum circuits as machine learning models. *Quantum Sci. Technol.* **4**, 043001 (2019).
38. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. & Madry, A. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations* (2019).
39. Sentís, G., Monras, A., Muñoz-Tapia, R., Calsamiglia, J. & Bagan, E. Unsupervised classification of quantum data. *Phys. Rev. X* **9**, 041029 (2019).

ACKNOWLEDGEMENTS

The authors are grateful to Ryan LaRose (Michigan State University), Zi-Wen Liu (Perimeter Institute for Theoretical Physics), Barry Sanders (University of Calgary), and Robert Pisarczyk (University of Oxford) for inspiring discussions on the question of robustness in quantum machine learning. N.L. acknowledges funding from the Shanghai Pujiang Talent Grant (no. 20PJ1408400) and the NSFC International Young Scientists Project (no. 12050410230). N.L. is also supported by the Innovation Program of the Shanghai Municipal Education Commission (no. 2021-01-07-00-02-E00087) and the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

AUTHOR CONTRIBUTIONS

The main idea was conceived by C.Z. in discussion with Z.Z. and M.W. Key insights to adversarial quantum learning were provided by N.L. while B.L. contributed to central insights to robustness in machine learning. The work on QHT, the resulting QHT condition, and the derivation for closed form bounds for pure and depolarized states was completed by M.W. The extension of the fidelity bound for the mixed state case

was completed by M.W. and Z.Z. The different trace distance bounds for mixed states were derived by Z.Z. The proof for optimality was done by M.W. and Z.Z. The noisy input scenario and the example were initiated by N.L. and completed by M.W. All authors contributed to the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41534-021-00410-5.

Correspondence and requests for materials should be addressed to C.Z. or Z.Z.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.