Information-theoretic generalization bounds for learning from quantum data

Matthias C. Caro MATTHIAS.CARO@FU-BERLIN.DE
Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Berlin, Germany
Institute for Quantum Information and Matter, Caltech, Pasadena, CA, USA

Tom Gur TOM.GUR@CL.CAM.AC.UK
Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

Cambyse Rouzé CAMBYSE.ROUZE@TELECOM-PARIS.FR
Inria, Télécom Paris - LTCI, Institut Polytechnique de Paris, Palaiseau, France
Zentrum Mathematik, Technische Universität München, Garching, Germany

Daniel Stilck França DANIEL.STILCK_FRANCA@ENS-LYON.FR
Univ Lyon, ENS Lyon, UCBL, CNRS, Inria, LIP, F-69342, Lyon Cedex 07, France

Sathyawageeswar Subramanian SS2310@CAM.AC.UK
Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
Department of Computer Science, University of Warwick, Coventry, UK

Editors: Shipra Agrawal and Aaron Roth

Abstract

Learning tasks play an increasingly prominent role in quantum information and computation. They range from fundamental problems such as state discrimination and metrology over the framework of quantum probably approximately correct (PAC) learning, to the recently proposed shadow variants of state tomography. However, the many directions of quantum learning theory have so far evolved separately. We propose a mathematical formalism for describing quantum learning by training on classical-quantum data and then testing how well the learned hypothesis generalizes to new data. In this framework, we prove bounds on the expected generalization error of a quantum learner in terms of classical and quantum information-theoretic quantities measuring how strongly the learner's hypothesis depends on the data seen during training. To achieve this, we use tools from quantum optimal transport and quantum concentration inequalities to establish non-commutative versions of decoupling lemmas that underlie classical information-theoretic generalization bounds.

Our framework encompasses and gives intuitive generalization bounds for a variety of quantum learning scenarios such as quantum state discrimination, PAC learning quantum states, quantum parameter estimation, and quantumly PAC learning classical functions. Thereby, our work lays a foundation for a unifying quantum information-theoretic perspective on quantum learning.

Keywords: quantum learning, generalization bounds, quantum mutual information, quantum optimal transport

1. Introduction

The intersection of machine learning and quantum physics has developed into a vibrant area of research. On the one hand, along the lines of using (at least partially) quantum learners for classical data, there are proposals for machine learning models based on quantum circuits (Biamonte et al., 2017; Dunjko and Briegel, 2018; Havlíček et al., 2019), such as the so-called variational quantum machine learning models and quantum kernel methods. On the other hand, there has been significant progress in learning from quantum data. Inspired by "pretty good tomography" (Aaronson, 2007), viewing quantum experiments through the lens of learning from quantum data has given rise to 'shadow' protocols (Aaronson, 2019; Huang et al., 2020) that use few copies of an unknown quantum state to predict many of its properties. The learning perspective has also led to insights into the potential for quantum advantage of fully quantum over conventional experiments (Huang et al., 2021; Aharonov et al., 2022; Chen et al., 2022b,a; Huang et al., 2022; Caro, 2022a; Chen et al., 2023a,b). Moreover, from the viewpoint of computer science, quantum theory allows for new kinds of oracular access to an unknown object that is to be learned (Bshouty and Jackson, 1998), and thus potentially (though not always) for more efficient learning algorithms (Arunachalam and de Wolf, 2017). Even fundamental problems of quantum information theory, such as state or process tomography (Haah et al., 2016; O'Donnell and Wright, 2016; Haah et al., 2023; Zhao et al., 2023) or state discrimination (Helstrom, 1969; Holevo, 1974; Yuen et al., 1975), can be interpreted as tasks of learning from quantum data (Guţă and Kotłowski, 2010; Sentís et al., 2019).

As quantum machine learning and quantum learning theory have grown, so has the number of different quantum learning scenarios and mathematical descriptions thereof. This is reminiscent of the plethora of approaches to generalization and sample complexity bounds in classical machine learning theory (Vapnik and Chervonenkis, 1971; Pollard, 1984; Littlestone and Warmuth, 1986; Kearns and Schapire, 1994; Dudley, 1999; McAllester, 1999; Bousquet and Elisseeff, 2002; Bartlett and Mendelson, 2002; Dwork et al., 2006). Recently, information-theoretic generalization bounds (Hellström et al., 2023), going back to (Xu and Raginsky, 2017; Russo and Zou, 2019), have emerged as a promising approach towards unifying these varied results. Furthermore, they may help overcome the limitations of uniform generalization bounds (Zhang et al., 2017, 2021), which have recently also been pointed out for quantum machine learning models (Gil-Fuster et al., 2024). However, a similarly unifying perspective on quantum learning has so far been lacking.

In the spirit of unification, we propose a mathematical framework for quantum learning procedures that train on data composed of classical samples as well as quantum data states, and then produce a classical and/or quantum hypothesis to be used for prediction on new classical-quantum data. We prove that the generalization behavior of such quantum learners – that is, how well they generalize from available training data to previously unseen data – can be controlled through classical and quantum information-theoretic quantities, which quantify how much information the learner's hypothesis contains about the data, combined with concentration properties of the loss observables used for training. We demonstrate several applications of this quantum version of the central insight from (Xu and Raginsky, 2017; Russo and Zou, 2019). To mention a few, it allows us to provide a new perspective on quantum state classification tasks (Guţă and Kotłowski, 2010), and recover the seminal result of (Aaronson, 2007) on PAC learning quantum states as well as the results of (Chung and Lin, 2021; Caro, 2021; Fanizza et al., 2022) on learning state preparation procedures.

1.1. Main results

Our first contribution is a unifying framework capable of capturing a wide variety of quantum learning problems. Having formulated the framework, we then use it to prove information-theoretic generalization bounds for quantum learners and demonstrate applications to learning quantum states, learning classical functions from entangled quantum data, and quantum state classification.

1.1.1. UNIFIED INFORMATION-THEORETIC FRAMEWORK

Learners as maps. Classical randomized (supervised) learning algorithms can be modeled as channels. They take as input *training data*, which is a set $S = (Z_1, \ldots, Z_m)$ of $|S| = m$ i.i.d. data points drawn from a probability distribution P over an *instance space* Z. The output of a learner is a random variable called the *hypothesis* taking values in a *hypothesis space* W. We often think of the input domain Z as being a Cartesian product $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$, and the hypothesis space W as a subset of Y^{X}, so that a hypothesis is in fact a (randomized) *function* $w : \mathsf{X} \to \mathsf{Y}$. The learner can then be identified with a conditional probability distribution $P(W|S)$ for the hypothesis given the data.

In analogy, we propose to think of *quantum* learning algorithms $\mathcal{A}$ as quantum procedures that take as input data represented by a quantum state ρ coming from a quantum instance space $\mathcal{Z}$. The output of a quantum learner is a hypothesis state taking values in a space $\mathcal{W}$. In particular, without loss of generality, we can take $\mathcal{Z}$ to be a space of *classical-quantum* "CQ states" of the form

$$\rho = \mathop{\mathbb{E}}_{S \sim P^m} \Big[|S\rangle\langle S| \otimes \rho(S) \Big] , \tag{1}$$

where $\rho(S)$ is a quantum state on the Hilbert space $\mathcal{H}_{\mathrm{train}}$. Typically, we consider $\mathcal{H}_{\mathrm{train}} \cong \bigotimes_{i=1}^{m} \mathbb{C}^d$, where d is the local dimension, and assume the quantum training data to factorize as $\rho(S) = \bigotimes_{i=1}^{m} \rho(Z_i)$ for d-dimensional states $\rho(Z_i)$. (Note that we consider an analogous classical "factorization" by working with i.i.d. data $S \sim P^m$.) Similarly $\mathcal{W}$ consists of states of the form

$$\sigma^{\mathcal{A}} = \mathop{\mathbb{E}}_{(S,W) \sim P^{\mathcal{A}}} \Big[|S, W\rangle\langle S, W| \otimes \sigma^{\mathcal{A}}(S, W) \Big] , \tag{2}$$

where $P^{\mathcal{A}}$ is a joint distribution over data and hypothesis induced by the learner, and $\sigma^{\mathcal{A}}(S, W)$ is a quantum state on the Hilbert space $\mathcal{H}_{\mathrm{hyp}}$. The learning procedure consists of two steps that can be iterated: measurement and post-processing. The measurements may be implemented by positive operator-valued measures (POVMs), and the associated instruments. Here, a POVM maps states to classical probability distributions over outcomes, and the instruments give the corresponding mappings to post-measurement states. We allow randomized classical post-processing of the measurement outcomes as well as quantum post-processing of the post-measurement states.

Risk for classical learners. In classical learning theory, the performance of a hypothesis on a data point is evaluated by a loss function $\ell : \mathsf{W} \times \mathsf{Z} \to \mathbb{R}_{\geq 0}$. Accordingly, the *true risk* of a hypothesis $w \in \mathsf{W}$ relative to the distribution P is

$$R_P(w) = \mathop{\mathbb{E}}_{Z \sim P} [\ell(w, Z)]. \tag{3}$$

The goal of a learner is to output a randomized hypothesis W that has small true risk $R_P(W)$, either in expectation or with high success probability. However, the data distribution P is typically unknown, so the learner cannot directly evaluate $R_P(w)$ for a candidate hypothesis w. Instead, the average loss of a hypothesis on available training data serves as a proxy for the true risk. For training data $S = (Z_1, \dots, Z_m)$ and hypothesis $w \in \mathsf{W}$, the *empirical risk* is defined by

$$\hat{R}_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, Z_i) \,. \tag{4}$$

In contrast to $R_P(w)$, a classical learner with access to S can in principle evaluate $\hat{R}_S(w)$ for any $w \in \mathsf{W}$. When the focus is on the average performance of a learner, the quality of $\hat{R}_S(W)$ as a proxy for $R_P(W)$ may be quantified by the *expected generalization error*

$$\mathrm{gen}_P(\mathcal{A}) = \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}} \left[R_P(W) - \hat{R}_S(W) \right] . \tag{5}$$

We refer to bounds on $\mathrm{gen}_P(\mathcal{A})$ simply as generalization bounds[1]. Such bounds then give rise to guarantees on when successful training, quantified by small empirical risk, leads to small true risk.

Risk for quantum learners. In translating the above recipe for evaluating the performance of a learner to the quantum scenario, we encounter a fundamental obstacle: Quantum data that has been used for training may be irreversibly modified by measurements and post-processing, and cannot be reused for evaluating the empirical risk of a hypothesis obtained at the end of the training process.

Therefore, we introduce an additional quantum system to capture test data. That is, we now allow $\rho(S)$ in the quantum data state of Equation (1) to be states on a composite Hilbert space $\mathcal{H}_{\mathrm{data}} = \mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{train}}$. Note that $\rho(S)$ can be correlated or even entangled across the test-train bipartition of the data Hilbert space. The action of the learner on the training data subsystem then leads to a hypothesis state as in Equation (2), with $\sigma^{\mathcal{A}}(S, W)$ now a quantum state on the Hilbert space $\mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{hyp}}$.

Lifting the notion of loss function to a quantum observable, we work with a family of (non-negative) loss observables $\{L(S, W)\} \subset \mathcal{B}(\mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{hyp}})$. We then define the *expected empirical risk* of the quantum learner $\mathcal{A}$ as the expectation value of the observable $L(S, W)$ on the hypothesis state $\sigma(S, W)$, further averaged over $P^{\mathcal{A}}$. That is,

$$\hat{R}_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}} \left[\mathrm{Tr}[L(S, W)\sigma^{\mathcal{A}}(S, W)] \right] . \tag{6}$$

In contrast, we define the *expected true risk* of $\mathcal{A}$ as

$$R_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{(\bar{S},\bar{W})\sim P^{\mathcal{A}}_{\mathsf{Z}^m} \otimes P^{\mathcal{A}}_{\mathsf{W}}} \left[\mathrm{Tr}\left[L(\bar{S}, \bar{W}) \left(\rho_{\mathrm{test}}(\bar{S}) \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}(\bar{S}, \bar{W}) \right) \right] \right] , \tag{7}$$

where we have "decoupled" the quantum test and training data systems before letting the learner act, and we have also decoupled the classical training data and hypothesis random variables. Here, a state with a subscript denotes a reduced density matrix obtained by tracing out the other subsystems. Mathematically, this is achieved by a partial trace, for example, we have $\rho_{\mathrm{test}}(\bar{S}) = \mathrm{Tr}_{\mathrm{train}}[\rho(\bar{S})]$ and $\sigma^{\mathcal{A}}_{\mathrm{hyp}}(\bar{S}, \bar{W}) = \mathrm{Tr}_{\mathrm{test}}[\sigma^{\mathcal{A}}(\bar{S}, \bar{W})]$. Finally, we define the expected generalization error as the difference between expected true and empirical risks,

$$\mathrm{gen}_\rho(\mathcal{A}) = R_\rho(\mathcal{A}) - \hat{R}_\rho(\mathcal{A}). \tag{8}$$

[1] Concentration bounds for the generalization error are also often of interest, but we primarily consider bounds in expectation in this article.

Our main goal is to bound $\mathrm{gen}_\rho(\mathcal{A})$ in terms of properties of the CQ data ρ, the loss observables $L(S, W)$, and the learner $\mathcal{A}$.

One may consider alternative notions of decoupling and alternative definitions for the quantum risks. These notions may also be tailored differently to capture the essence of the learning task at hand. In the next paragraph, we motivate our decoupling approach to the definition of true risk and generalization error by a comparison to the classical framework, and demonstrate how it extends established notions from classical learning theory. In addition to reducing to the expected empirical and true risk in the classical case, these choices give rise to natural notions of risks and generalization error for a variety of quantum learning tasks (see Appendix C). Moreover, our definitions account for the desiderata that $\hat{R}_\rho(\mathcal{A})$ should incorporate all aspects in which the learner's actions "contaminate" the test data, whereas the test data in $R_\rho(\mathcal{A})$, both classical and quantum, must be completely untarnished by the learner. This justifies Equations (7) and (8) as the quantum extension of (Xu and Raginsky, 2017)'s change-of-measure/decoupling perspective on classical generalization analysis.

Classical → quantum: motivating our framework and bounds. Before formally stating our bounds on the generalization error in quantum learning, it is natural to wonder how our framework for describing quantum learners and their generalization behavior compare with existing work. Here, we provide such a comparison to information-theoretic generalization bounds in classical learning theory. We begin with (Xu and Raginsky, 2017) and arrive at our quantum framework via an intermediate (classical) extension, which is reminiscent of information-theoretic approaches to out-of-distribution generalization (Hellström et al., 2023, Section 9.2)

First, we recall the main result of (Xu and Raginsky, 2017): Assuming that for $Z_i \sim P$ the random variable $\ell(w, Z_i)$ is β-sub-gaussian for all $w \in \mathsf{W}$ – a special case of our assumption (locCMGF) below – (Xu and Raginsky, 2017, Theorem 1) proved that the classical expected generalization error defined in Equation (5) is bounded as

$$|\mathrm{gen}_P(\mathcal{A})| \leq \sqrt{\frac{2\beta^2}{m} I(S; W)}\,, \tag{9}$$

where $I(S; W)$ is the mutual information (MI) between training data and hypothesis random variables. A simple but crucial observation underlying this bound: It amounts to a statement about *decoupling* two random variables. Namely, we can rewrite the expected true risk as $\mathbb{E}_{W \sim P_{\mathsf{W}}^{\mathcal{A}}}[R_P(W)] = \mathbb{E}_{\bar{S} \sim P^m} \mathbb{E}_{\bar{W} \sim P_{\mathsf{W}}^{\mathcal{A}}}[\hat{R}_{\bar{S}}(\bar{W})]$. This has the same form as the expected empirical risk, given by the expression $\mathbb{E}_{(S,W) \sim P^{\mathcal{A}}}[\hat{R}_S(W)]$, but the training data and hypothesis random variables have been replaced by independent copies thereof. Informally speaking, (Xu and Raginsky, 2017, Theorem 1) thus tells us that decoupling training data and hypothesis comes at a cost depending on the mutual information $I(S; W)$.

Next, as an intermediate step towards our quantum framework, we introduce a variant of this result by adding test data to the classical learning-theoretic framework discussed above. Concretely, suppose we have test data $S_{\mathrm{te}} = (Z_{\mathrm{te},i})_{i=1}^m$ and training data $S_{\mathrm{tr}} = (Z_{\mathrm{tr},i})_{i=1}^m$, where the pairs $(Z_{\mathrm{te},i}, Z_{\mathrm{tr},i})$ are drawn i.i.d. from some probability distribution P over $\mathsf{Z} \times \mathsf{Z}$. Note that while different pairs are independent, the two random variables $Z_{\mathrm{te},i}, Z_{\mathrm{tr},i}$ within any single pair may not be. During training, a learner $\mathcal{A}$ has access to S_{tr} but not to S_{te}, so its output behaviour may still be described by a conditional distribution $P(W|S_{\mathrm{tr}})$. However, the relevant performance measures are now taken w.r.t. *test* instead of training data. That is, we now consider the expected empirical *testing* risk $\mathbb{E}_{S_{\mathrm{te}},S_{\mathrm{tr}},W}[\hat{R}_{S_{\mathrm{te}}}(W)]$ and the expected true *testing* risk $\mathbb{E}_W[R_{P_{\mathrm{te}}}(W)]$, where P_{te} denotes the marginal of P on the first subsystem. Two extreme examples illustrate the utility of this setup: First, if $Z_{\mathrm{te},i}$ and $Z_{\mathrm{tr},i}$ are perfectly correlated, we recover exactly the setting considered in (Xu and Raginsky, 2017). In contrast, if $Z_{\mathrm{te},i}$ and $Z_{\mathrm{tr},i}$ are independent and have the same distribution, then the expected generalization error trivially vanishes.

Also in this setting, the expected true risk can be obtained from the expected empirical risk via decoupling as before, starting with the rewriting $\mathbb{E}_W[R_{P_{\mathrm{te}}}(W)] = \mathbb{E}_{\bar{S}_{\mathrm{tr}},\bar{S}_{\mathrm{te}},\bar{W}}[\hat{R}_{\bar{S}_{\mathrm{te}}}(\bar{W})]$. However, W depends on S_{te} only through S_{tr}, so decoupling can now be achieved in different ways: We can decouple W from S_{tr} as before, or we can decouple S_{te} from S_{tr}, or we can (unnecessarily) decouple both pairs simultaneously. More rigorously, using (Xu and Raginsky, 2017, Lemma 1), we can show that if $\ell(w, Z_{\mathrm{te},i})$, with $Z_{\mathrm{te},i} \sim P_{\mathrm{te}}$, is β-sub-gaussian, then the expected generalization error satisfies

$$\left|\mathop{\mathbb{E}}_{W}[R_{P_{\mathrm{te}}}(W)] - \mathop{\mathbb{E}}_{S_{\mathrm{te}},S_{\mathrm{tr}},W}[\hat{R}_{S_{\mathrm{te}}}(W)]\right| \leq \sqrt{\frac{2\beta^2}{m} I(S_{\mathrm{te}};W)} \leq \sqrt{\frac{2\beta^2}{m}\min\{I(S_{\mathrm{tr}};W), I(S_{\mathrm{tr}};S_{\mathrm{te}})\}}\,, \tag{10}$$

where the last inequality follows from the data processing inequality and the chain rule.

We can now describe the final step towards our quantum framework. To do so, we return to the setting of (Xu and Raginsky, 2017) on the classical side, assuming only training data but no test data. This is for simplicity of presentation, our bounds can be extended to the case with classical training and test data. On the quantum side, however, we assume both a test and a training data system, which may share classical correlations or quantum entanglement. Thus, going from the expected empirical risk $\hat{R}_\rho(\mathcal{A})$ to the expected true risk $R_\rho(\mathcal{A})$ now requires two decoupling steps, the first quantum – from a general bipartite state $\sigma^{\mathcal{A}}(S,W)$ to a tensor product state $\tau^{\mathcal{A}}(S,W)$ by decoupling the quantum test and train systems before the action of the learner – and the second classical – going from correlated random variables S, W to independent copies $\bar{S}, \bar{W}$. Our generalization bounds below show that the first decoupling step contributes an expected *quantum* mutual information (QMI) plus Holevo information and the second a classical MI. Notably, whereas a single decoupling step was sufficient in the case of classical test data, our classical-quantum decoupling indeed consists of two non-trivial decoupling steps.

1.1.2. GENERALIZATION ERROR BOUNDS

Assumptions. The framework and formalism described above can capture a variety of learning scenarios. In order to prove bounds on the generalization error, we will assume mild properties to be satisfied by the learner, the data, and the loss observables. To avoid clutter, in the following we suppress dependencies on S, W in the notation where it is clear from the context. That is, we write σ instead of $\sigma(S,W)$ and L instead of $L(S,W)$. Additionally, we will frequently denote $\tau^{\mathcal{A}} = \tau^{\mathcal{A}}(s,w) = \rho_{\mathrm{test}}(s) \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w) = \rho_{\mathrm{test}} \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}$. Here and throughout, a state with a subsystem subscript denotes the reduced state on that subsystem.

As in many classical works on this subject, bounds on the moment generating function (MGF) allow for characterizing the concentration properties of the value of the loss observable around its expectation value. However, due to the noncommutative setting at hand, we consider the following two generalizations:

(QMGF) **Quantum MGF/tail bound**: For every $(s,w) \in \mathsf{Z}^m \times \mathsf{W}$, let the logarithmic quantum moment generating function of the loss observable L with respect to the product state $\tau^{\mathcal{A}}$ be bounded by convex functions $\psi_+, \psi_- : \mathbb{R} \to \mathbb{R}$ which satisfy $\psi_\pm(0) = \psi'_\pm(0) = 0$, i.e.

$$\log \operatorname{Tr}\left[\tau^{\mathcal{A}} e^{\lambda\left(L - \operatorname{Tr}[L\tau^{\mathcal{A}}]\mathbb{1}\right)}\right] \leq \begin{cases} \psi_+(\lambda) & \text{if } \lambda \geq 0 \\ \psi_-(\lambda) & \text{if } \lambda < 0 \end{cases} . \tag{QMGF}$$

(CMGF) **Classical MGF/tail bound:** For every $w \in \mathsf{W}$, let the logarithmic moment generating function of the expectation value $\operatorname{Tr}[L\tau^{\mathcal{A}}]$ of the loss observable L in the product state $\tau^{\mathcal{A}}$, viewed as a random variable, be bounded by convex functions $\phi_+, \phi_- : \mathbb{R} \to \mathbb{R}$ which satisfy $\phi_\pm(0) = \phi'_\pm(0) = 0$, i.e.,

$$\log \mathop{\mathbb{E}}_{S\sim P^m}\left[e^{\lambda\left(\operatorname{Tr}[L\tau^{\mathcal{A}}] - \mathbb{E}_{S\sim P^m}\left[\operatorname{Tr}[L\tau^{\mathcal{A}}]\right]\right)}\right] \leq \begin{cases} \phi_+(\lambda) & \text{if } \lambda \geq 0 \\ \phi_-(\lambda) & \text{if } \lambda < 0 \end{cases} . \tag{CMGF}$$

If the convex functions $\psi_\pm$ and $\phi_\pm$ are of the form $\lambda \mapsto \frac{\alpha^2\lambda^2}{2}$ and $\lambda \mapsto \frac{\beta^2\lambda^2}{2}$, respectively, then we speak of an α-sub-gaussian QMGF and a β-sub-gaussian CMGF. We describe some scenarios of interest in which these assumptions are satisfied in Section 1.1.3.

Generalization bounds. Can the generalization error of the quantum learner $\mathcal{A}$ on the data ρ be controlled in terms of quantities that we can interpret, giving us a handle on how one can produce a hypothesis that attains a balance between fitting the training data and performing well on unseen data? We answer this question in the affirmative, and show that assuming classical and quantum MGF bounds allows us to control the generalization error via quantities measuring the classical and quantum information shared between data and hypothesis.

Our first main result is the following generalization bound for quantum learners.

Theorem 1 (Informally stated; see Theorem 17) *If the classical-quantum data state ρ and the loss observable satisfy* (QMGF) *and* (CMGF)*, then the expected generalization error of $\mathcal{A}$ satisfies*

$$\begin{aligned}\pm\, \mathrm{gen}_\rho(\mathcal{A}) \leq \psi_\mp^{*-1}&\left(\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}}\right] + \mathop{\mathbb{E}}_{S\sim P^m}\left[\chi\left(\{P^{\mathcal{A}}_{\mathsf{W}|S}(w), \rho^{\mathcal{A}}_{\mathrm{test}}(S,w)\}_w\right)\right]\right)\\ &+ \phi_\mp^{*-1}(I(S;W))\,,\end{aligned} \tag{11}$$

where $\psi_\mp^{*-1}$ and $\phi_\mp^{*-1}$ denote the inverses of the Legendre transforms of $\psi_\mp$ and $\phi_\mp$.

In Equation (11), the following quantities from classical and quantum information theory appear (see Appendix A for formal definitions): $I(S;W)$ is the mutual information (MI) between the training data S and the hypothesis W. $I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}} = I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}(S,W)}$ denotes the quantum mutual information (QMI) between test and hypothesis systems in the output state $\sigma^{\mathcal{A}}(S,W)$ produced by the learner. Finally, $\chi(\{P(x),\rho(x)\}_{x\in\mathsf{X}})$ denotes the Holevo information of an ensemble of states, which is connected to how much information about $x \sim P$ can be extracted from $\rho(x)$. It is given by $H(\mathbb{E}_{x\sim P}[\rho(x)]) - \mathbb{E}_{x\sim P}[H(\rho(x))]$, the difference between the (von Neumann) entropy of the average state and the expected (von Neumann) entropy of a state drawn from the ensemble.

Theorem 1 provides a theoretical guideline for designing a learner $\mathcal{A}$. Namely, we expect better generalization performance for learners whose measurements and post-processing do not induce strong correlations between the available data set and the output hypothesis. Naturally, we inherit a caveat from classical learning theory: Learning typically requires both good performance on the training data and good generalization. Thus, our bounds provide an information-theoretic perspective on the bias-variance trade-off in quantum learning. On the one hand, for good training performance, a learner has to extract information about the underlying concept from the available classical-quantum data. On the other hand, for good generalization, the amount of extracted/accessible classical and quantum information should be limited.

In the sub-gaussian case, the inverse Legendre transforms can be computed explicitly and the generalization error bound takes an appealingly simple form.

Corollary 2 (Informally stated; see Corollary 23) *If the classical-quantum data and the loss observable satisfy an α-sub-gaussian (QMGF) and a β-sub-gaussian (CMGF) condition, then*

$$|\mathrm{gen}_\rho(\mathcal{A})| \leq \sqrt{2\alpha^2 \left(\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}} \left[I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}} \right] + \mathop{\mathbb{E}}_{S\sim P^m} \left[\chi\left(\{ P^{\mathcal{A}}_{\mathsf{W}|S}(w), \rho^{\mathcal{A}}_{\mathrm{test}}(S,w) \}_w \right) \right] \right)} + \sqrt{2\beta^2 I(S;W)}\,. \tag{12}$$

We have already assumed that $S = (Z_i)_{i=1}^m$ consists of i.i.d. examples. Now, we additionally assume that $\rho(S) = \bigotimes_{i=1}^m \rho_i(Z_i)$ is a tensor product of quantum data states, and that the measurements and channels performed by the learner $\mathcal{A}$ also factorize. (In fact, if $\mathcal{A}$ produces only a classical but no quantum hypothesis, we can drop this factorization requirement on $\mathcal{A}$.) Then, our states after the action of the learner also factorize as $\sigma^{\mathcal{A}} = \bigotimes_{i=1}^m \sigma_i^{\mathcal{A}}$ and $\tau^{\mathcal{A}} = \bigotimes_{i=1}^m \tau_i^{\mathcal{A}}$, with $\sigma_i^{\mathcal{A}} = \sigma_i^{\mathcal{A}}(z_i, w)$ and $\tau_i^{\mathcal{A}} = \tau_i^{\mathcal{A}}(z_i, w)$. Finally, we assume that the loss observable is an average of local losses, $L = \frac{1}{m}\sum_{i=1}^m L_i$, with $L_i = L_i(z_i, w)$ acting non-trivially only on the i^{th} tensor factor. The natural analogues of (QMGF) and (CMGF) become:

(locQMGF) **Local QMGF:** For every $i \in [m]$, for every (z_i, w), each local L_i satisfies (QMGF) w.r.t. $\tau_i^{\mathcal{A}}$ with bound $\psi_{\pm,i}$.

(locCMGF) **Local CMGF:** For every $i \in [m]$, for every w, each local $\mathrm{Tr}[L_i \tau_i^{\mathcal{A}}]$ satisfies (CMGF) w.r.t. P with bound $\phi_{\pm,i}$.

In addition to serving as natural quantum counterparts of common assumptions used to derive classical information-theoretic generalization bounds, we identify several scenarios in which (locQMGF) and (locCMGF) are satisfied in Appendix B.2. As before, we speak of α_i-sub-gaussian (locQMGF) and β_i-sub-gaussian (locCMGF) if the convex functions $\psi_{\pm,i}$ and $\phi_{\pm,i}$ are of the form $\lambda \mapsto \frac{\alpha_i^2\lambda^2}{2}$ and $\lambda \mapsto \frac{\beta_i^2\lambda^2}{2}$, respectively. If the sub-gaussianity parameters are the same for all i, that is, if $\alpha_i = \alpha$ and $\beta_i = \beta$ for all i, then we simply speak of α-sub-gaussian (locQMGF) and β-sub-gaussian (locCMGF). In this scenario, Corollary 2 becomes:

Corollary 3 (Informally stated; see Corollary 24) *If the classical-quantum data and the loss observable satisfy an α-sub-gaussian (locQMGF) and a β-sub-gaussian (locCMGF) condition, then*

$$|\mathrm{gen}_\rho(\mathcal{A})| \leq \sqrt{\frac{2\alpha^2}{m}\left(\underset{(S,W)\sim P^{\mathcal{A}}}{\mathbb{E}}\left[I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}}\right] + \underset{S\sim P^m}{\mathbb{E}}\left[\chi\left(\{P^{\mathcal{A}}_{\mathsf{W}|S}(w), \rho^{\mathcal{A}}_{\mathrm{test}}(S,w)\}_w\right)\right]\right)} + \sqrt{\frac{2\beta^2}{m} I(S;W)}\,. \tag{13}$$

Corollary 3 tells us: We can control the expected generalization error by choosing the training data size m to be on the order of the maximum between the classical and quantum information shared between the data and the learner's hypothesis. Conversely, if only a limited amount of data is available, then to guarantee good generalization, we have to limit the classical and quantum information that the learner accumulates about the data accordingly. As we explain in Appendix B, Corollary 3 can be extended to the case of different local loss observables, which also have different sub-gaussianity parameters α_i and β_i (see Corollary 24), and to stable learners employing channels that approximately preserve locality (see Corollary 25).

1.1.3. APPLICATIONS

Our framework and generalization bounds capture a variety of settings. Therefore, we envision that our approach can lead to new insights by providing a novel perspective on diverse quantum learning problems. Here, we highlight only three applications, but to fundamental problems in quantum learning. For further examples in quantum parameter estimation, variational quantum machine learning, approximate quantum membership problems, learning quantum state preparation procedures, quantum differential privacy, and inductive quantum learning see Appendix C.

PAC learning quantum states. (Aaronson, 2007) pioneered the use of learning-theoretic perspectives for quantum information problems. The seminal contribution of this work was to formulate "pretty good state tomography" in a PAC learning sense and to analyze its sample complexity. Here, instead of aiming for a (classically described) approximation to an unknown quantum state in trace distance, one considers the relaxed task of producing a (classically described) hypothesis state that accurately approximates the expectation value on a test measurement drawn from an underlying data distribution, with high success probability. While full state tomography requires resources scaling exponentially with the number n of qubits (O'Donnell and Wright, 2016; Haah et al., 2017), this PAC relaxation has sample complexity scaling linearly in n (Aaronson, 2007).

In Appendix C.1, we use Corollary 3 to reproduce this fundamental insight into learning quantum states within our framework of in-expectation learning. Concretely, we give a simple learning strategy achieving an in-expectation version of (Aaronson, 2007, Theorem 1.1) with the same dependence on the Hilbert space dimension d and the approximation accuracy ε. Our formulation allows us to naturally describe an end-to-end learning strategy that starts from (possibly entangled) copies of the unknown quantum state. As part of our derivation, we extend an argument due to (Xu and Raginsky, 2017) to prove that information-theoretic generalization guarantees reproduce classical in-expectation excess risk bounds for regression based on the fat-shattering dimension (Kearns and Schapire, 1994; Bartlett and Long, 1998; Anthony and Bartlett, 2000).

We highlight that our in-expectation guarantees show that for each observable seen during training, a number of copies independent of d is sufficient to achieve overall reliable expectation value estimates. In essence, there are distinct classical ("how many observables") and quantum ("how many copies of ρ per observable") aspects to the sample complexity. Only the first is d-dependent. Our perspective thus provides a natural intermediary between the "measure $\log\log(d)$ many times" setting of (Aaronson, 2007, Objection 6) and the "measure once" scenario of (Aaronson, 2007, Theorem 1.3). This illustrates how studying in-expectation bounds can complement studying the concentration properties of the generalization error.

Quantum PAC learning from entangled data. A central question in quantum learning theory (Arunachalam and de Wolf, 2017) is whether and when quantum access to data allows one to learn an unknown classical object (typically a function) more sample- and/or computationally efficiently than is possible purely classically. A prominent way of modeling quantum (access to) data is via *superposition examples* (Bshouty and Jackson, 1998), which then admit questions of PAC learning from quantum oracle access.

We propose a variant of quantum superposition examples: Viewing a single classical training example as a mixed state $\rho = \sum_z P(z) |z\rangle\langle z|$ diagonal in the computational basis, we take a *purification* and consider the resulting entangled state $|\phi\rangle = \sum_z \sqrt{P(z)}\, |z\rangle_{\mathrm{test}} \otimes |z\rangle_{\mathrm{train}}$ as describing the joint system of a single quantum test and training example. Multiple copies of this bipartite state then form the overall data. The entanglement between test and training data is an inherently quantum analogue to a classical scenario with perfectly correlated test and training data.

For this notion of quantum data access, we study learners that perform simple measurements followed by classical post-processing. We show how to analyze the generalization performance of such learners purely quantumly by describing the measurement and post-processing jointly by a quantum channel acting on the training data. In particular, we demonstrate that Corollary 3 in this case reproduces the main result of (Xu and Raginsky, 2017). Notably, it does so *via the QMI contribution* in the upper bound, which highlights the relevance and necessity of this term.

Quantum state discrimination and classification. Distinguishing between different candidate states when given copies of an unknown quantum state is a fundamental task in quantum information science (Bae and Kwek, 2015). The optimal measurement for binary state discrimination, the case of two candidates, is well understood (Helstrom, 1969; Holevo, 1973). For distinguishing between multiple states, necessary and sufficient optimality criteria are known (Holevo, 1974; Yuen et al., 1975), but in general do not give rise to an explicit construction for the optimal POVM. Only in certain symmetric cases can the optimal measurement be made explicit (Ban et al., 1997; Eldar and Forney, 2001; Eldar et al., 2004), often via the pretty good (or square root) measurement (Hausladen and Wootters, 1994; Hausladen et al., 1996). These results, however, presuppose that classical descriptions for the possible candidate states are known in advance.

More recently, distinguishing between two a priori unknown quantum states was considered as a classification problem inspired by machine learning approaches to pattern recognition (Guţă and Kotłowski, 2010; Sentís et al., 2019; Rosati, 2022). Here, the goal is to learn a distinguishing POVM from (labelled) copies of the unknown states. We formulate a PAC version of quantum state classification (see Appendix B). Then, our information-theoretic generalization guarantees yield bounds on the sample size sufficient to ensure that a learned POVM, which performs well on available training data, will also successfully classify previously unseen state pairs in-expectation over an underlying distribution over pairs. These may serve as a guiding principle for avoiding overfitting in quantum state classification. In particular, our results imply that limiting the complexity of the admissible hypothesis POVMs and thus the maximum information content of a hypothesis, for instance by imposing locality restrictions, will favorably affect the required amount of quantum data.

1.2. Discussion and outlook

In this work, we have established a mathematical framework for reasoning about tasks of learning from data that is part classical and part quantum. In addition to proving generalization error bounds for quantum learners in such scenarios, we have also demonstrated a variety of applications that our framework encompasses. Importantly, our bounds are information-theoretic in nature. Thus, they come with an intuitive interpretation and provide a perspective on quantum learning that can benefit from insights in quantum information theory.

The average-case and in-expectation generalization bounds give an insightful perspective that is complementary to worst-case analyses, which have thus far been more widespread in the literature on quantum learning. The former illuminates certain features that are not apparent in the latter, raising the question of re-examining established results in a new or different light. We hope our work motivates future work on quantum learning to also consider in-expectation generalization alongside worst-case behavior.

With part of our contribution being the formulation of a novel framework, our work raises many interesting follow-up questions. In the following, we highlight some of them.

Average-case vs. worst-case. As is typical in PAC learning, our results address the average performance on instances drawn from an (unknown) underlying distribution. For instance, our risk bounds for "pretty good tomography" (Aaronson, 2007) hold w.r.t. a distribution over 2-outcome POVMs. In contrast, recent progress in shadow tomography (Aaronson, 2019; Bădescu and O'Donnell, 2021; Huang et al., 2021) and classical shadows (Huang et al., 2020; Elben et al., 2022) has focused on making correct predictions in the worst-case simultaneously over many observables. Moreover, recent work (Huang et al., 2021) has drawn attention to the notable contrast between the average-case and the worst-case when it comes to the potential for a quantum advantage in learning. Extending our information-theoretic perspective on quantum learning to these worst-case scenarios could give us novel ways of probing this frontier.

Open Problem 1 *Establish a quantum information-theoretic characterization of the performance of learners for shadow tomography.*

Quantum-quantum learners. A recent spate of results (Aharonov et al., 2022; Chen et al., 2022a,b; Huang et al., 2022; Caro, 2022a; Huang et al., 2023b; Dutkiewicz et al., 2023) has emphasized the role of quantum-enhanced experiments for learning quantum channels. In particular, the ability to coherently and sequentially query the unknown channel on input states of our choice is an example of quantum enhancement. Can our framework be further developed to incorporate learning from such *query access* to a quantum-to-quantum channel?

Open Problem 2 *Establish a quantum information-theoretic characterization of the performance in learning quantum-to-quantum channels in a query input model.*

Optimality and technical improvements. One might raise the question of whether information-theoretic bounds on the expected generalization error are tight. This is already a non-trivial open question in the classical setting. In the quantum world, the problem of state discrimination is very well understood information-theoretically. We speculate that a notion of average-case state discrimination may be an approach towards understanding the optimality of our bounds.

Finally, (Xu and Raginsky, 2017; Russo and Zou, 2019) have led to a series of follow-up works, including techniques to tighten information-theoretic generalization bounds (Asadi et al., 2018; Bu et al., 2019), improvements relying on (evaluated) sample-wise and/or conditional mutual information (Steinke and Zakynthinou, 2020; Haghifam et al., 2020; Hellström and Durisi, 2021; Harutyunyan et al., 2021; Hellström and Durisi, 2022; Chu and Raginsky, 2023; Hellström et al., 2023), and connections to optimal transport (Esposito and Gastpar, 2022) and convex analysis (Lugosi and Neu, 2022, 2023). These results may inspire improvements to our quantum generalization bounds and potentially connections to quantum optimal transport (De Palma and Trevisan, 2021; De Palma and Rouzé, 2022; De Palma and Trevisan, 2023).

In spite of the rich structure and wealth of open problems in this area of research, simply *translating* these ideas to quantum learning is fraught with pitfalls: for example, there is no unique quantum analogue to the classical notion of conditioning. Breakthrough progress in our quantum information-theoretic understanding of learning will require proving *genuinely quantum* statements which may not have classical analogues.

2. Technical overview

In this section we give an outline of the ideas involved in the development of our framework, and a taste of the techniques that we use in proving our generalization bounds. The proofs of classical information-theoretic generalization bounds, starting from assumptions analogous to Equation (CMGF), typically proceed as follows[2]: First, the mutual information between data and hypothesis can be expressed as the expected relative entropy between the the distribution of the data conditioned on the hypothesis and the unconditioned distribution of the data, $I(S;W) = \mathbb{E}_{W\sim P^{\mathcal{A}}_{\mathsf{W}}}[D(P^{\mathcal{A}}_{\mathsf{Z}^m|W}\|P^{\mathcal{A}}_{\mathsf{Z}^m})]$. Next, the relative entropies are rewritten via the Donsker-Varadhan representation of the relative entropy (see, for example, (Boucheron et al., 2013, Corollary 4.15)), which in this case in particular implies

$$D(P^{\mathcal{A}}_{\mathsf{Z}^m|W}\|P^{\mathcal{A}}_{\mathsf{Z}^m}) \geq \mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathsf{Z}^m|W}}[\lambda f(W,S)] - \log \mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathsf{Z}^m}}\left[e^{\lambda f(W,S)}\right] \quad \forall \lambda\in\mathbb{R}\,, \tag{14}$$

for $f(W,S) = \frac{1}{m}\sum_{i=1}^m \ell(W,Z_i)$. The second term is controlled based on assumptions on the logarithmic MGF, which in particular introduces a term $-\mathbb{E}_{S\sim P^{\mathcal{A}}_{\mathsf{Z}^m}}[\lambda f(W,S)] = -\lambda\,\mathbb{E}_{S\sim P^{\mathcal{A}}_{\mathsf{Z}^m}}[\hat{R}_S(W)]$. After an optimization over λ, one can rearrange and average over the hypothesis to obtain an information-theoretic generalization bound.

An obstacle to extending this argument to our setting with classical-quantum data is the lack of a quantum analogue of conditioning the data on the hypothesis. To circumvent this obstacle, we decompose the generalization error into a classical and a quantum part. As highlighted in the previous subsection, this decomposition is a feature inherent to our classical-quantum setup: Even

[2] Starting from different sub-gaussianity assumptions, variations on this reasoning can be successful, see, e.g., (Bu et al., 2019, Proposition 1).

after extending the classical learning framework to include test data, it still admits a generalization bound with a single classical mutual information term, no decomposition into separate terms is needed. In our decomposition, the classical part is a difference of two terms that differ only in whether the underlying classical data and hypothesis random variables are correlated or decoupled. Thus, it can be controlled with the classical proof strategy outlined above. The quantum part takes the form of a classical expectation value of the difference between the quantum expectation values $\mathrm{Tr}[L\sigma]$, of the loss observable on the state σ, and $\mathrm{Tr}[L\tau^{\mathcal{A}}]$, the expectation value on its decoupled counterpart $\tau^{\mathcal{A}} = \rho_{\text{test}} \otimes \sigma^{\mathcal{A}}_{\text{hyp}}$. To control this quantum decoupling, we lift the classical proof strategy to our non-commutative quantum setting, replacing Donsker-Varadhan by a combination of Petz's variational characterization of the relative entropy (Petz, 1988) and the Golden-Thompson inequality. Assuming (QMGF), this yields the quantum relative entropy lower bound

$$D(\sigma^{\mathcal{A}} \| \tau^{\mathcal{A}}) \geq \lambda \left(\mathrm{Tr}[L\sigma^{\mathcal{A}}] - \mathrm{Tr}[L\tau^{\mathcal{A}}] \right) - \begin{cases} \psi_+(\lambda) & \text{if } \lambda \geq 0 \\ \psi_-(\lambda) & \text{if } \lambda < 0 \end{cases} . \tag{15}$$

Now, we can optimize over λ and rearrange to obtain a bound on $\mathrm{Tr}[L\sigma] - \mathrm{Tr}[L\tau^{\mathcal{A}}]$ in terms of $\mathbb{E}_{(S,W)\sim P^{\mathcal{A}}}[D(\sigma^{\mathcal{A}} \| \tau^{\mathcal{A}})]$. After showing that this expected relative entropy equals the expression $\mathbb{E}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\text{test}; \text{hyp})_{\sigma^{\mathcal{A}}}\right] + \mathbb{E}_{S\sim P^m}\left[\chi\left(\{P^{\mathcal{A}}_{\mathsf{W}|S}(w), \rho^{\mathcal{A}}_{\text{test}}(S,w)\}_w\right)\right]$, we combine this quantum decoupling bound with the bound on the classical part to obtain Theorem 1.

The usefulness of classical generalization bounds depends on whether and how quickly they decay as the training data size m increases. Typically, such a decrease is proved under an i.i.d. assumption on the data. To strengthen Theorem 1 for i.i.d. quantum data, adhering to a tensor product structure, we invoke tools from quantum optimal transport. On the one hand, (De Palma and Trevisan, 2023, Theorem 8.1), restated as Lemma 30, shows that Lipschitz observables have subgaussian QMGFs w.r.t. any tensor product state:

$$\mathrm{Tr}\left[\exp\left(\log\left(\bigotimes_{i=1}^{m} \rho_i\right) + \lambda H\right)\right] \leq \exp\left(\frac{\lambda^2 m \|H\|^2_{\mathrm{Lip}}}{2}\right) . \tag{16}$$

While this is weaker than bounds of the form (QMGF) due to Golden-Thompson, we demonstrate that such a QMGF bound is still sufficient for our above proof strategy. This then allows us to improve Theorem 1 achieve a bound that decays with $1/\sqrt{m}$ if both quantum data and learner factorize, assuming local loss observables (Corollary 3). On the other hand, the machinery of quantum Lipschitz constants allows us to go beyond quantum learners that factorize. In particular, it guides us to define a stability criterion for quantum learners in terms of Wasserstein-1 distances, a quantum version of classical replace-one stability (Bousquet and Elisseeff, 2000, 2002; Shalev-Shwartz et al., 2010). Namely, if the underlying classical data sets differ in only few data points, then the associated quantum processing channels employed by a stable learner differ only by a small amount, measured in terms of a Schatten-1–to–Wasserstein-1 norm. Combining our newly established sub-gaussianity of Lipschitz observables w.r.t. tensor products with the classical bounded differences concentration inequality (McDiarmid, 1989), we can then extend our generalization guarantees to stable quantum learners with a controlled increase in Wasserstein-1 distances (Corollary 25).

Acknowledgments

The authors thank Marco Fanizza, Fredrik Hellström, and Matteo Rosati for feedback on an earlier draft. Moreover, the authors thank Mark M. Wilde for pointing out the connection to measured relative entropy discussed in Remark 20 and Giacomo de Palma for pointing out the reference (De Palma and Trevisan, 2023). Finally, the authors thank the anonymous COLT 2024 reviewers for valuable feedback that in particular helped improve the presentation. MCC was supported by a DAAD PRIME fellowship. The Institute for Quantum Information and Matter is an NSF Physics Frontiers Center. TG is supported by UKRI Future Leaders Fellowship MR/S031545/1, EPSRC New Horizons Grant EP/X018180/1, and EPSRC Robust and Reliable Quantum Computing Grant EP/W032635/1. DSF is supported by France 2030 under the French National Research Agency award number "ANR-22-PNCQ-0002". CR acknowledges financial support from the ANR project QTraj (ANR-20-CE40-0024-01) of the French National Research Agency (ANR), as well as from the Humboldt Foundation. SS is supported by a Royal Commission for the Exhibition of 1851 Research Fellowship.

References

Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, 2007. doi: 10.1098/rspa.2007.0113.

Scott Aaronson. Shadow tomography of quantum states. *SIAM Journal on Computing*, 49(5): STOC18–368, 2019. doi: 10.1137/18M120275X.

Scott Aaronson and Guy N Rothblum. Gentle measurement of quantum states and differential privacy. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 322–333, 2019. doi: 10.1145/3313276.3316378.

Amira Abbas, David Sutter, Alessio Figalli, and Stefan Woerner. Effective dimension of machine learning models. *arXiv preprint arXiv:2112.04807*, 2021a. URL https://arxiv.org/abs/2112.04807.

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, 2021b. doi: 10.1038/s43588-021-00084-1.

Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. Quantum algorithmic measurement. *Nature Communications*, 13(1):1–9, 2022. doi: 10.1038/s41467-021-27922-0. URL https://www.nature.com/articles/s41467-021-27922-0.

Armando Angrisani and Elham Kashefi. Quantum local differential privacy and quantum statistical query model. *arXiv preprint arXiv:2203.03591*, 2022. URL https://arxiv.org/abs/2203.03591.

Armando Angrisani, Mina Doosti, and Elham Kashefi. A unifying framework for differentially private quantum algorithms. *arXiv preprint arXiv:2307.04733*, 2023. URL https://arxiv.org/abs/2307.04733.

Anurag Anshu. Concentration bounds for quantum states with finite correlation length on quantum spin lattice systems. *New Journal of Physics*, 18(8):083011, 2016. doi: 10.1088/1367-2630/18/8/083011.

Anurag Anshu and Tony Metger. Concentration bounds for quantum states and limitations on the QAOA from polynomial approximations. *Quantum*, 7:999, May 2023. ISSN 2521-327X. doi: 10.22331/q-2023-05-11-999.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Martin Anthony and Peter L. Bartlett. Function learning from interpolation. *Combinatorics, Probability and Computing*, 9(3):213–225, 2000. doi: 10.1017/S0963548300004247.

Srinivasan Arunachalam and Ronald de Wolf. Guest column: A survey of quantum learning theory. *SIGACT News*, 48, 2017. doi: 10.1145/3106700.3106710.

Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7245–7254, Red Hook, NY, USA, 2018. Curran Associates Inc. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8d7628dd7a710c8638dbd22d4421ee46-Paper.pdf.

Shahab Asoodeh and Huanyu Zhang. Contraction of locally differentially private mechanisms. *arXiv preprint arXiv:2210.13386*, 2022. URL https://arxiv.org/abs/2210.13386.

Costin Bădescu and Ryan O'Donnell. Improved quantum data analysis. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1398–1411, 2021. doi: 10.1145/3406325.3451109.

Joonwoo Bae and Leong-Chuan Kwek. Quantum state discrimination and its applications. *Journal of Physics A: Mathematical and Theoretical*, 48(8):083001, 2015. doi: 10.1088/1751-8113/48/8/083001.

Masashi Ban, Keiko Kurokawa, Rei Momose, and Osamu Hirota. Optimum measurements for discrimination among symmetric quantum states and parameter estimation. *International Journal of Theoretical Physics*, 36:1269–1288, 1997. doi: 10.1007/BF02435921.

Leonardo Banchi, Jason Pereira, and Stefano Pirandola. Generalization in quantum machine learning: A quantum information standpoint. *PRX Quantum*, 2(4):040321, 2021. doi: 10.1103/PRXQuantum.2.040321.

Leonardo Banchi, Jason Luke Pereira, Sharu Theresa Jose, and Osvaldo Simeone. Statistical complexity of quantum learning. *Advanced Quantum Technologies*, page 2300311, 2024. doi: 10.1002/qute.202300311.

Ivan Bardet, Ángela Capel, Angelo Lucia, David Pérez-García, and Cambyse Rouzé. On the modified logarithmic sobolev inequality for the heat-bath dynamics for 1d systems. *Journal of Mathematical Physics*, 62(6):061901, June 2021. doi: 10.1063/1.5142186.

Peter L Bartlett and Philip M Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998. doi: 10.1006/jcss.1997.1557.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL https://jmlr.org/papers/v3/bartlett02a.html.

Mario Berta, Omar Fawzi, and Marco Tomamichel. On variational expressions for quantum relative entropies. *Letters in Mathematical Physics*, 107:2239–2265, 2017. doi: 10.1007/s11005-017-0990-7.

Rajendra Bhatia. *Matrix analysis*. Springer Science & Business Media, 1997. doi: 10.1007/978-1-4612-0653-8.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017. doi: 10.1038/nature23474. URL https://www.nature.com/articles/nature23474.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002. URL https://jmlr.org/papers/v2/bousquet02a.html.

Nader H. Bshouty and Jeffrey C. Jackson. Learning DNF over the uniform distribution using a quantum example oracle. *SIAM Journal on Computing*, 28(3):1136–1153, 1998. ISSN 0097-5397. doi: 10.1137/S0097539795293123.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591, 2019. doi: 10.1109/ISIT.2019.8849590.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020. doi: 10.1109/JSAIT.2020.2991139.

Zhenyu Cai. Quantum error mitigation using symmetry expansion. *Quantum*, 5:548, 2021. doi: 10.22331/q-2021-09-21-548.

Ángela Capel, Cambyse Rouzé, and Daniel Stilck França. The modified logarithmic sobolev inequality for quantum spin systems: classical and commuting nearest neighbour interactions. *arXiv preprint arXiv:2009.11817*, 2020. URL https://arxiv.org/abs/2009.11817.

Matthias C. Caro. Binary classification with classical instances and quantum labels. *Quantum Machine Intelligence*, 3:18, 2021. doi: 10.1007/s42484-021-00043-z.

Matthias C Caro. Learning quantum processes and Hamiltonians via the Pauli transfer matrix. *arXiv preprint arXiv:2212.04471*, 2022a. URL https://arxiv.org/abs/2212.04471.

Matthias C. Caro. *Quantum Learning Theory*. Dissertation, Technische Universität München, München, 2022b. URL http://mediatum.ub.tum.de/node?id=1634443.

Matthias C. Caro and Ishaun Datta. Pseudo-dimension of quantum circuits. *Quantum Machine Intelligence*, 2:14, 2020. doi: 10.1007/s42484-020-00027-5.

Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum*, 5:582, 2021. doi: 10.22331/q-2021-11-17-582.

Matthias C Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. Generalization in quantum machine learning from few training data. *Nature Communications*, 13:4919, 2022. doi: 10.1038/s41467-022-32550-3.

Matthias C. Caro, Hsin-Yuan Huang, Nicholas Ezzell, Joe Gibbs, Andrew T. Sornborger, Lukasz Cincio, Patrick J. Coles, and Zoë Holmes. Out-of-distribution generalization for learning quantum dynamics. *Nature Communications*, 14:3751, 2023. doi: 10.1038/s41467-023-39381-w.

Matthias C. Caro, Marcel Hinsche, Marios Ioannou, Alexander Nietner, and Ryan Sweke. Classical Verification of Quantum Learning. In Venkatesan Guruswami, editor, *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*, volume 287 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 24:1–24:23, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-309-6. doi: 10.4230/LIPIcs.ITCS.2024.24. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2024.24.

Chih-Chieh Chen, Masaya Watabe, Kodai Shiba, Masaru Sogabe, Katsuyoshi Sakamoto, and Tomah Sogabe. On the expressibility and overfitting of quantum circuit learning. *ACM Transactions on Quantum Computing*, 2(2):1–24, 2021. doi: 10.1145/3466797.

Kean Chen, Qisheng Wang, Peixun Long, and Mingsheng Ying. Unitarity estimation for quantum channels. *IEEE Transactions on Information Theory*, 69(8):5116–5134, 2023a. doi: 10.1109/TIT.2023.3263645.

Senrui Chen, Sisi Zhou, Alireza Seif, and Liang Jiang. Quantum advantages for Pauli channel estimation. *Physical Review A*, 105(3):032435, 2022a. doi: 10.1103/PhysRevA.105.032435.

Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. In *2021 IEEE 62nd Annual Symposium on Foundations* *of Computer Science (FOCS)*, pages 574–585. IEEE, 2022b. doi: 10.1109/FOCS52979.2021.00063.

Sitan Chen, Brice Huang, Jerry Li, Allen Liu, and Mark Sellke. When does adaptivity help for quantum state learning? In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 391–404. IEEE, 2023b. doi: 10.1109/FOCS57990.2023.00029.

Hao-Chung Cheng, Min-Hsiu Hsieh, and Ping-Cheng Yeh. The learnability of unknown quantum measurements. *Quantum Info. Comput.*, 16(7–8):615–656, May 2016. ISSN 1533-7146. doi: 10.5555/3179466.3179470.

Yifeng Chu and Maxim Raginsky. A unified framework for information-theoretic generalization bounds. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 79260–79278. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fa67d13ba6c73637593bbcc92f6400ff-Paper-Conference.pdf.

Kai-Min Chung and Han-Hsuan Lin. Sample Efficient Algorithms for Learning Quantum Channels in PAC Model and the Approximate State Discrimination Problem. In Min-Hsiu Hsieh, editor, *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)*, volume 197 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 3:1–3:22, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-198-6. doi: 10.4230/LIPIcs.TQC.2021.3.

Giacomo De Palma and Cambyse Rouzé. Quantum concentration inequalities. *Annales Henri Poincaré*, 23(9):3391–3429, 2022. doi: 10.1007/s00023-022-01181-1.

Giacomo De Palma and Dario Trevisan. Quantum optimal transport with quantum channels. In *Annales Henri Poincaré*, volume 22, pages 3199–3234. Springer, 2021. doi: 10.1007/s00023-021-01042-3.

Giacomo De Palma and Dario Trevisan. The wasserstein distance of order 1 for quantum spin systems on infinite lattices. *Annales Henri Poincaré*, 24:4237—-4282, 2023. doi: 10.1007/s00023-023-01340-y.

Giacomo De Palma, Milad Marvian, Dario Trevisan, and Seth Lloyd. The quantum wasserstein distance of order 1. *IEEE Transactions on Information Theory*, 2021. doi: 10.1109/TIT.2021.3076442.

Giacomo De Palma, Milad Marvian, Cambyse Rouzé, and Daniel Stilck França. Limitations of variational quantum algorithms: A quantum optimal transport approach. *PRX Quantum*, 4:010309, 2023. doi: 10.1103/PRXQuantum.4.010309.

Christian L Degen, Friedemann Reinhard, and Paola Cappellaro. Quantum sensing. *Reviews of modern physics*, 89(3):035002, 2017. doi: 10.1103/RevModPhys.89.035002.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Dacheng Tao, and Nana Liu. Quantum noise protects quantum classifiers against adversaries. *Physical Review Research*, 3(2):023153, 2021. doi: 10.1103/PhysRevResearch.3.023153.

Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. Efficient measure for the expressivity of variational quantum algorithms. *Physical Review Letters*, 128(8):080506, 2022. doi: 10.1103/PhysRevLett.128.080506.

Richard M. Dudley. *Uniform central limit theorems*. Cambridge University Press, 1999. doi: 10.1017/CBO9780511665622.

Vedran Dunjko and Hans J Briegel. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81(7):074001, 2018. doi: 10.1088/1361-6633/aab406.

Alicja Dutkiewicz, Thomas E O'Brien, and Thomas Schuster. The advantage of quantum control in many-body Hamiltonian learning. *arXiv preprint arXiv:2304.07172*, 2023. URL https://arxiv.org/abs/2304.07172.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006. doi: 10.1007/11681878_14.

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.

Andreas Elben, Steven T Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. *Nature Review Physics*, 2022. doi: 10.1038/s42254-022-00535-2.

Yonina C Eldar and G David Forney. On quantum detection and the square-root measurement. *IEEE Transactions on Information Theory*, 47(3):858–872, 2001. doi: 10.1109/18.915636.

Yonina C Eldar, Alexandre Megretski, and George C Verghese. Optimal detection of symmetric mixed quantum states. *IEEE Transactions on Information Theory*, 50(6):1198–1207, 2004. doi: 10.1109/TIT.2004.828070.

Amedeo Roberto Esposito and Michael Gastpar. From generalisation error to transportation-cost inequalities and back. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 294–299. IEEE, 2022. doi: 10.1109/ISIT50566.2022.9834354.

Marco Fanizza, Yihui Quek, and Matteo Rosati. Learning quantum processes without input control. *arXiv preprint arXiv:2211.05005*, 2022. URL https://arxiv.org/abs/2211.05005.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 535–544. PMLR, 2018. URL https://proceedings.mlr.press/v75/feldman18a.html.

Christopher A. Fuchs and Jeroen Van De Graaf. Cryptographic distinguishability measures for quantum-mechanical states. *IEEE Transactions on Information Theory*, 45(4):1216–1227, 1999. doi: 10.1109/18.761271.

Elies Gil-Fuster, Jens Eisert, and Carlos Bravo-Prieto. Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1):1–12, 2024. doi: 10.1038/s41467-024-45882-z.

Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum metrology. *Physical review letters*, 96(1):010401, 2006. doi: 10.1103/PhysRevLett.96.010401.

Oded Goldreich. On the average-case complexity of property testing. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 124–135. Springer, 2011. doi: 10.1007/978-3-642-22670-0_15.

Peter Grunwald, Thomas Steinke, and Lydia Zakynthinou. Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2217–2247. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/grunwald21a.html.

Mădălin Guţă and Wojciech Kotłowski. Quantum learning: asymptotically optimal classification of qubit states. *New Journal of Physics*, 12(12):123032, 2010. doi: 10.1088/1367-2630/12/12/123032.

Casper Gyurik, Dyon Vreumingen, van, and Vedran Dunjko. Structural risk minimization for quantum linear classifiers. *Quantum*, 7:893, 2023. ISSN 2521-327X. doi: 10.22331/q-2023-01-13-893.

Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 913–925, 2016. doi: 10.1145/2897518.2897585.

Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63(9):5628–5641, 2017. doi: 10.1109/TIT.2017.2719044.

Jeongwan Haah, Robin Kothari, Ryan O'Donnell, and Ewin Tang. Query-optimal estimation of unitary channels in diamond distance. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 363–390. IEEE, 2023. doi: 10.1109/FOCS57990.2023.00028.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9925–9935. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/712a3c9878efeae8ff06d57432016ceb-Paper.pdf.

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=L_cN8vD0XdT.

Tobias Haug and MS Kim. Generalization with quantum geometry for learning unitaries. *arXiv preprint arXiv:2303.13462*, 2023. URL https://arxiv.org/abs/2303.13462.

Paul Hausladen and William K Wootters. A 'pretty good'measurement for distinguishing quantum states. *Journal of Modern Optics*, 41(12):2385–2390, 1994. doi: 10.1080/09500349414552221.

Paul Hausladen, Richard Jozsa, Benjamin Schumacher, Michael Westmoreland, and William K Wootters. Classical information capacity of a quantum channel. *Physical Review A*, 54(3):1869, 1996. doi: 10.1103/PhysRevA.54.1869.

Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. doi: 10.1038/s41586-019-0980-2. URL https://www.nature.com/articles/s41586-019-0980-2.

Teiko Heinosaari and Mário Ziman. *The Mathematical Language of Quantum Theory*. Cambridge University Press, 2011. doi: 10.1017/cbo9781139031103.

Fredrik Hellström and Giuseppe Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *2021 IEEE International Symposium on Information Theory (ISIT)*, page 952–957. IEEE Press, 2021. doi: 10.1109/ISIT45174.2021.9517731. URL https://doi.org/10.1109/ISIT45174.2021.9517731.

Fredrik Hellström and Giuseppe Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 952–957. IEEE, 2021. doi: 10.1109/ISIT45174.2021.9517731.

Fredrik Hellström and Giuseppe Durisi. A new family of generalization bounds using samplewise evaluated cmi. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10108–10121. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/41b6674c28a9b93ec8d22a53ca25bc3b-Paper-Conference.pdf.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and pac-bayes. *arXiv preprint arXiv:2309.04381*, 2023. URL https://arxiv.org/abs/2309.04381.

Carl W Helstrom. Quantum detection and estimation theory. *Journal of Statistical Physics*, 1: 231–252, 1969. doi: 10.1007/BF01007479.

Christoph Hirche, Cambyse Rouzé, and Daniel Stilck França. Quantum differential privacy: An information theory perspective. *IEEE Transactions on Information Theory*, 2023. doi: 10.1109/TIT.2023.3272904.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

A. S. Holevo. Statistical decision theory for quantum systems. *J. Multivariate Anal.*, 3:337–394, 1973. ISSN 0047-259X. doi: 10.1016/0047-259X(73)90028-6. URL https://doi.org/10.1016/0047-259X(73)90028-6.

Alexander Semenovich Holevo. Remarks on optimal quantum measurements. *Problemy Peredachi Informatsii*, 10(4):51–55, 1974. URL https://www.mathnet.ru/eng/ppi1057.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020. doi: 10.1038/s41567-020-0932-7.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021. doi: 10.1103/PhysRevLett.126.190505.

Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean. Quantum advantage in learning from experiments. *Science*, 376(6598):1182–1186, 2022. doi: 10.1126/science.abn7293. URL https://www.science.org/doi/10.1126/science.abn7293.

Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *Science*, 376(6598):1182–1186, 2022. doi: 10.1126/science.abn7293.

Hsin-Yuan Huang, Sitan Chen, and John Preskill. Learning to predict arbitrary quantum processes. *PRX Quantum*, 4(4):040337, 2023a. doi: 10.1103/PRXQuantum.4.040337.

Hsin-Yuan Huang, Yu Tong, Di Fang, and Yuan Su. Learning many-body hamiltonians with heisenberg-limited scaling. *Physical Review Letters*, 130(20):200403, 2023b. doi: 10.1103/PhysRevLett.130.200403.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479. IEEE, 2017. doi: 10.1109/ISIT.2017.8006774.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994. doi: 10.1016/S0022-0000(05)80062-5.

Tomotaka Kuwahara and Keiji Saito. Gaussian concentration bound and ensemble equivalence in generic quantum many-body systems including long-range interactions. *Annals of Physics*, 421:168278, 2020. doi: 10.1016/j.aop.2020.168278.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability, 1986. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.7589&rep=rep1&type=pdf.

Gabor Lugosi and Gergely Neu. Generalization bounds via convex analysis. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3524–3546. PMLR, 2022. URL https://proceedings.mlr.press/v178/lugosi22a.html.

Gábor Lugosi and Gergely Neu. Online-to-pac conversions: Generalization bounds via regret analysis. *arXiv preprint arXiv:2305.19674*, 2023. URL https://arxiv.org/abs/2305.19674.

The Tien Mai and Pierre Alquier. Pseudo-bayesian quantum tomography with rank-adaptation. *Journal of Statistical Planning and Inference*, 184:62–76, 2017. doi: 10.1016/j.jspi.2016.11.003.

David A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. doi: 10.1023/A:1007618624809.

Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989. doi: 10.1017/CBO9781107359949.008.

Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1):37–55, 2003. doi: 10.1007/s00222-002-0266-3.

Alex Monras, Gael Sentís, and Peter Wittek. Inductive supervised quantum learning. *Physical review letters*, 118(19):190503, 2017. doi: 10.1103/PhysRevLett.118.190503.

Ashley Montanaro and Ronald de Wolf. *A Survey of Quantum Property Testing*. Number 7 in Graduate Surveys. Theory of Computing Library, 2016. doi: 10.4086/toc.gs.2016.007.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/05ae14d7ae387b93370d142d82220f1b-Paper.pdf.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2010. ISBN 9780511976667. doi: 10.1017/CBO9780511976667.

Theshani Nuradha, Ziv Goldfeld, and Mark M Wilde. Quantum pufferfish privacy: a flexible privacy framework for quantum systems. *arXiv preprint arXiv:2306.13054*, 2023. URL https://arxiv.org/abs/2306.13054.

Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 899–912, 2016. doi: 10.1145/2897518.2897544.

Dénes Petz. A variational expression for the relative entropy. *Communications in Mathematical Physics*, 114(2):345–349, 1988. doi: 10.1007/BF01225040.

David Pollard. *Convergence of stochastic processes*. Springer, 1984.

Claudiu Marius Popescu. Learning bounds for quantum circuits in the agnostic setting. *Quantum Information Processing*, 20(9):1–24, 2021. doi: 10.1007/s11128-021-03225-7.

Maxim Raginsky. Information, concentration, and learning, 2019. URL http://iss.bu.edu/bobak/nasit/Maxim_Raginsky_Tutorial_Slides.pdf. 2019 IEEE North American School of Information Theory (NASIT 2019).

Matteo Rosati. A learning theory for quantum photonic processors and beyond. *arXiv preprint arXiv:2209.03075*, 2022. URL https://arxiv.org/abs/2209.03075.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019. doi: 10.1109/TIT.2019.2945779.

Gael Sentís, Alex Monràs, Ramon Muñoz Tapia, John Calsamiglia, and Emilio Bagan. Unsupervised classification of quantum data. *Phys. Rev. X*, 9:041029, 2019. doi: 10.1103/PhysRevX.9.041029.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010. URL http://jmlr.org/papers/v11/shalev-shwartz10a.html.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020. URL https://proceedings.mlr.press/v125/steinke20a.html.

Vladimir N. Vapnik and Alexei Ya. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Mathukumalli Vidyasagar. *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer, 2003.

Qisheng Wang, Zhicheng Zhang, Kean Chen, Ji Guan, Wang Fang, Junyi Liu, and Mingsheng Ying. Quantum algorithm for fidelity estimation. *IEEE Transactions on Information Theory*, 69(1):273–282, 1 2023. doi: 10.1109/tit.2022.3203985.

Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 36111–36137. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wang23w.html.

Mark M Wilde. *Quantum information theory*. Cambridge University Press, 2013. doi: 10.1017/CBO9781139525343.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://dl.acm.org/doi/abs/10.5555/3294996.3295013.

Horace Yuen, Robert Kennedy, and Melvin Lax. Optimum testing of multiple hypotheses in quantum detection theory. *IEEE transactions on information theory*, 21(2):125–134, 1975. doi: 10.1109/TIT.1975.1055351.

Behnoosh Zamanlooy and Shahab Asoodeh. Strong data processing inequalities for locally differentially private mechanisms. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1794–1799. IEEE, 2023. doi: 10.1109/ISIT54713.2023.10206578.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021. ISSN 0001-0782. doi: 10.1145/3446776.

Haimeng Zhao, Laura Lewis, Ishaan Kannan, Yihui Quek, Hsin-Yuan Huang, and Matthias C Caro. Learning quantum states and unitaries of bounded gate complexity. *arXiv preprint arXiv:2310.19882*, 2023. URL https://arxiv.org/abs/2310.19882.

Li Zhou and Mingsheng Ying. Differential privacy in quantum computation. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 249–262. IEEE, 2017. doi: 10.1109/CSF.2017.23.

Appendix A. Preliminaries and Notation

We establish some minimal preliminaries and notation regarding quantum information and computing, and refer the reader to textbooks such as (Wilde, 2013; Nielsen and Chuang, 2010) for details.

We use $\mathcal{H}$ to denote a Hilbert space, and different Hilbert spaces are distinguished by subscripts. We denote the set of bounded operators on $\mathcal{H}$ by $\mathcal{B}(\mathcal{H})$ and the trace class operators on $\mathcal{H}$ by $\mathcal{T}_1(\mathcal{H})$. The space of density operators (i.e., positive semidefinite trace class operators with trace 1) on $\mathcal{H}$ is denoted by $\mathcal{S}(\mathcal{H})$. It describes the space of quantum states on $\mathcal{H}$, we will use the terms 'density operator' and 'quantum state' interchangeably. Throughout the paper, we work with finite-dimensional Hilbert spaces $\mathcal{H} \cong \mathbb{C}^d$, but as we sometimes consider states with classical subsystems on a continuous alphabet, we nevertheless employ the notion of trace class operators. When viewing multiple quantum systems with associated Hilbert spaces $\mathcal{H}_1, \ldots, \mathcal{H}_m$ as a single composite quantum system, the associated Hilbert space is the tensor product $\bigotimes_{i=1}^m \mathcal{H}_i$. We obtain the reduced density matrix ρ_j on subsystem j of a multipartite state $\rho_{1,\ldots,m} \in \mathcal{S}(\bigotimes_{i=1}^m \mathcal{H}_i)$ via a partial trace over the remaining subsystems, $\rho_j = \mathrm{Tr}_{1,\ldots,j-1,j+1,\ldots,m}[\rho_{1,\ldots,m}]$. A trace with a subscript always indicates a partial trace over the Hilbert space with the same subscript.

We next define states that have both classical and quantum subsystems:

Definition 1 (Classical-Quantum (CQ) States) *Let X be a (classical) measurable space, let $\mathcal{H}$ be a Hilbert space. Let P be a probability measure on X and let $\mathsf{X} \ni x \mapsto \rho(x) \in \mathcal{S}(\mathcal{H})$ be a (Borel-)measurable mapping from elements of the alphabet to quantum states. The associated* classical-quantum (CQ) *state is given by*

$$\mathop{\mathbb{E}}_{x\sim P}\Big[|x\rangle\langle x| \otimes \rho(x)\Big] . \tag{17}$$

Here, the expectation value can be understood as a Bochner integral of a function mapping to a Banach space. If X is a finite alphabet, then the expression in Equation (17) simplifies to

$$\mathop{\mathbb{E}}_{x\sim P}\Big[|x\rangle\langle x| \otimes \rho(x)\Big] = \sum_{x\in\mathsf{X}} P(x)\, |x\rangle\langle x| \otimes \rho(x) . \tag{18}$$

Quantum information theory is a rich field and has successfully "quantized" a variety of notions from classical information theory. We will make use of the quantum counterpart of the classical relative entropy (also known as Kullback-Leibler divergence).

Definition 2 (Quantum relative entropy) *The* quantum relative entropy *between a density operator $\rho \in \mathcal{S}(\mathcal{H})$ and a positive semi-definite $\sigma \in \mathcal{B}(\mathcal{H})$ is given by*

$$D(\rho\|\sigma) = \begin{cases} \operatorname{Tr}[\rho(\log\rho - \log\sigma)] & \textit{if } \operatorname{supp}(\rho) \subseteq \operatorname{supp}(\sigma) \\ +\infty & \textit{else} \end{cases} . \tag{19}$$

Here, the support of ρ is, by Hermiticity, the orthogonal complement of its kernel, that is, $\operatorname{supp}(\rho) = (\ker(\rho))^{\perp}$.

From the quantum relative entropy, we can now obtain the quantum mutual information. It measures how much information one subsystem in a bipartite quantum state carries about the other subsystem.

Definition 3 (Quantum mutual information) *Let $\rho = \rho_{AB} \in \mathcal{S}(\mathcal{H}_A \otimes \mathcal{H}_B)$ be a bipartite quantum state. The* quantum mutual information (QMI) *between subsystems A and B in the quantum state $\rho = \rho_{AB}$ is given by*

$$I(A;B)_\rho = D(\rho_{AB}\|\rho_A \otimes \rho_B) = H(\rho_A) + H(\rho_B) - H(\rho_{AB}) , \tag{20}$$

where $H(\sigma) = -\operatorname{Tr}[\sigma \log(\sigma)]$ denotes the von Neumann entropy.

When applied to a CQ state, the QMI gives rise to the so-called Holevo information:

Definition 4 (Holevo information) *Let $\{P(x), \rho(x)\}_{x\in\mathsf{X}}$ be an ensemble of quantum states. The* Holevo information *is given by the QMI between the classical and quantum registers in the associated CQ state:*

$$\chi\left(\{P(x), \rho(x)\}_{x\in\mathsf{X}}\right) = I(C:Q)_{\mathbb{E}_{x\sim P}[|x\rangle\langle x|\otimes\rho(x)]} . \tag{21}$$

The Holevo information can equivalently be expressed as

$$\chi\left(\{P(x),\rho(x)\}_{x\in\mathsf{X}}\right) = H\left(\mathop{\mathbb{E}}_{x\sim P}[\rho(x)]\right) - \mathop{\mathbb{E}}_{x\sim P}[H(\rho(x))] = \mathop{\mathbb{E}}_{x\sim P}\left[D\left(\rho(x)\middle\|\mathop{\mathbb{E}}_{\tilde{x}\sim P}[\rho(\tilde{x})]\right)\right]. \tag{22}$$

In addition to quantum states, we need mathematical descriptions for measurements as well as for general processing of quantum systems. To describe measurements, we use positive operator-valued measures (POVMs).

Definition 5 (POVMs and post-measurement states) *The set of* effect operators $\mathcal{E}(\mathcal{H})$ *on a Hilbert space* $\mathcal{H}$ *is given by* $\mathcal{E}(\mathcal{H}) = \{E\in\mathcal{B}(\mathcal{H})\mid E = E^\dagger \wedge 0\leq E\leq \mathbb{1}_{\mathcal{H}}\}$. *A collection* $\{E_k\}_{k=1}^K\subset\mathcal{E}(\mathcal{H})$ *with* $\sum_{k=1}^K E_k = \mathbb{1}_{\mathcal{H}}$ *is called* K-outcome POVM. *When measuring a POVM* $\{E_k\}_{k=1}^K$ *on a state* $\rho\in\mathcal{S}(\mathcal{H})$, *the probability of observing outcome* k *is* $\mathrm{Tr}[E_k\rho]$. *Moreover, conditioned on observing outcome* k, *the post-measurement state is given by*

$$\rho_k := \frac{\sqrt{E_k}\rho\sqrt{E_k}}{\mathrm{Tr}[E_k\rho]}. \tag{23}$$

The dynamics of quantum systems can be mathematically described by quantum channels.

Definition 6 (Quantum channels – Schrödinger picture) *A linear map* $\Lambda:\mathcal{T}_1(\mathcal{H}_{\mathrm{in}})\to\mathcal{T}_1(\mathcal{H}_{\mathrm{out}})$ *between trace class operators on Hilbert spaces* $\mathcal{H}_{\mathrm{in}}$ *and* $\mathcal{H}_{\mathrm{out}}$ *is called a* quantum channel (in the Schrödinger picture) *if it is completely positive (CP) and trace-preserving (TP). Here, we call* Λ *completely positive if, for any* $\mathcal{H}_{\mathrm{aux}}$, $(\mathrm{id}_{\mathcal{T}_1(\mathcal{H}_{\mathrm{aux}})}\otimes\Lambda)(\rho)$ *is positive-semidefinite whenever* $\rho\in\mathcal{T}_1(\mathcal{H}_{\mathrm{aux}}\otimes\mathcal{H}_{\mathrm{in}})$ *is positive semidefinite, and we call* Λ *trace-preserving if* $\mathrm{Tr}[\Lambda(\rho)] = \mathrm{Tr}[\rho]$ *holds for all* $\rho\in\mathcal{T}_1(\mathcal{H}_{\mathrm{in}})$.

According to Definition 6, we describe a general quantum process with a CPTP map. This is the Schrödinger picture perspective, in which we view states as evolving. Complementary to this, we can define the dual $\Lambda^*:\mathcal{B}(\mathcal{H}_{\mathrm{out}})\to\mathcal{B}(\mathcal{H}_{\mathrm{in}})$ of Λ via the requirement $\mathrm{Tr}[E\Lambda(\rho)] = \mathrm{Tr}[\Lambda^*(E)\rho]$ $\forall\rho\in\mathcal{S}(\mathcal{H}_{\mathrm{in}}), E\in\mathcal{E}(\mathcal{H}_{\mathrm{out}})$. The *Heisenberg picture* map Λ^* is completely positive if and only if Λ is. Also, Λ being TP is equivalent to Λ^* being unital (U), i.e., $\Lambda^*(\mathbb{1}_{\mathcal{H}_{\mathrm{out}}}) = \mathbb{1}_{\mathcal{H}_{\mathrm{in}}}$. Thus, quantum channels in the Heisenberg picture are linear CPU maps.

Finally, we recall two recently introduced notions from quantum optimal transport. These constitute alternatives to the trace distance between multi-qudit states and the operator norm for multi-qudit observables, respectively, and take locality into account.

Definition 7 (Quantum Wasserstein-1 distance (De Palma et al., 2021)) *Let* $\rho,\sigma\in\mathcal{S}((\mathbb{C}^d)^{\otimes m})$ *be two* m*-qudit states. The* quantum Wasserstein-1 distance $\|\rho-\sigma\|_{W_1}$ *between* ρ *and* σ *is defined as*

$$\|\rho-\sigma\|_{W_1} = \min\left\{\sum_{i=1}^m c_i \,\middle|\, \begin{array}{l} c_i\geq 0:\exists\rho^{(i)},\sigma^{(i)}\in\mathcal{S}((\mathbb{C}^d)^{\otimes m}), 1\leq i\leq m \text{ s.t.}\\ \mathrm{Tr}_i[\rho^{(i)}] = \mathrm{Tr}_i[\sigma^{(i)}]\forall i \wedge \rho-\sigma = \sum_{i=1}^m c_i\left(\rho^{(i)}-\sigma^{(i)}\right)\end{array}\right\}. \tag{24}$$

The quantum Wasserstein-1 distance between quantum states induces a notion of quantum Lipschitz constant for observables via duality.

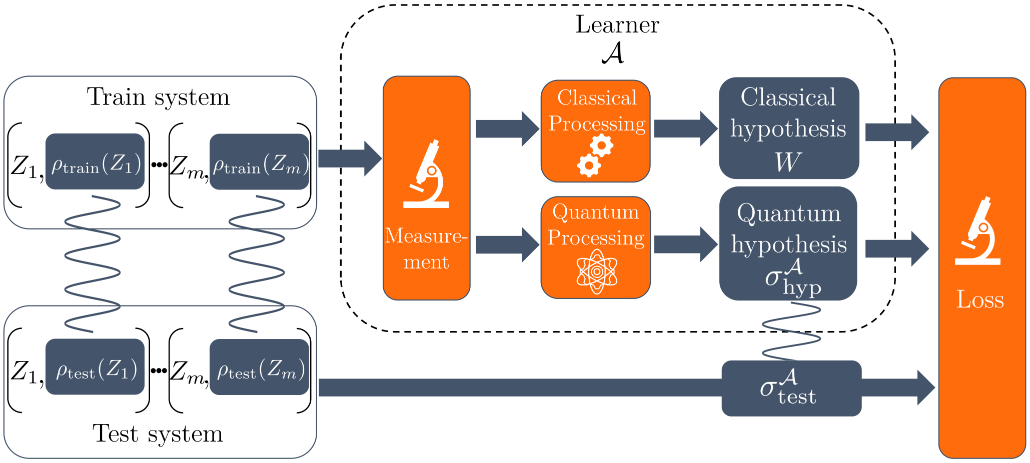
Learner
$\mathcal{A}$
Train system
$Z_1, \rho_{\text{train}}(Z_1)$
$Z_m, \rho_{\text{train}}(Z_m)$
Measurement
Classical Processing
Quantum Processing
Classical hypothesis W
Quantum hypothesis $\sigma^{\mathcal{A}}_{\text{hyp}}$
Loss
$Z_1, \rho_{\text{test}}(Z_1)$
$Z_m, \rho_{\text{test}}(Z_m)$
Test system
$\sigma^{\mathcal{A}}_{\text{test}}$

Figure 1: **Framework for learning from classical-quantum data:** The quantum learner $\mathcal{A}$ acts on the classical data and on the training subsystem of the quantum data via a measurement followed by classical and quantum post-processing. The performance of the resulting classical and quantum hypotheses are then evaluated via a loss measurement that also takes the testing subsystem of the quantum data into account. The training and testing subsystems may initially be correlated or even entangled.

Definition 8 (Quantum Lipschitz constant (De Palma et al., 2021)) *Let $H = H^\dagger \in \mathcal{B}((\mathbb{C}^d)^{\otimes m})$ be an m-qudit observable. The* quantum Lipschitz constant $\|H\|_{\text{Lip}}$ *of H is defined as*

$$\|H\|_{\text{Lip}} = \max\left\{\text{Tr}[HX] \mid X = X^\dagger \in \mathcal{B}((\mathbb{C}^d)^{\otimes m}) : \text{Tr}[X] = 0 \wedge \|X\|_{W_1} \leq 1\right\} \tag{25}$$

$$= \max_{1\leq i\leq m} \max\left\{\text{Tr}[H(\rho-\sigma)] \mid \rho,\sigma \in \mathcal{S}((\mathbb{C}^d)^{\otimes m}) : \text{Tr}_i[\rho] = \text{Tr}_i[\sigma]\right\}. \tag{26}$$

Appendix B. Framework and main result

B.1. Framework for learning from classical-quantum data

We aim to provide a formalism for learning from quantum data given as a classical-quantum (CQ) state. Our framework is visualized in Figure 1. We suppose that the data comes in the form of a CQ state

$$\rho = \mathop{\mathbb{E}}_{S\sim P^m}\left[|S\rangle\langle S| \otimes \rho(S)\right], \tag{27}$$

with P a probability measure over a classical measurable instance space Z, and with $\rho(s)$ a density operator on a (typically composite) data Hilbert space $\mathcal{H}_{\text{data}}$, $\rho(s) \in \mathcal{S}(\mathcal{H}_{\text{data}})$, for each $s \in \mathsf{Z}^m$.

A quantum learner $\mathcal{A}$ now consists of:

(i) a (possibly trivial) decomposition of the data Hilbert space into a tensor product of a test data and a training data Hilbert space, $\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{test}} \otimes \mathcal{H}_{\text{train}}$,

(ii) a measurable hypothesis space W,

(iii) POVMs $\{E_s^{\mathcal{A}}(w)\}_{w\in\mathsf{W}}$ on $\mathcal{H}_{\text{train}}$ for each $s\in\mathsf{Z}^m$, describing the measurements used by the learner to extract classical information from the training data state and leading to probability distributions[3] $Q_s^{\mathcal{A}}$ on W defined via $Q_s^{\mathcal{A}}(w)=\operatorname{Tr}[E_s(w)\operatorname{Tr}_{\text{test}}[\rho(s)]]$,

(iv) a quantum hypothesis Hilbert space $\mathcal{H}_{\text{hyp}}$,

(v) a family of quantum channels $\{\Lambda_{s,w}^{\mathcal{A}}:\mathcal{T}_1(\mathcal{H}_{\text{train}})\to\mathcal{T}_1(\mathcal{H}_{\text{hyp}})\}_{(s,w)\in\mathsf{Z}^m\times\mathsf{W}}$.

That is, the learner $\mathcal{A}$ proceeds as follows: First, conditioned on the classical data s, $\mathcal{A}$ performs the measurement described by the POVM $\{E_s^{\mathcal{A}}(w)\}_{w\in\mathsf{W}}$ on the training data subsystem of $\rho(s)$ and classically records the measurement outcome. Second, conditioned on both the classical data s and the observed measurement outcome w, $\mathcal{A}$ applies the quantum channel $\Lambda_{s,w}^{\mathcal{A}}$ to the post-measurement state of the training data subsystem. This way, the action of the learner $\mathcal{A}$ on the CQ data state ρ leads to the CQ output state

$$\begin{aligned}\sigma^{\mathcal{A}} &= \mathop{\mathbb{E}}_{S\sim P^m}\left[|S\rangle\langle S|\otimes\mathop{\mathbb{E}}_{W\sim Q_S^{\mathcal{A}}}\left[(\mathrm{id}_{\text{test}}\otimes\Lambda_{S,W}^{\mathcal{A}})\left(\rho^{\mathcal{A}}(S,W)\right)\otimes|W\rangle\langle W|\right]\right] &(28)\\ &= \mathop{\mathbb{E}}_{S\sim P^m}\mathop{\mathbb{E}}_{W\sim Q_S^{\mathcal{A}}}\left[|S\rangle\langle S|\otimes(\mathrm{id}_{\text{test}}\otimes\Lambda_{S,W}^{\mathcal{A}})\left(\rho^{\mathcal{A}}(S,W)\right)\otimes|W\rangle\langle W|\right] &(29)\\ &= \mathop{\mathbb{E}}_{S\sim P^m}\mathop{\mathbb{E}}_{W\sim Q_S^{\mathcal{A}}}\left[|S\rangle\langle S|\otimes\sigma^{\mathcal{A}}(S,W)\otimes|W\rangle\langle W|\right], &(30)\end{aligned}$$

where we have defined the post-measurement state

$$\rho^{\mathcal{A}}(s,w)=\frac{\left(\mathbb{1}_{\text{test}}\otimes\sqrt{E_s^{\mathcal{A}}(w)}\right)\rho(s)\left(\mathbb{1}_{\text{test}}\otimes\sqrt{E_s^{\mathcal{A}}(w)}\right)}{\operatorname{Tr}[E_s^{\mathcal{A}}(w)\rho_{\text{train}}(s)]}\,, \tag{31}$$

Note that $\sigma^{\mathcal{A}}(s,w)\in\mathcal{S}(\mathcal{H}_{\text{test}}\otimes\mathcal{H}_{\text{hyp}})$ for every $(s,w)\in\mathsf{Z}^m\times\mathsf{W}$. If we denote by $P^{\mathcal{A}}$ the induced probability distribution over $\mathsf{Z}^m\times\mathsf{W}$ with

$$P^{\mathcal{A}}(s,w)=P^m(s)\cdot Q_s^{\mathcal{A}}(w), \tag{32}$$

denote its marginal on W by $P_{\mathsf{W}}^{\mathcal{A}}$, and its conditional distribution for the data given the hypothesis W by $P_{\text{data}}^{\mathcal{A}}|W$, we can interchange the order of the expectations in Equation (30) and rewrite $\sigma^{\mathcal{A}}$ as

$$\sigma^{\mathcal{A}}=\mathop{\mathbb{E}}_{W\sim P_{\mathsf{W}}^{\mathcal{A}}}\mathop{\mathbb{E}}_{S\sim P_{\text{data}}^{\mathcal{A}}|W}\left[|S\rangle\langle S|\otimes\sigma^{\mathcal{A}}(S,W)\otimes|W\rangle\langle W|\right]. \tag{33}$$

Remark 9 *In the language used in Section 1.1.1 to compare our framework to the classical one, the setup described above assumes perfectly correlated classical training and test data. This choice was made to simplify the presentation. However, one may extend the framework by considering the classical part of the data to consist of (in general correlated) training and test data. Naturally, the POVMs and channels performed by the learner should only depend on the training data but not on the test data. This straightforward extension of our framework then also encompasses the*

[3] This formulation implicitly assumes that W is discrete. If W is continuous, we can instead work with associated probability densities $q_s^{\mathcal{A}}(w)=\operatorname{Tr}[E_s(w)\operatorname{Tr}_{\text{test}}[\rho(s)]]$. Our framework and results encompass both the discrete and the continuous case. We choose discrete-case notation merely for simplicity.

classical extension of the (Xu and Raginsky, 2017) framework with test data that we describe in Section 1.1.1. Note that including separate classical test data also enables us to describe tasks in which the training data distribution is different from the test data distribution, for example in scenarios of covariate shift, where out-of-distribution generalization becomes relevant. This may allow for connecting our framework to recent work on out-of-distribution generalization in learning quantum processes (Caro et al., 2023; Huang et al., 2023a).

Remark 10 *Instead of describing $\mathcal{A}$ in terms of POVMs and channels, we could merge these objects into a description in terms of quantum instruments (compare (Heinosaari and Ziman, 2011, Chapter 5)). We have chosen a formulation based on POVMs and channels in order to make the presentation more concrete and widely accessible.*

Example 1 (Quantum state classification) *As an illustrative example, we consider a task of* quantum state classification*, in which the quantum learner should PAC learn a two-outcome POVM that distinguishes between pairs of d-dimensional states weighted according to prior probabilities. This can be viewed as a version of the problem studied in (Guţă and Kotłowski, 2010) but with an underlying distribution over weighted pairs of states. To formalize this problem, we consider a probability distribution $P_{\text{weight}} \otimes P_{\text{pair}}$ over the space $[0,1] \times \big(\mathcal{S}(\mathbb{C}^d) \times \mathcal{S}(\mathbb{C}^d)\big)$ of weights and pairs of states. If the learner has access to m labeled quantum examples generated from this distribution, the overall classical-quantum data is described by the state*

$$\rho = \mathop{\mathbb{E}}_{\{\pi_0^{(i)},(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m \sim (P_{\text{weight}} \otimes P_{\text{pair}})^m} \left[\bigotimes_{i=1}^m \left(\pi_0^{(i)} \, |0\rangle\!\langle 0| \otimes (\sigma_0^{(i)})^{\otimes 2} + (1-\pi_0^{(i)}) \, |1\rangle\!\langle 1| \otimes (\sigma_1^{(i)})^{\otimes 2} \right) \right] \tag{34}$$

$$= \mathop{\mathbb{E}}_{\{\pi_0^{(i)}\}_{i=1}^m \sim P_{\text{weight}}^m} \left[\sum_{s=(z_1,\ldots,z_m)\in\{0,1\}^m} \left(\prod_{i=1}^m \pi_{z_i}^{(i)} \right) |s\rangle\!\langle s| \otimes \mathop{\mathbb{E}}_{\{(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m \sim P_{\text{pair}}^m} \left[\left(\bigotimes_{i=1}^m \sigma_{z_i}^{(i)} \right)^{\otimes 2} \right] \right], \tag{35}$$

where we used the notation $\pi_1^{(i)} = 1 - \pi_0^{(i)}$. If we define $\mathsf{Z} = \{0,1\}$, if we let P be the probability distribution on Z defined via

$$P(z_i) = \mathop{\mathbb{E}}_{\{\pi_0^{(i)}\}_{i=1}^m \sim P_{\text{weight}}^m} \left[\pi_{z_i}^{(i)} \right] \quad \forall z_i \in \{0,1\}, \tag{36}$$

and if we further define the density operators $\rho(s)$ acting on the Hilbert space $\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{test}} \otimes \mathcal{H}_{\text{train}} = (\mathbb{C}^d)^{\otimes m} \otimes (\mathbb{C}^d)^{\otimes m}$ as

$$\rho(s) = \mathop{\mathbb{E}}_{\{(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m \sim P_{\text{pair}}^m} \left[\left(\bigotimes_{i=1}^m \sigma_{z_i}^{(i)} \right)^{\otimes 2} \right] \quad \forall s = (z_1,\ldots,z_m) \in \{0,1\}^m, \tag{37}$$

then, we see that

$$\rho = \mathop{\mathbb{E}}_{S\sim P^m} \Big[|S\rangle\!\langle S| \otimes \rho(S) \Big] \tag{38}$$

in accordance with Equation (27).

To describe a quantum learner $\mathcal{A}$ in this setting, take the quantum hypothesis space $\mathcal{H}_{\mathrm{hyp}} = \mathbb{C}$ to be trivial and consider a measurable hypothesis space W. Here, we imagine each $w \in \mathsf{W}$ to be associated to a two-outcome qudit POVM $\{F(w), \mathbb{1}_d - F(w)\}$, which describes a measurement that the learner could use for the distinguishing task. Now, to every $s \in \{0,1\}^m$ we associate a POVM $\{E_s^{\mathcal{A}}(w)\}_{w\in\mathsf{W}}$. Note: As $\mathcal{T}_1(\mathcal{H}_{\mathrm{hyp}}) = \mathcal{T}_1(\mathbb{C}) = \mathbb{C}$ is trivial, so is the family of quantum channels $\{\Lambda_{s,w}^{\mathcal{A}} : \mathcal{T}_1(\mathcal{H}_{\mathrm{train}}) \to \mathcal{T}_1(\mathcal{H}_{\mathrm{hyp}})\}$ in this setting. That is, $\Lambda_{s,w}^{\mathcal{A}}(\cdot) = \mathrm{Tr}[(\cdot)]$ for all s, w. Thus, according to Equation (33), the action of the learner $\mathcal{A}$ on ρ leads to the output state

$$\sigma^{\mathcal{A}} = \mathop{\mathbb{E}}_{W\sim P_{\mathsf{W}}^{\mathcal{A}}} \mathop{\mathbb{E}}_{S\sim P_{\mathrm{data}|W}^{\mathcal{A}}} \left[|S\rangle\langle S| \otimes \sigma^{\mathcal{A}}(S,W) \otimes |W\rangle\langle W|\right], \tag{39}$$

with the probability distribution $P^{\mathcal{A}}$ on $\{0,1\}^m \times \mathsf{W}$ given by

$$P^{\mathcal{A}}(s,w) = P^m(s) \cdot \mathrm{Tr}\left[E_s(w)\rho_{\mathrm{train}}(s)\right] \tag{40}$$

and with the post-measurement subsystem states

$$\rho_{\mathrm{test}}(s) = \rho_{\mathrm{train}}(s) = \mathop{\mathbb{E}}_{\{(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m \sim P_{\mathrm{pair}}^m} \left[\bigotimes_{i=1}^m \sigma_{z_i}^{(i)}\right] \quad \forall s = (z_1,\ldots,z_m) \in \{0,1\}^m. \tag{41}$$

This concludes the example, we now return to the discussion of our general framework.

Given a learner $\mathcal{A}$ and a data CQ state as described above, we now define relevant notions of risk/error. In classical notion theory, the most commonly used such notions are those of empirical and true risk. As discussed in Section 1.1.1, the expected empirical risk arises as an average of losses with correlated training data and hypothesis random variables. In contrast, the expected true risk can be understood as an average of losses after decoupling training data and hypothesis. To define analogous notions for quantum learning, we go from loss functions to loss observables. Moreover, we extend the intuition that decoupling makes the difference between empirical and true risk to a decoupling on both the classical and the quantum level.

For the next three definitions, ρ, $\mathcal{A}$, $P^{\mathcal{A}}$, and $\sigma^{\mathcal{A}}$ are as introduced above, and we consider a family of self-adjoint loss observables $\{L(s,w)\}_{(s,w)\in\mathsf{Z}^m\times\mathsf{W}}$ with $L(s,w) \in \mathcal{B}(\mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{hyp}})$.

Definition 11 (Expected empirical risk) *The* expected empirical risk *of $\mathcal{A}$ w.r.t. ρ as measured by $\{L(s,w)\}_{(s,w)\in\mathsf{Z}^m\times\mathsf{W}}$ is defined as*

$$\hat{R}_\rho(\mathcal{A}) := \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}} \left[\mathrm{Tr}[L(S,W)\sigma^{\mathcal{A}}(S,W)]\right]. \tag{42}$$

Definition 12 (Expected true risk) *The* expected true risk *of $\mathcal{A}$ w.r.t. ρ and $\{L(s,w)\}_{(s,w)\in\mathsf{Z}^m\times\mathsf{W}}$ is defined as*

$$R_\rho(\mathcal{A}) := \mathop{\mathbb{E}}_{(\bar{S},\bar{W})\sim P^m\otimes P_{\mathsf{W}}^{\mathcal{A}}} \left[\mathrm{Tr}\left[L(\bar{S},\bar{W})\left(\rho_{\mathrm{test}}(\bar{S}) \otimes \sigma_{\mathrm{hyp}}^{\mathcal{A}}(\bar{S},\bar{W})\right)\right]\right]. \tag{43}$$

As before, here we let $\bar{S}, \bar{W}$ denote independent copies of S and W.

Definition 13 (Expected generalization error) *The* expected generalization error *of $\mathcal{A}$ w.r.t. ρ as measured by $\{L(s,w)\}_{(s,w)\in \mathsf{Z}^m\times \mathsf{W}}$ is defined to be*

$$\mathrm{gen}_\rho(\mathcal{A}) := R_\rho(\mathcal{A}) - \hat{R}_\rho(\mathcal{A}), \tag{44}$$

the difference between the expected true and empirical risks of $\mathcal{A}$ w.r.t. ρ and $\{L(s,w)\}_{(s,w)\in \mathsf{Z}^m\times \mathsf{W}}$.

Remark 14 *Note that there is some freedom in our definition of channels $\Lambda^{\mathcal{A}}_{s,w}$ and loss observables $L(s,w)$ because of the duality of Schrödinger and Heisenberg pictures. Concretely, if $\Lambda^{\mathcal{A}}_{s,w} = \Lambda^{''\mathcal{A}}_{s,w} \circ \Lambda^{'\mathcal{A}}_{s,w}$ and if we define $L^{'}(s,w) = (\mathrm{id}_{\text{test}} \otimes \Lambda^{''\mathcal{A}}_{s,w})^*(L(s,w))$, then the expected empirical and true risks obtained by considering $\Lambda^{'\mathcal{A}}_{s,w}$ and $L^{'}(s,w)$ coincide with those originally obtained from $\Lambda^{\mathcal{A}}_{s,w}$ and $L(s,w)$.*

Next, we illustrate these definitions in two concrete examples. First, we demonstrate how they recover the classical case, before continuing the discussion of our state classification application.

Example 2 *Starting from Definitions 11 to 13, we can reproduce the corresponding classical notions of expected empirical risk, expected true risk, and expected generalization error in (at least) the following two ways: On the one hand, if we assume all involved quantum systems to be trivial (i.e., $\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{hyp}} = \mathbb{C}$), then the loss observables are real scalars. Interpreting these as classical loss functions, we recover the notions familiar from the classical case. On the other hand, even when (some of) the involved quantum systems are non-trivial, if we consider loss observables $L(s,w) = \ell(s,w)\cdot \mathbb{1}_{\text{test,hyp}}$ given by multiples of the identity, with classical loss function values $\ell(s,w) = \frac{1}{m}\sum_{i=1}^m \ell(w,z_i)$, then the trace-normalization of $\sigma^{\mathcal{A}}(s,w)$ and $\rho_{\text{test}}(s)\otimes \sigma^{\mathcal{A}}_{\text{hyp}}(s,w)$ ensures that we again obtain the same quantities as in the classical case. As we will see later, our results for this latter setting indeed reproduce the classical bounds of (Xu and Raginsky, 2017).*

Example 3 (Quantum state classification – Example 1 continued) *To obtain reasonable notions of risk in the quantum state classification setting of Example 1, we can take the loss observables for $s = (z_1,\dots,z_m)\in\{0,1\}^m$ and $w\in\mathsf{W}$ to be*

$$L(s,w) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_d^{\otimes(i-1)} \otimes ((1-z_i)(\mathbb{1}_d - F(w)) + z_i F(w)) \otimes \mathbb{1}_d^{\otimes(m-i)}. \tag{45}$$

With this choice, the expected empirical risk from Definition 11 becomes

$$\begin{aligned}\hat{R}_\rho(\mathcal{A}) &= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[\mathrm{Tr}[L(S,W)\sigma^{\mathcal{A}}(S,W)]\right] &(46)\\ &= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[\frac{1}{m}\sum_{i=1}^m \mathrm{Tr}\left[((1-Z_i)(\mathbb{1}_d - F(W)) + Z_i F(W)) \mathop{\mathbb{E}}_{(\sigma_0^{(i)},\sigma_1^{(i)})\sim P_{\text{pair}}}\left[\sigma^{(i)}_{Z_i}\right]\right]\right] &(47)\\ &= \mathop{\mathbb{E}}_{(\{\pi_0^{(i)},(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m, W)\sim P^{\mathcal{A}}}\left[\frac{1}{m}\sum_{i=1}^m \left(\pi_0^{(i)}\,\mathrm{Tr}[(\mathbb{1}_d - F(W))\sigma_0^{(i)}] + \pi_1^{(i)}\,\mathrm{Tr}[F(W)\sigma_1^{(i)}]\right)\right], &(48)\end{aligned}$$

where the last step uses the definition of P from Example 1 and, in a slight abuse of notation, uses $P^{\mathcal{A}}$ to also denote the induced joint distribution over weighted pairs of states and hypotheses. This *induced distribution can explicitly be written as*

$$\begin{aligned}&P^{\mathcal{A}}\left(\{\pi_0^{(i)},(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m, w\right)\\&=\left(\prod_{i=1}^m P_{\mathrm{weight}}(\pi_0^{(i)})\right)\cdot\left(\prod_{i=1}^m P_{\mathrm{pair}}(\sigma_0^{(i)},\sigma_1^{(i)})\right)\cdot\sum_{s\in\{0,1\}^m}\left(\prod_{i=1}^m \pi_{z_i}^{(i)}\right)\cdot\operatorname{Tr}\left[E_s^{\mathcal{A}}(w)\rho_{\mathrm{train}}(s)\right].\end{aligned}\tag{49}$$

Thus, Equation (48) is exactly the expected probability that the quantum learner misclassifies an unknown state, where the average is over the joint distribution of training data and hypothesis. This is the natural notion of expected empirical risk in this scenario.

The expected true risk according to Definition 12 is

$$R_\rho(\mathcal{A})=\mathop{\mathbb{E}}_{(\bar{S},\bar{W})\sim P^m\otimes P_{\mathsf{W}}^{\mathcal{A}}}\left[\operatorname{Tr}\left[L(\bar{S},\bar{W})\left(\rho_{\mathrm{test}}(\bar{S})\otimes\sigma_{\mathrm{hyp}}^{\mathcal{A}}(\bar{S},\bar{W})\right)\right]\right]\tag{50}$$

$$=\mathop{\mathbb{E}}_{(\{\bar{\pi}_0^{(i)},(\bar{\sigma}_0^{(i)},\bar{\sigma}_1^{(i)})\}_{i=1}^m,\bar{W})\sim(P_{\mathrm{weight}}\otimes P_{\mathrm{pair}})^m\otimes P_{\mathsf{W}}^{\mathcal{A}}}\left[\frac{1}{m}\sum_{i=1}^m\left(\bar{\pi}_0^{(i)}\operatorname{Tr}[(\mathbb{1}_d-F(\bar{W}))\bar{\sigma}_0^{(i)}]+\bar{\pi}_1^{(i)}\operatorname{Tr}[F(\bar{W})\bar{\sigma}_1^{(i)}]\right)\right]\tag{51}$$

$$=\mathop{\mathbb{E}}_{(\bar{\pi}_0,(\bar{\sigma}_0,\bar{\sigma}_1),\bar{W})\sim P_{\mathrm{weight}}\otimes P_{\mathrm{pair}}\otimes P_{\mathsf{W}}^{\mathcal{A}}}\left[\bar{\pi}_0\operatorname{Tr}\left[(\mathbb{1}_d-F(\bar{W}))\bar{\sigma}_0\right]+\bar{\pi}_1\operatorname{Tr}[F(\bar{W})\bar{\sigma}_1]\right],\tag{52}$$

where the random variables with bars again denote independent copies of the respective unbarred random variables. Thus, the expected true risk is exactly the expected probability that the quantum learner misclassifies an unknown state from a new, independently drawn weighted pair, a natural choice of expected true risk in this setting. Hence, the expected generalization error from Definition 13 indeed reproduces the natural expression, namely the difference between the expected misclassification probability on a randomly drawn new data point and the expected average misclassification probability over the training data. This concludes the discussion of risks for our state classification tasks.

In Appendix C, we demonstrate that the general notions of risks introduced in Definitions 11 to 13 reproduce further natural performance measures for suitably chosen loss observable L in learning scenarios such as PAC learning quantum states, learning classical functions from entangled quantum data, and quantum parameter estimation, among others.

B.2. Generalization bounds for learning from classical-quantum data

The remainder of this section is concerned with proving that classical and quantum moment generating function assumptions lead to expected generalization error bounds in terms of quantities measuring the classical and quantum information between the data and the output of the learner. This lifts the following intuition from classical to quantum learning: Learners generalize well (in distribution) if they produce hypotheses that do not depend too strongly on the specific dataset that they were trained on.

Table 1 compiles relevant notation for the formulation of our results. Before stating them, we recall the following definition from convex analysis and a lemma about the quantum relative entropy:

Object	Notation
Probability density of classical data	P
Input data CQ state	$\rho = \mathbb{E}_{S\sim P^m}[\lvert S\rangle\langle S\rvert \otimes \rho(S)]$
POVMs associated with learner $\mathcal{A}$	$E_s^{\mathcal{A}}(w)$
Joint distribution induced by learner $\mathcal{A}$	$P^{\mathcal{A}}$
CPTP maps associated with learner $\mathcal{A}$	$\Lambda_{s,w}^{\mathcal{A}}$
Learner output	$\sigma^{\mathcal{A}} = \mathbb{E}_{(S,W)\sim P^{\mathcal{A}}}\left[\lvert S\rangle\langle S\rvert \otimes \sigma^{\mathcal{A}}(S,W) \otimes \lvert W\rangle\langle W\rvert\right]$
Loss observables	$L(s,w)$
Quantum mutual information	$I(\cdot;\cdot)_\bullet$
Holevo information	$\chi(\{\cdot,\cdot\})$
Quantum log-MGF bound	$\psi_\pm$
Classical log-MGF bound	$\phi_\pm$

Table 1: Notation for the various mathematical objects appearing in this section.

Definition 15 (Fenchel-Legendre dual) *Let $\psi : \mathbb{R} \to \mathbb{R}$ be lower-semi-continuous and convex. The Fenchel-Legendre dual $\psi^* : \mathbb{R} \to \mathbb{R}$ is defined as*

$$\psi^*(t) = \sup_{\lambda\in\mathbb{R}}\{\lambda t - \psi(\lambda)\}. \tag{53}$$

Lemma 16 (Petz's variational characterization of the quantum relative entropy (Petz, 1988)) *Let $\sigma_1, \sigma_2 \in \mathcal{S}(\mathcal{H})$ be two quantum states. Then, the relative entropy between σ_1 and σ_2 can be rewritten as follows:*

$$D(\sigma_1\|\sigma_2) = \sup_{H=H^\dagger\in\mathcal{B}(\mathcal{H})}\{\mathrm{Tr}[\sigma_1 H] - \log \mathrm{Tr}[\exp\left(\log(\sigma_2) + H\right)]\}. \tag{54}$$

We can now state and prove our main result:

Theorem 17 (Expected generalization error bound via quantum mutual information) *Assume that, for every $(s,w) \in \mathsf{Z}^m \times \mathsf{W}$,*

$$\log \mathrm{Tr}\left[(\rho_{\mathrm{test}}(s) \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w))e^{\lambda\left(L(s,w)-\mathrm{Tr}[L(s,w)(\rho_{\mathrm{test}}(s)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w))]\mathbb{1}_{\mathrm{test,hyp}}\right)}\right] \le \begin{cases}\psi_+(\lambda) & \textit{if } \lambda \ge 0\\ \psi_-(\lambda) & \textit{if } \lambda < 0\end{cases}, \tag{QMGF}$$

where $\psi_+, \psi_- : \mathbb{R} \to \mathbb{R}$ are convex, differentiable at 0, and satisfy $\psi_\pm(0) = \psi'_\pm(0) = 0$. Moreover, assume that, for every $w \in \mathsf{W}$,

$$\log \mathop{\mathbb{E}}_{S\sim P^m}\left[e^{\lambda(\mathrm{Tr}[L(S,w)(\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,w))]-\mathbb{E}_{\tilde{S}\sim P^m}\left[\mathrm{Tr}[L(\tilde{S},w)(\rho_{\mathrm{test}}(\tilde{S})\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(\tilde{S},w))]\right])}\right] \le \begin{cases}\phi_+(\lambda) & \textit{if } \lambda \ge 0\\ \phi_-(\lambda) & \textit{if } \lambda < 0\end{cases}, \tag{CMGF}$$

where $\phi_+, \phi_- : \mathbb{R} \to \mathbb{R}$ *are convex, differentiable at* 0*, and satisfy* $\phi_\pm(0) = \phi'_\pm(0) = 0$*. Then,*

$$\pm\,\mathrm{gen}_\rho(\mathcal{A}) \leq \psi_\mp^{*-1}\left(\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}[I(\text{test};\text{hyp})_{\sigma(S,W)}] + \mathop{\mathbb{E}}_{S\sim P^m}\left[\chi\left(\{P^{\mathcal{A}}_{\mathsf{W}|S}(w), \rho^{\mathcal{A}}_{\text{test}}(S,w)\}_{w\in\mathsf{W}}\right)\right]\right) \\ + \phi_\mp^{*-1}\left(I(S;W)\right)\,. \tag{55}$$

Our proof is inspired by the reasoning used in the classical case, for instance in (Xu and Raginsky, 2017; Raginsky, 2019; Bu et al., 2020), but differs from it in three non-trivial ways. First, one central ingredient in the classical argument, namely the Donsker-Varadhan representation of the classical relative entropy, has to be replaced by its quantum counterpart, Lemma 16. Second, to deal with potential complications about matrix exponentials arising from non-commutativity, we rely on the Golden-Thompson inequality. Finally, while there is only one decoupling step in the classical proof, our scenario requires both a classical and a quantum decoupling. Thus, our analysis uses an additional decomposition of the expected generalization error compared to the classical case.

Proof When combined with the Golden-Thompson inequality (see, e.g., Bhatia, 1997, Section IX.3), which tells us that $\mathrm{Tr}[e^{A+B}] \leq \mathrm{Tr}[e^A e^B]$ for Hermitian matrices A and B, Lemma 16 implies, for every $(s,w) \in \mathsf{Z}^m \times \mathsf{W}$ and for all $\lambda \in \mathbb{R}$,

$$D(\sigma^{\mathcal{A}}(s,w)\|\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w)) \tag{56}$$

$$\geq \lambda\,\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \log\mathrm{Tr}\left[\exp\left(\log(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w)) + \lambda L(s,w)\right)\right] \tag{57}$$

$$\geq \lambda\,\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \log\mathrm{Tr}\left[(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))\exp\left(\lambda L(s,w)\right)\right] \tag{58}$$

$$= \lambda\left(\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \mathrm{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))]\right) \\ \quad - \log\mathrm{Tr}\left[(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))e^{\lambda\left(L(s,w)-\mathrm{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))]\mathbb{1}_{\text{test,hyp}}\right)}\right] \tag{59}$$

$$\geq \lambda\left(\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \mathrm{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))]\right) - \begin{cases}\psi_+(\lambda) & \text{if } \lambda\geq 0\\ \psi_-(\lambda) & \text{if } \lambda < 0\end{cases}\,. \tag{60}$$

Here, the first step uses Lemma 16, the second is due to the Golden-Thompson inequality, the third step is a simple rewriting, and the final step consists in plugging in Equation (QMGF).

We can now rearrange this inequality and optimize over λ to obtain:

$$\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \mathrm{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))] \leq \inf_{\lambda>0}\frac{D(\sigma^{\mathcal{A}}(s,w)\|\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w)) + \psi_+(\lambda)}{\lambda}, \tag{61}$$

$$-\left(\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \mathrm{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))]\right) \leq \inf_{\lambda<0}\frac{D(\sigma^{\mathcal{A}}(s,w)\|\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w)) + \psi_-(\lambda)}{\lambda}. \tag{62}$$

Using (Boucheron et al., 2013, Lemma 2.4), we can rewrite the infima in terms of the generalized inverses $\psi_\pm^{*-1}(s) = \inf\{t\geq 0 \mid \psi_\pm^*(t) > s\}$ of the Fenchel-Legendre duals of $\psi_\pm$ to obtain

$$\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \mathrm{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))] \leq \psi_+^{*-1}(D(\sigma^{\mathcal{A}}(s,w)\|\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))), \tag{63}$$

$$-\left(\mathrm{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \mathrm{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))]\right) \leq \psi_-^{*-1}(D(\sigma^{\mathcal{A}}(s,w)\|\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w))). \tag{64}$$

Next, we rewrite the expression of interest as

$$\begin{aligned}
&\pm \operatorname{gen}_\rho(\mathcal{A}) && (65)\\
&= \pm \mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathrm{W}}}\left[\mathop{\mathbb{E}}_{\bar{S}\sim P^m}\left[\mathrm{Tr}[L(\bar{S},W)\left(\rho_{\mathrm{test}}(\bar{S})\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(\bar{S},W)\right)]\right] - \mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathrm{data}}|W}\left[\mathrm{Tr}[L(S,W)\sigma^{\mathcal{A}}(S,W)]\right]\right] && (66)\\
&= \mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathrm{W}}}\mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathrm{data}}|W}\left[\pm\left(\mathrm{Tr}[L(S,W)\left(\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,W)\right)] - \mathrm{Tr}[L(S,W)\sigma^{\mathcal{A}}(S,W)]\right)\right] && (67)\\
&\quad + \mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathrm{W}}}\Bigg[\pm\Bigg(\mathop{\mathbb{E}}_{\bar{S}\sim P^m}\left[\mathrm{Tr}[L(\bar{S},W)\left(\rho_{\mathrm{test}}(\bar{S})\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(\bar{S},W)\right)]\right]\\
&\qquad\qquad - \mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathrm{data}}|W}\left[\mathrm{Tr}[L(S,W)\left(\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,W)\right)]\right]\Bigg)\Bigg] && (68)
\end{aligned}$$

For the first summand, we can use Equations (63) and (64) to obtain:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathrm{W}}}\mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathrm{data}}|W}\left[\pm\left(\mathrm{Tr}[L(S,W)\left(\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,W)\right)] - \mathrm{Tr}[L(S,W)\sigma^{\mathcal{A}}(S,W)]\right)\right] && (69)\\
&\leq \mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathrm{W}}}\mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathrm{data}}|W}\left[\psi^{*-1}_{\mp}\left(D(\sigma^{\mathcal{A}}(S,W)\|\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,W))\right)\right]. && (70)
\end{aligned}$$

For the second summand, thanks to Equation (CMGF), we can apply (Jiao et al., 2017, Theorem 2) or (Bu et al., 2020, Theorem 1) (see also (Raginsky, 2019, p. 22) for a pedagogical presentation) to the classical random variable $\mathrm{Tr}[L(S,W)\left(\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,W)\right)]$ and obtain:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathrm{W}}}\Bigg[\pm\Bigg(\mathop{\mathbb{E}}_{\bar{S}\sim P^m}\left[\mathrm{Tr}[L(\bar{S},W)\left(\rho_{\mathrm{test}}(\bar{S})\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(\bar{S},W)\right)]\right]\\
&\qquad\qquad - \mathop{\mathbb{E}}_{S\sim P^{\mathcal{A}}_{\mathrm{data}}|W}\left[\mathrm{Tr}[L(S,W)\left(\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,W)\right)]\right]\Bigg)\Bigg] && (71)\\
&\leq \phi^{*-1}_{\mp}(I(S;W)). && (72)
\end{aligned}$$

Thus, we have shown the inequalities

$$\pm \operatorname{gen}_\rho(\mathcal{A}) \leq \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[\psi^{*-1}_{\mp}\left(D(\sigma^{\mathcal{A}}(S,W)\|\rho_{\mathrm{test}}(S)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(S,W))\right)\right] + \phi^{*-1}_{\mp}\left(I(S;W)\right). \tag{73}$$

As the $\psi_{\mp}^{*-1}$ are concave (since $\psi_{\mp}^{*}$ are convex), we can pull the expectation value inside the $\psi_{\mp}^{*-1}$ without making the right-hand side smaller, by Jensen's inequality. Then, it remains to observe that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}[D(\sigma^{\mathcal{A}}(S,W)\|\rho_{\text{test}}(S)\otimes\sigma_{\text{hyp}}^{\mathcal{A}}(s,W))] && (74)\\
&= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[-H(\sigma^{\mathcal{A}}(S,W)) + H(\sigma_{\text{hyp}}^{\mathcal{A}}(S,W)) - \operatorname{Tr}\left[\sigma_{\text{test}}^{\mathcal{A}}(S,W)\log\left(\rho_{\text{test}}(S)\right)\right]\right] && (75)\\
&= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\text{test};\text{hyp})_{\sigma^{\mathcal{A}}(S,W)} - H(\sigma_{\text{test}}^{\mathcal{A}}(S,W)) - \operatorname{Tr}\left[\sigma_{\text{test}}^{\mathcal{A}}(S,W)\log\left(\rho_{\text{test}}(S)\right)\right]\right] && (76)\\
&= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\text{test};\text{hyp})_{\sigma^{\mathcal{A}}(S,W)} - H(\rho_{\text{test}}^{\mathcal{A}}(S,W)) - \operatorname{Tr}\left[\rho_{\text{test}}^{\mathcal{A}}(S,W)\log\left(\rho_{\text{test}}(S)\right)\right]\right] && (77)\\
&= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\text{test};\text{hyp})_{\sigma(S,W)}\right] - \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[H(\rho_{\text{test}}^{\mathcal{A}}(S,W))\right] + \mathop{\mathbb{E}}_{S\sim P^m}\left[H(\rho_{\text{test}}(S))\right] && (78)\\
&= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\text{test};\text{hyp})_{\sigma(S,W)}\right] + \mathop{\mathbb{E}}_{S\sim P^m}\left[\chi\left(\left\{P_{\mathsf{W}|S}^{\mathcal{A}}(w),\rho_{\text{test}}^{\mathcal{A}}(S,w)\right\}_{w\in\mathsf{W}}\right)\right]. && (79)
\end{aligned}$$

Here, the third equality used that $\sigma_{\text{test}}^{\mathcal{A}}(s,w) = \rho_{\text{test}}^{\mathcal{A}}(s,w)$, because $\sigma^{\mathcal{A}}(s,w)$ and $\rho^{\mathcal{A}}(s,w)$ differ only by a CPTP map applied on the $\mathtt{train}$ subsystem. The fourth and fifth equalities used that $\mathbb{E}_{W\sim P_{\mathsf{W}|S}^{\mathcal{A}}}[\rho_{\text{test}}^{\mathcal{A}}(S,W)] = \rho_{\text{test}}(S)$. This holds because the state $\mathbb{E}_{W\sim P_{\mathsf{W}|S}^{\mathcal{A}}}[\rho^{\mathcal{A}}(S,W)]$ is obtained from $\rho(S)$ by applying the CPTP map $\text{id}_{\text{test}}\otimes\left(\sum_w\sqrt{E_S^{\mathcal{A}}(w)}(\cdot)\sqrt{E_S^{\mathcal{A}}(w)}\right)$, which acts non-trivially only on the training data register and thus leaves the test data marginal invariant. The fifth step also used Equation (22). Thus, after using Jensen to pull the expectation value inside $\psi_{\mp}^{*-1}$ and then rewriting the expected relative entropy as above, we have completed the proof. ■

Remark 18 *Our framework and Theorem 17 also encompass cases where classical and quantum side information can be generated during the learning process. If the risks and sub-gaussianity assumptions depend only on the data and hypothesis but not on the side information random variables and quantum registers, then we recover Equation (55). That is, despite having more objects to take into account, the final bound remains the same and in particular only depends on the data and the hypothesis, not on additional side information.*

Remark 19 *Having presented the proof of Theorem 17, we comment on some modifications. On the one hand, if we change the assumed Equation (QMGF) by allowing for (s,w)-dependent functions $\psi_{\pm;s,w}$, we can follow the same proof strategy. The obtained expected generalization error bound will differ from Equation (55) only in the first term on the r.h.s., which gets replaced by $\mathbb{E}_{(S,W)\sim P^{\mathcal{A}}}\left[\psi_{\mp;S,W}^{*-1}\left(D(\sigma^{\mathcal{A}}(S,W)\|\rho_{\text{test}}(S)\otimes\sigma_{\text{hyp}}^{\mathcal{A}}(S,W))\right)\right]$.*

On the other hand, if we change Equation (QMGF) to the (by Golden-Thompson weaker) assumption that

$$\operatorname{Tr}\left[e^{\log(\rho_{\text{test}}(s)\otimes\sigma_{\text{hyp}}^{\mathcal{A}}(s,w))+\lambda\left(L(s,w)-\operatorname{Tr}[L(s,w)(\rho_{\text{test}}(s)\otimes\sigma_{\text{hyp}}^{\mathcal{A}}(s,w))]\mathbb{1}_{\text{test},\text{hyp}}\right)}\right] \leq \begin{cases}\psi_+(\lambda) & \textit{if } \lambda\geq 0\\ \psi_-(\lambda) & \textit{if } \lambda<0\end{cases}, \tag{80}$$

we can still recover Equation (55). This can be seen by noticing that the second step in the proof of Theorem 17 was exactly to apply Golden-Thompson.

Finally, Theorem 17 and its proof simplify in different scenarios, for instance for learners that produce either only a classical or only a quantum hypothesis. Concretely, if W *is trivial, then we obtain a variant of Equation (55) without the Holevo information term and without the second summand on the r.h.s. In this case, the assumption Equation (CMGF) is not needed. Furthermore, if* $\mathcal{H}_{\mathrm{hyp}}$ *is trivial, then we obtain a variant of Equation (55) without the first summand on the right-hand side. In this case, the assumption Equation (QMGF) is not needed. Similarly, if* Z *is trivial, the second summand vanishes, whereas if* $\mathcal{H}_{\mathrm{data}}$ *is trivial, the first summand vanishes, so that we recover (Xu and Raginsky, 2017, Lemma 1). Moreover, if* $\sigma^{\mathcal{A}}(s,w) = \sigma^{\mathcal{A}}_{\mathrm{test}}(s,w) \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w)$ *is already a tensor product state – for example if each* $\rho^{\mathcal{A}}(s,w)$ *factorizes or if each* $E^{\mathcal{A}}_s(w)$ *is a pure state projector (so that monogamy of entanglements forbids the pure post-measurement state on the training system from being correlated or entangled with the test system) –, then we get a variant of Equation (55) without the QMI term. Finally, if* $\rho(s) = \rho_{\mathrm{test}}(s) \otimes \rho_{\mathrm{train}}(s)$ *factorizes, then both the QMI and the Holevo information contribution vanish and the assumption Equation (QMGF) is not needed.*

Remark 20 *As a consequence of (Berta et al., 2017, Lemma 1 and Theorem 2) – who applied Golden-Thompson in (Berta et al., 2017, Proposition 5) similarly to our use in the proof of Theorem 17 –, we have in fact established an expected generalization error bound in terms of measured quantum information quantities. Namely, relying on (Berta et al., 2017), we can tighten the initial inequality in our proof to*

$$D^{\mathbb{M}}(\sigma^{\mathcal{A}}(s,w) \| \rho_{\mathrm{test}}(s) \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w)) \tag{81}$$

$$\geq \lambda \operatorname{Tr}[L(s,w)\sigma^{\mathcal{A}}(s,w)] - \log \operatorname{Tr}\left[(\rho_{\mathrm{test}}(s) \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w)) \exp\left(\lambda L(s,w)\right)\right] , \tag{82}$$

where $D^{\mathbb{M}}(\rho\|\sigma)$ *denotes the measured relative entropy. The quantum relative entropy* $D(\rho\|\sigma)$ *upper bounds* $D^{\mathbb{M}}(\rho\|\sigma)$*, but there can be a gap between these two quantities.*

Example 4 (Example 2 continued) *The loss observables* $L(s,w) = \ell(s,w) \cdot \mathbb{1}_{\mathrm{test,hyp}}$ *considered in Example 2 trivially satisfy Equation (QMGF) even for* $\psi_{\pm}$ *given by the* 0*-function. With this choice,* $\psi^{*}_{\pm}(t) = +\infty$ *for all* t *and* $\psi^{*-1}_{\pm}(s) = 0$ *for all* s*, so the first term in our bound vanishes. Thus, Theorem 17 reproduces (Xu and Raginsky, 2017, Lemma 1) in this special case.*

Theorem 17 takes a particularly simple and appealing form if the assumptions on the moment-generating functions are sub-gaussianity assumptions. Before stating the corresponding result, we recall the notions of sub-gaussianity in the cases of observables and random variables:

Definition 21 (Sub-gaussianity for observables) *Let* $\alpha > 0$*. A self-adjoint loss observable* $L \in \mathcal{B}(\mathcal{H})$ *is called* α*-sub-gaussian with respect to a quantum state* $\sigma \in \mathcal{S}(\mathcal{H})$ *if*

$$\log \operatorname{Tr}\left[\sigma \cdot e^{\lambda(L - \operatorname{Tr}[L\sigma]\mathbb{1})}\right] \leq \frac{\alpha^2\lambda^2}{2} \tag{83}$$

holds for all $\alpha \in \mathbb{R}$*.*

Example 5 *Quantum concentration inequalities recently received considerable attention in the literature. Prominent examples of classes of states for which bounds on the MGF are known include the following:*

1. *Local observables w.r.t. high-temperature Gibbs states (Kuwahara and Saito, 2020) and, more generally, Lipschitz observables w.r.t. high temperature commuting Gibbs states (De Palma and Rouzé, 2022; Capel et al., 2020) or $1D$-commuting Gibbs states (Bardet et al., 2021), are known to satisfy sub-gaussianity with $\alpha = \mathcal{O}(1)$.*

2. *Local observables w.r.t. outcomes of shallow circuits also satisfy sub-gaussianity with $\alpha = \mathcal{O}(1)$ (Anshu and Metger, 2023).*

3. *Lipschitz observables w.r.t. tensor product states, up to a weakening à la Golden-Thompson analogously to Equation (80), satisfy sub-gaussianity with $\alpha = \mathcal{O}(1)$ (De Palma and Trevisan, 2023, Theorem 8.1).*

4. *More generally, (Anshu, 2016) proved concentration bounds for local observables w.r.t. states with finite correlation length by bounding the MGF. However, they are weaker than sub-gaussian concentration and depend on the dimension of the underlying lattice.*

Definition 22 (Sub-gaussianity for random variables (Vershynin, 2018, Section 2.5)) *Let $\alpha > 0$. A real-valued random variable X is α-sub-gaussian if*

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \frac{\alpha^2\lambda^2}{2} \tag{84}$$

holds for all $\alpha \in \mathbb{R}$.

Example 6 *Trivially, a gaussian random variable with variance β^2 is β-sub-gaussian. By Hoeffding's Lemma (Hoeffding, 1963), any random variable that almost surely takes values in a bounded interval $[a,b]$ is $(\frac{b-a}{2})$-sub-gaussian. Finally, any L-Lipschitz function of a Haar-random variable on the unit sphere in $\mathbb{R}^n$ is $(\frac{CL}{\sqrt{n}})$-sub-gaussian for a suitable $C > 0$ (see, e.g., (Vershynin, 2018, Chapter 5)).*

With these definitions, we can now compactly state the sub-gaussian versions of Theorem 17:

Corollary 23 *Let $\alpha, \beta > 0$. Assume that the loss observable $L(s,w)$ is α-sub-gaussian w.r.t. $\rho_{\text{test}}(s) \otimes \sigma^{\mathcal{A}}_{\text{hyp}}(s,w)$ for every $(s,w) \in \mathsf{Z}^m \times \mathsf{W}$. Moreover, assume that the random variable $\operatorname{Tr}[L(S,w)(\rho_{\text{test}}(S) \otimes \sigma^{\mathcal{A}}_{\text{hyp}}(S,w))]$, with $S \sim P^m$, is β-sub-gaussian for every $w \in \mathsf{W}$. Then,*

$$\begin{aligned}|\mathrm{gen}_\rho(\mathcal{A})| \leq &\sqrt{2\alpha^2\left(\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}[I(\text{test};\text{hyp})_{\sigma(S,W)}] + \mathop{\mathbb{E}}_{S\sim P^m}\left[\chi\left(\{P^{\mathcal{A}}_{\mathsf{W}|S}(w), \rho^{\mathcal{A}}_{\text{test}}(S,w)\}_{w\in\mathsf{W}}\right)\right]\right)}\\ &+\sqrt{2\beta^2 I(S;W)}\,.\end{aligned} \tag{85}$$

Proof This follows from Theorem 17 with the log-MGF bounds $\psi_\pm : \mathbb{R} \to \mathbb{R}$, $\psi_\pm(x) = \frac{\alpha^2x^2}{2}$ and $\phi_\pm : \mathbb{R} \to \mathbb{R}$, $\phi_\pm(x) = \frac{\beta^2x^2}{2}$. This leads to $\psi_\pm^{*-1}(\xi) = \sqrt{2\alpha^2\xi}$ and $\phi_\pm^{*-1}(\xi) = \sqrt{2\beta^2\xi}$. ■

So far, our generalization error bounds do not explicitly depend on the training data size m. To achieve such a dependence, we now impose an i.i.d. structure on the quantum data, in addition to the already assumed (but not yet fully exploited) i.i.d. structure on the classical training data $S \sim P^m$. Namely, we assume that the data Hilbert space and states factorize as

$$\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{test}} \otimes \mathcal{H}_{\text{train}} = \bigotimes_{i=1}^{m} (\mathcal{H}_{\text{test},i} \otimes \mathcal{H}_{\text{train},i}) = \bigotimes_{i=1}^{m} \mathcal{H}_{\text{data},i}\,, \tag{86}$$

$$\rho(s) = \rho(z_1, \dots, z_m) = \bigotimes_{i=1}^{m} \rho_i(z_i)\,, \quad \text{with } \rho_i(z_i) \in \mathcal{S}(\mathcal{H}_{\text{test},i} \otimes \mathcal{H}_{\text{train},i})\,. \tag{87}$$

For our next result, we consider learners and loss observables that adhere to this factorization. On the one hand, we assume that the POVMs and channels used by the learner $\mathcal{A}$ factorize as $E_s^{\mathcal{A}}(w) = E_{z_1,\dots,z_m}^{\mathcal{A}}(w) = \bigotimes_{i=1}^{m} E_{z_i}^{\mathcal{A}}(w)$ and $\Lambda_{s,w}^{\mathcal{A}} = \Lambda_{z_1,\dots,z_m,w}^{\mathcal{A}} = \bigotimes_{i=1}^{m} \Lambda_{z_i,w}^{\mathcal{A}}$ with $E_{z_i}^{\mathcal{A}}(w) \in \mathcal{E}(\mathcal{H}_{\text{train},i})$ and $\Lambda_{z_i,w}^{\mathcal{A}} : \mathcal{T}_1(\mathcal{H}_{\text{train},i}) \to \mathcal{T}_1(\mathcal{H}_{\text{hyp},i})$. Note that this in particular comes with factorizations $\mathcal{H}_{\text{hyp}} = \bigotimes_{i=1}^{m} \mathcal{H}_{\text{hyp},i}$ of the hypothesis Hilbert space and $\sigma^{\mathcal{A}}(s,w) = \bigotimes_{i=1}^{m} \sigma_i^{\mathcal{A}}(z_i,w)$ of the state after the action of $\mathcal{A}$, with $\sigma_i^{\mathcal{A}}(z_i,w) \in \mathcal{S}(\mathcal{H}_{\text{test},i} \otimes \mathcal{H}_{\text{hyp},i})$. On the other hand, we assume the loss observables to be of the local form $L(s,w) = \frac{1}{m}\sum_{i=1}^{m} L_i(z_i,w)$, with $L_i(z_i,w) \in \mathcal{B}(\mathcal{H}_{\text{test},i} \otimes \mathcal{H}_{\text{hyp},i})$ acting only on the ith test and hypothesis subsystems. (For readability, we notationally suppress identities on the remaining subsystems when convenient.) In this setting, Corollary 23 gives the following result:

Corollary 24 *Assume the above factorization for the quantum data and the learner $\mathcal{A}$ as well as the above local structure of the loss observables. Moreover, assume that $L_i(z_i,w)$ is α_i-sub-gaussian w.r.t. $\rho_{\text{test},i}(z_i) \otimes \sigma_{\text{hyp},i}^{\mathcal{A}}(z_i,w)$ for every $(z_i,w) \in \mathsf{Z} \times \mathsf{W}$ and $1 \leq i \leq m$, and that the random variable $\operatorname{Tr}\left[L_i(Z_i,w)(\rho_{\text{test},i}(Z_i) \otimes \sigma_{\text{hyp},i}^{\mathcal{A}}(Z_i,w))\right]$, with $Z_i \sim P$, is β_i-sub-gaussian for every $w \in \mathsf{W}$ and $1 \leq i \leq m$. Then,*

$$|\text{gen}_\rho(\mathcal{A})| \leq \sqrt{\frac{2\sum_{i=1}^{m} \alpha_i^2}{m^2}\left(\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[\sum_{i=1}^{m} I(\text{test};\text{hyp})_{\sigma_i^{\mathcal{A}}(Z_i,W)}\right] + \mathop{\mathbb{E}}_{S\sim P^m}\left[\chi\left(\{P_{\mathsf{W}|S}^{\mathcal{A}}(w), \rho_{\text{test}}^{\mathcal{A}}(S,w)\}_{w\in\mathsf{W}}\right)\right]\right)} + \sqrt{\frac{2\sum_{i=1}^{m}\beta_i^2}{m^2} I(S;W)}\,. \tag{88}$$

In particular, if $\alpha_i = \alpha_0$ and $\beta_i = \beta_0$ for all $1 \leq i \leq m$, then

$$|\text{gen}_\rho(\mathcal{A})| \leq \sqrt{\frac{2\alpha_0^2}{m}\left(\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[\sum_{i=1}^{m} I(\text{test};\text{hyp})_{\sigma_i^{\mathcal{A}}(Z_i,W)}\right] + \mathop{\mathbb{E}}_{S\sim P^m}\left[\chi\left(\{P_{\mathsf{W}|S}^{\mathcal{A}}(w), \rho_{\text{test}}^{\mathcal{A}}(S,w)\}_{w\in\mathsf{W}}\right)\right]\right)} + \sqrt{\frac{2\beta_0^2}{m} I(S;W)}\,. \tag{89}$$

Proof See Appendix D. ■

We point out that the factorization assumption on the POVM elements $E_s^{\mathcal{A}}(w)$ is not needed if $\mathcal{A}$ produces only a classical hypothesis. In this case, the hyp quantum system is trivial. Thus, $\rho_{\text{test}}(s) \otimes \sigma_{\text{hyp}}^{\mathcal{A}}(s,w) = \rho_{\text{test}}(s) = \bigotimes_{i=1}^{m} \rho_{\text{test},i}(z_i)$ factorizes by assumption, which is sufficient for the proof of Corollary 24. Even in this setting, the Holevo information term in the bound is an in general non-trivial quantum contribution.

If, however, the learner produces a non-trivial quantum hypothesis, our current proof strategy does rely on the factorization assumption. Notice, however, that Example 5 already contains QMGF bounds w.r.t. non-product states. Thus, insights into CMGF bounds w.r.t. non-product states may allow future work to improve upon our proof of Corollary 24, extending it to more general (non-product) learners.

Let us return to our continuing example of quantum state classification and see the implications of our generalization bounds in that setting.

Example 7 (Quantum state classification – Examples 1 and 3 continued) *As the learner $\mathcal{A}$ in our quantum state classification example produces only a classical hypothesis and as the initial quantum data states $\rho(s)$ factorize across the test-train bipartition, it suffices to verify a suitable classical sub-gaussianity assumption. Observe that, for every $(s,w) \in \{0,1\}^m \times \mathsf{W}$, the state*

$$\rho_{\text{test}}(s) = \mathop{\mathbb{E}}_{\{(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m \sim P_{\text{pair}}^m}\left[\bigotimes_{i=1}^m \sigma_{z_i}^{(i)}\right] = \bigotimes_{i=1}^m \mathop{\mathbb{E}}_{\{(\sigma_0^{(i)},\sigma_1^{(i)})\}_{i=1}^m \sim P_{\text{pair}}^m}\left[\sigma_{z_i}^{(i)}\right] = \bigotimes_{i=1}^m \rho_{\text{test},i}(z_i) \tag{90}$$

is an m-fold tensor product. Moreover, the loss observables defined in Example 3 are local w.r.t. this tensor factorization. So, to apply Corollary 24, we consider the sub-gaussianity parameter β_0 of the random variable $\operatorname{Tr}[((1-Z_i)(\mathbb{1}_d - F(w)) + Z_i F(w))\rho_{\text{test}}(Z_i)]$, with $Z_i \sim P$. Without any prior assumptions on the distribution P and on the mapping $z \mapsto \rho(z)$, the random variable of interest takes values in $[0,1]$ because $0 \leq F(w), 1-F(w) \leq 1$. Thus, Hoeffding's Lemma (Hoeffding, 1963) implies $\beta_0 \leq 1/2$ for every $w \in \mathsf{W}$. Therefore, Corollary 24 yields

$$|R_\rho(\mathcal{A}) - \hat{R}_\rho(\mathcal{A})| \leq \sqrt{\frac{1}{2m} I(S;W)}\,. \tag{91}$$

If W is finite, then we immediately have the mutual information upper bound $I(S;W) \leq \log|\mathsf{W}|$. Hence, our above bound implies that we can guarantee a small expected generalization error as soon as the training data size m is of the same order as the number of bits needed to describe the classical hypotheses. If W is infinite, we may first discretize and then apply the bound. Concretely, if $\varepsilon > 0$ and if $\mathsf{W}_\varepsilon \subseteq \mathsf{W}$ is an ε-covering net for W w.r.t. the sup-norm, then $|R_\rho(\mathcal{A}) - \hat{R}_\rho(\mathcal{A})| \leq \varepsilon + \sqrt{\frac{1}{2m}\log|\mathsf{W}_\varepsilon|}$. If there are no prior assumptions on the admissible effect operators $\{F(w)\}_{w\in\mathsf{W}}$, then we cannot expect better bounds on the cardinality of an ε-covering net for W than $\log|\mathsf{W}_\varepsilon| \leq \tilde{\mathcal{O}}\left(\min\{d/\varepsilon^2, d^2\log(1/\varepsilon)\}\right)$ (Cheng et al., 2016, Section 4)[4]. In the case of n qubits, we have $d = 2^n$ and the resulting bound scales exponentially with n. This can be improved if $\{F(w)\}_{w\in\mathsf{W}}$ is limited. For example, if $F(w)$ is a sum of k-local Pauli terms for every $w \in \mathsf{W}$, where $k = \mathcal{O}(1)$, then, since there are at most $\mathcal{O}(n^k)$ such terms, one can obtain an improved covering number bound of $\log|\mathsf{W}_\varepsilon| \leq \tilde{\mathcal{O}}(n^k \log(1/\varepsilon))$, which scales polynomially in n. This can be improved further if the locality assumption is strengthened to geometric locality. Note that these bounds on $I(S;W)$ are worst-case and we expect tighter algorithm-dependent bounds to be possible when taking the POVMs $\{E_s^{\mathcal{A}}(w)\}_{w\in\mathsf{W}}$ chosen by the learner into account. This concludes the discussion of our state classification example.

[4] Here, the $\tilde{\mathcal{O}}$ hides non-leading logarithmic factors.

As it concerns a special case with only a classical hypothesis, Equation (91) can already be deduced from the classical generalization bounds of (Xu and Raginsky, 2017). In the next section, we demonstrate the applicability of our general framework and our generalization error bounds for a variety of quantum learning problems, including scenarios that cannot be studied with the purely classical framework. Before this discussion, we conclude this section with an extension of Corollary 24 to stable learners that use channels leading to a controlled increase of Lipschitz constants:

Corollary 25 *Assume the above factorization for the quantum data and the POVMs used by the learner as well as the above local structure for the loss observables. Furthermore, assume that the Heisenberg picture duals* $(\Lambda^{\mathcal{A}}_{s,w})^*$ *of the channels* $\Lambda^{\mathcal{A}}_{s,w}$ *used by* $\mathcal{A}$ *satisfy* $\|(\Lambda^{\mathcal{A}}_{s,w})^*\|_{\mathrm{Lip}\to\mathrm{Lip}} \leq C_1$ *as well as* $\max_{s\sim s',w}\|(\Lambda^{\mathcal{A}}_{s,w} - \Lambda^{\mathcal{A}}_{s',w})^*\|_{\mathrm{Lip}\to\infty} \leq C_2$*, where* $s \sim s'$ *denotes neighboring training data sets (i.e., training data sets that differ only in a single data point). Then,*

$$\begin{aligned}
&|\mathrm{gen}_\rho(\mathcal{A})| \\
&\leq \frac{2\sqrt{2}\max\limits_{i,z_i,w}\|L_i(z_i,w)\|}{\sqrt{m}}\left(\sqrt{C_1\left(\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}(S,W)}\right] + \mathop{\mathbb{E}}_{S\sim P^m}\left[\chi\left(\{P^{\mathcal{A}}_{\mathsf{W}|S}(w), \rho^{\mathcal{A}}_{\mathrm{test}}(S,w)\}_{w\in\mathsf{W}}\right)\right]\right)} \right.\\
&\qquad\qquad\qquad \left. + \sqrt{(1+C_1(1+C_2))I(S;W)}\right).
\end{aligned} \tag{92}$$

In the assumed bound $\|(\Lambda^{\mathcal{A}}_{s,w})^*\|_{\mathrm{Lip}\to\mathrm{Lip}} \leq C_1$, the Lipschitz constants considered are w.r.t. the factorizations $\mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{hyp}} = \bigotimes_{i=1}^m(\mathcal{H}_{\mathrm{test},i} \otimes \mathcal{H}_{\mathrm{hyp},i})$ and $\mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{train}} = \bigotimes_{i=1}^m(\mathcal{H}_{\mathrm{test},i} \otimes \mathcal{H}_{\mathrm{train},i})$. Similarly, the Lipschitz constants relevant for the stability assumption $\max_{s\sim s',w}\|(\Lambda^{\mathcal{A}}_{s,w} - \Lambda^{\mathcal{A}}_{s',w})^*\|_{\mathrm{Lip}\to\infty} \leq C_2$ are w.r.t. $\mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{hyp}} = \bigotimes_{i=1}^m(\mathcal{H}_{\mathrm{test},i} \otimes \mathcal{H}_{\mathrm{hyp},i})$. Again, the POVM factorization assumption is not needed if the learner only produces a classical hypothesis.

Proof Recall from Remark 14 that we obtain the same notions of risk when absorbing the channels $\Lambda^{\mathcal{A}}_{s,w}$ into the loss observables via the Heisenberg picture. Thus, instead of proving sub-gaussianity of $L(s,w)$ w.r.t. $\rho_{\mathrm{test}}(s) \otimes \sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w)$, we can also establish sub-gaussianity of $(\Lambda^{\mathcal{A}}_{s,w})^*(L(s,w))$ w.r.t. $\rho_{\mathrm{test}}(s) \otimes \rho^{\mathcal{A}}_{\mathrm{train}}(s,w)$. We do this in the first part of the proof. As $\|(\Lambda^{\mathcal{A}}_{s,w})^*\|_{\mathrm{Lip}\to\mathrm{Lip}} \leq C_1$, we have

$$\|(\Lambda^{\mathcal{A}}_{s,w})^*(L(s,w))\|_{\mathrm{Lip}} \leq C_1\|L(s,w)\|_{\mathrm{Lip}} \leq \frac{2C_1\max_{i,z_i,w}\|L_i(z_i,w)\|}{m}, \tag{93}$$

where the last step used (De Palma et al., 2021, Proposition 8). Therefore, according to (De Palma and Trevisan, 2023, Theorem 8.1), which we restate as Lemma 30, the observable $(\Lambda^{\mathcal{A}}_{s,w})^*(L(s,w))$ satisfies a version of $(m^{-1/2} \cdot 2C_1\max_{i,z_i,w}\|L_i(z_i,w)\|)$-sub-gaussianity w.r.t. the m-fold tensor product $\bigotimes_{i=1}^m \rho_{\mathrm{test},i}(z_i) \otimes \rho^{\mathcal{A}}_{\mathrm{train},i}(z_i,w)$ weakened analogously to Equation (80). As argued in Remark 19, this weaker version is a sufficient quantum sub-gaussianity for our purposes.

Next, we establish a suitable classical sub-gaussianity. To this end, take two training data sets $s = (z_1,\ldots,z_m)$, $s' = (z_1',\ldots,z_m') \in \mathsf{Z}^m$ that differ in exactly one data point, i.e., $\exists 1 \leq i \leq m$ such that $z_i \neq z_i'$ and $z_j = z_j'$ for all $j \neq i$. For this relation, we use the shorthand $s \sim s'$. Then, because of our assumed factorization of the quantum data states and of the POVMs used by the learner, the post-measurement states $\rho_{\mathrm{test}}(s) \otimes \rho^{\mathcal{A}}_{\mathrm{train}}(s,w)$ and $\rho_{\mathrm{test}}(s) \otimes \rho_{\mathrm{train}}(s',w)$ agree after tracing out the ith subsystem, i.e., $\mathrm{Tr}_{\mathrm{test},i;\mathrm{hyp},i}[\rho_{\mathrm{test}}(s) \otimes \rho^{\mathcal{A}}_{\mathrm{train}}(s,w)] = \mathrm{Tr}_{\mathrm{test},i;\mathrm{hyp},i}[\rho_{\mathrm{test}}(s') \otimes$

$\rho^{\mathcal{A}}_{\text{train}}(s', w)]$ for all $w \in \mathsf{W}$. Hence, by definition of the quantum Lipschitz constant (compare (De Palma et al., 2021, Definition 8)), we obtain the bound

$$\left|\mathrm{Tr}[L(s,w)\left(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w)\right)] - \mathrm{Tr}[L(s',w)\left(\rho_{\text{test}}(s')\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s',w)\right)]\right| \tag{94}$$

$$= \left|\mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s,w})^*(L(s,w))\left(\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\right)] - \mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s',w})^*(L(s',w))\left(\rho_{\text{test}}(s')\otimes\rho^{\mathcal{A}}_{\text{train}}(s',w)\right)]\right| \tag{95}$$

$$\leq \left|\mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s,w})^*(L(s,w))\left(\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\right)] - \mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s,w})^*(L(s',w))\left(\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\right)]\right| \tag{96}$$

$$+ \left|\mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s,w})^*(L(s',w))\left(\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\right)] - \mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s',w})^*(L(s',w))\left(\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\right)]\right| \tag{97}$$

$$+ \left|\mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s',w})^*(L(s',w))\left(\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\right)] - \mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s',w})^*(L(s',w))\left(\rho_{\text{test}}(s')\otimes\rho^{\mathcal{A}}_{\text{train}}(s',w)\right)]\right| \tag{98}$$

$$\leq \left|\mathrm{Tr}[(L(s,w)-L(s',w))\left(\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w)\right)]\right| \tag{99}$$

$$+ \left|\mathrm{Tr}[(\Lambda^{\mathcal{A}}_{s,w}-\Lambda^{\mathcal{A}}_{s',w})^*(L(s',w))\left(\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\right)]\right| \tag{100}$$

$$+ \frac{2C_1 \max_{i,z_i,w}\|L_i(z_i,w)\|}{m} \tag{101}$$

$$\leq \|L(s,w)-L(s',w)\|\cdot\|\rho_{\text{test}}(s)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(s,w)\|_1 \tag{102}$$

$$+ \|(\Lambda^{\mathcal{A}}_{s,w}-\Lambda^{\mathcal{A}}_{s',w})^*(L(s',w))\|\cdot\|\rho_{\text{test}}(s)\otimes\rho^{\mathcal{A}}_{\text{train}}(s,w)\|_1 \tag{103}$$

$$+ \frac{2C_1 \max_{i,z_i,w}\|L_i(z_i,w)\|}{m} \tag{104}$$

$$\leq \frac{2\max_{i,z_i,w}\|L_i(z_i,w)\|}{m} + \frac{2C_1 \max_{i,z_i,w}\|L_i(z_i,w)\|}{m}\cdot\max_{s\sim s',w}\|(\Lambda^{\mathcal{A}}_{s,w}-\Lambda^{\mathcal{A}}_{s',w})^*\|_{\mathrm{Lip}\to\infty} \tag{105}$$

$$+ \frac{2C_1 \max_{i,z_i,w}\|L_i(z_i,w)\|}{m} \tag{106}$$

$$= \frac{2\max_{i,z_i,w}\|L_i(z_i,w)\|}{m}\left(1 + C_1\left(1+\max_{s\sim s',w}\|(\Lambda^{\mathcal{A}}_{s,w}-\Lambda^{\mathcal{A}}_{s',w})^*\|_{\mathrm{Lip}\to\infty}\right)\right) \tag{107}$$

$$\leq \frac{2\max_{i,z_i,w}\|L_i(z_i,w)\|}{m}\left(1+C_1\left(1+C_2\right)\right). \tag{108}$$

Therefore, the random variable $\mathrm{Tr}[L(S,w)\left(\rho_{\text{test}}(S)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(S,w)\right)]$, with $S\sim P^m$, is sub-gaussian with sub-gaussianity parameter $\left(m^{-1/2}\cdot 2\max_{i,z_i,w}\|L_i(z_i,w)\|\left(1+C_1\left(1+C_2\right)\right)\right)$, by McDiarmid's bounded differences inequality (McDiarmid, 1989).

We can now apply Corollary 23 with the classical and quantum sub-gaussianity parameters established above and obtain the claimed generalization bound. ■

A short discussion of the assumptions made on the channels $\Lambda^{\mathcal{A}}_{s,w}$ is in order. On the one hand, we assume that their Heisenberg duals $(\Lambda^{\mathcal{A}}_{s,w})^*$ lead to a bounded increase in quantum Lipschitz constants, namely that $\|(\Lambda^{\mathcal{A}}_{s,w})^*\|_{\mathrm{Lip}\to\mathrm{Lip}}\leq C_1$. Equivalently, the maps $\Lambda^{\mathcal{A}}_{s,w}$ should lead to a limited increase of quantum Wasserstein-1 norms, that is, $\|\Lambda^{\mathcal{A}}_{s,w}\|_{W_1\to W_1}\leq C_1$. This is satisfied for approximately locality-preserving channels such as constant-depth circuits or short-time evolutions under a local Lindblad generator (De Palma and Trevisan, 2023; De Palma et al., 2023)(with associated Lieb-Robinson bound). Also, as the proof of (De Palma and Trevisan, 2023, Theorem 8.1) shows, this property is satisfied with $C_1 = 1$ for m-fold tensor products of single-qudit channels. Moreover, for channels described by quantum circuits with local depolarizing noise, we obtain a

C_1 that decays exponentially with the circuit depth for large enough noise strength compared to the size of the light-cone of each layer (compare the proof of (Hirche et al., 2023, Proposition IV.8.)).

On the other hand, we assume that $\|(\Lambda^{\mathcal{A}}_{s,w} - \Lambda^{\mathcal{A}}_{s',w})^*\|_{\mathrm{Lip}\to\infty} \leq C_2$ for any neighboring data sets $s \sim s'$. Note that, as a consequence of (De Palma et al., 2021, Proposition 9) we can rewrite this as $\|(\Lambda^{\mathcal{A}}_{s,w} - \Lambda^{\mathcal{A}}_{s',w})^*\|_{\mathrm{Lip}\to\infty} = \|\Lambda^{\mathcal{A}}_{s,w} - \Lambda^{\mathcal{A}}_{s',w}\|_{1\to W_1} \leq C_2$. This is a stability assumption: When the two classical data sets s, s' differ in only a single data point, the quantum channels $\Lambda^{\mathcal{A}}_{s,w}$ and $\Lambda^{\mathcal{A}}_{s',w}$ employed by the learner $\mathcal{A}$ must not differ too much. It is reminiscent of classical replace-one stability (Bousquet and Elisseeff, 2000, 2002; Shalev-Shwartz et al., 2010). Using (De Palma et al., 2021, Corollary 2), we see that this quantum stability assumption is for example satisfied if, for every $s \sim s'$ and for every w, we can write $\Lambda^{\mathcal{A}}_{s,w} - \Lambda^{\mathcal{A}}_{s',w} = (\mathcal{N}_{s,w} - \mathcal{N}_{s',w})\mathcal{M}_w$ with $\mathcal{M}_w$ an arbitrary CPTP map and with CPTP maps $\mathcal{N}_{s,w}, \mathcal{N}_{s',w}$ that act non-trivially only on a constant number of training data subsystems. As this is in particular satisfied for learners that factorize, we can indeed view Corollary 25 as an extension of Corollary 24.

One strength of the results presented in this section is that they encompass a variety of learning tasks. However, when applied to a specific scenario, they do not necessarily lead to optimal bounds. For instance, our bounds in Corollaries 24 and 25 have a "slow rate" of $1/\sqrt{m}$, which is to be contrasted with the "fast rate" of $1/m$ recently achieved by, among others, (Hellström and Durisi, 2021; Grunwald et al., 2021; Wang and Mao, 2023) for classical information-theoretic generalization bounds and by (Mai and Alquier, 2017) in the context of PAC-Bayesian quantum state tomography w.r.t. squared Frobenius norm. We leave proving improved quantum information-theoretic generalization bounds with fast rates to future work.

Appendix C. Applications

C.1. PAC learning quantum states

For our first application, we consider a setting of PAC learning quantum states, going back to (Aaronson, 2007). Here, the goal is to predict expectation values w.r.t. an unknown state on average over an unknown distribution over effect operators. Take the data Hilbert space

$$\mathcal{H}_{\mathrm{data}} = \mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{train}} = ((\mathbb{C}^d)^{\otimes m_{\mathrm{test}}})^{\otimes m} \otimes ((\mathbb{C}^d)^{\otimes m_{\mathrm{train}}})^{\otimes m} \tag{C.1.1}$$

for some $d \in \mathbb{N}$ and $m, m_{\mathrm{test}}, m_{\mathrm{train}} \in \mathbb{N}$. Let the quantum data state ρ be the CQ state given by

$$\rho = \mathop{\mathbb{E}}_{S=(Z_1,\ldots,Z_{2m})\sim P^{2m}} \left[\left(\bigotimes_{i=1}^{2m} |Z_i\rangle\langle Z_i| \right) \otimes (\rho_0^{\otimes m_{\mathrm{test}}})^{\otimes m} \otimes (\rho_0^{\otimes m_{\mathrm{train}}})^{\otimes m} \right], \tag{C.1.2}$$

where $\rho_0 \in \mathcal{S}(\mathbb{C}^d)$ is the unknown qudit state to be PAC-learned, we imagine that each $z \in \mathsf{Z}$ comes with an associated qudit effect operator $E(z) \in \mathcal{E}(\mathbb{C}^d)$, and P is an unknown probability distribution over Z. That is, the CQ data consists of independent copies of an unknown state that we are trying to learn, as well as of (classical descriptions of) random two-outcome POVM measurements drawn i.i.d. from P.

We describe a simple quantum learner $\mathcal{A}$ for this scenario as follows: Take $\mathcal{H}_{\mathrm{hyp}}$ to be trivial, and take W to be some measurable hypothesis space. Here, we imagine each classical hypothesis $w \in \mathsf{W}$ to be associated to some hypothesis state $\rho_0(w) \in \mathcal{S}(\mathbb{C}^d)$ that the learner could output. Upon seeing the classical data $s = (z_1, \ldots, z_{2m}) \in \mathsf{Z}^{2m}$, the learner performs a two-step procedure:

Let $\tilde{\varepsilon} > 0$ be an auxiliary accuracy parameter, which we determine later. First, the learner takes $\mathsf{W}_1 \subset \mathsf{W}$ to be a $\tilde{\varepsilon}$-covering of the hypothesis space W w.r.t. the empirical seminorm $\|\cdot\|_{2,\{z_j\}_{j=1}^m}$ defined as

$$\|w\|_{2,\{z_j\}_{j=1}^m} = \sqrt{\frac{1}{m}\sum_{j=1}^m |\mathrm{Tr}[E(z_j)\rho_0(w)]|^2}\,. \tag{C.1.3}$$

Second, for each $m+1 \leq i \leq 2m$, the learner measures the 2-outcome POVM $\{E(z_i), \mathbb{1} - E(z_i)\}$ separately on m_{train} copies of ρ_0, obtaining outcomes $b_\ell^{(i)}, 1 \leq \ell \leq m_{\mathrm{train}}$, and then uses the empirical average $\tilde{b}^{(i)} := \frac{1}{m_{\mathrm{train}}}\sum_{\ell=1}^{m_{\mathrm{train}}} b_\ell^{(i)}$ as an estimate of $\mathrm{Tr}[E(z_i)\rho_0]$. The quantum learner then outputs an empirical risk minimizing hypothesis

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathsf{W}_1} \frac{1}{m}\sum_{i=m+1}^{2m} \left|\mathrm{Tr}[E(z_i)\rho_0(w)] - \tilde{b}^{(i)}\right| =: \operatorname*{argmin}_{w\in\mathsf{W}_1} \hat{R}^{\mathrm{train}}_{s_{(m+1):2m}, b_\ell^{(i)}}(w). \tag{C.1.4}$$

If there are multiple empirical risk minimizers, the tie is broken arbitrarily (but, for simplicity of notation, deterministically). Note: Both building the empirical covering net and performing empirical risk minimization over that net are computationally inefficient in general. Here, we focus on information-theoretic aspects and ignore computational complexity.

As in Example 1, the family of quantum channels associated to this quantum learner is trivial, since there is no quantum hypothesis. Thus, following Equation (33), when letting the learner $\mathcal{A}$ act on the quantum data state ρ, we obtain the output state

$$\sigma = \mathop{\mathbb{E}}_{W\sim P_{\mathsf{W}}^{\mathcal{A}}} \mathop{\mathbb{E}}_{S\sim P_{\mathrm{data}}^{\mathcal{A}}|W} [|S\rangle\langle S| \otimes \rho_{\mathrm{test}} \otimes |\hat{w}\rangle\langle\hat{w}|]\,, \tag{C.1.5}$$

with quantum test state $\rho_{\mathrm{test}} = (\rho_0^{\otimes m_{\mathrm{test}}})^{\otimes m}$ and with the probability distribution $P^{\mathcal{A}}$ on $\mathsf{Z}^m \times \mathsf{W}$ given by

$$P^{\mathcal{A}}(s,\hat{w}) = P^m(s)\cdot P^{\mathcal{A}}(\hat{w}|s) = P^m(s)\cdot \mathbb{P}_{B_\ell^{(i)}|s}\left[\hat{w} \in \operatorname*{argmin}_{w\in\mathsf{W}_1} \hat{R}^{\mathrm{train}}_{s_{(m+1):2m}, B_\ell^{(i)}}(w)\right], \tag{C.1.6}$$

where the $B_\ell^{(i)}$ are $\{0,1\}$-valued random variables which become independent when conditioned on s, with probability distributions

$$\mathbb{P}_{B_\ell^{(i)}|s}[B_\ell^{(i)} = 1] = \mathrm{Tr}[E(z_i)\rho_0] = 1 - \mathbb{P}_{B_\ell^{(i)}|s}[B_\ell^{(i)} = 0]$$

for all $1 \leq \ell \leq m_{\mathrm{train}}$ and for all $m+1 \leq i \leq 2m$. While more general quantum learners are possible, for instance by allowing for general s-dependent POVM elements, the simple quantum learner presented here is similar in spirit to (Aaronson, 2007) and (Xu and Raginsky, 2017, Section 4.2). As we show below, we can make guarantees on its performance based on Corollary 24.

Given that our quantum learner is based on empirical risk minimization, we define the loss observables in analogy to the notion of empirical risk used above. Namely, for each $1 \leq i \leq m$ and $c_\ell^{(i)}$, we set

$$L^{(i)}_{c_\ell^{(i)}}(z_i, w) = L^{(i)}_{c_\ell^{(i)}}(z_i) = \bigotimes_{\ell=1}^{m_{\mathrm{test}}} \left(c_\ell^{(i)} E(z_i) + (1 - c_\ell^{(i)})(\mathbb{1}_d - E(z_i))\right) \tag{C.1.7}$$

and

$$L(s,w) = L(s) = \sum_{c_\ell^{(i)} \in \{0,1\}} \hat{R}^{\text{test}}_{s_{(m+1):2m}, c_\ell^{(i)}}(w) \cdot (\mathbb{1}_d^{\otimes m_{\text{test}}})^{\otimes(i-1)} \otimes L^{(i)}_{c_\ell^{(i)}}(z_i) \otimes (\mathbb{1}_d^{\otimes m_{\text{test}}})^{\otimes(m-i)}, \tag{C.1.8}$$

with $\hat{R}^{\text{test}}_{s_{(m+1):2m}, c_\ell^{(i)}}(w) = \frac{1}{m}\sum_{i=m+1}^{2m} \left| \text{Tr}[E(z_i)\rho_0(w)] - \frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}} c_\ell^{(i)} \right|$, Plugging these choices into Definition 11, we obtain the expected empirical risk

$$\hat{R}_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{S\sim P^{2m}} \mathop{\mathbb{E}}_{C_\ell^{(i)}|S} \mathop{\mathbb{E}}_{\hat{W}} \left[\hat{R}^{\text{test}}_{S_{(m+1):2m}, C_\ell^{(i)}}(\hat{W}) \right], \tag{C.1.9}$$

where the $C_\ell^{(i)}$ are $\{0,1\}$-valued random variables which become independent when conditioned on s, with probability distributions

$$\mathbb{P}_{C_\ell^{(i)}|s}[C_\ell^{(i)} = 1] = \text{Tr}[E(z_i)\rho_0] = 1 - \mathbb{P}_{C_\ell^{(i)}|s}[C_\ell^{(i)} = 0] \tag{C.1.10}$$

for all $1 \leq \ell \leq m_{\text{test}}$ and for all $m+1 \leq i \leq 2m$. Note that the $\hat{W}$ in this expression depends on the random variables $B_\ell^{(i)}$, which in turn depend on the random variables z_i. Similarly, by Definition 12, the expected true risk is

$$\begin{aligned} R_\rho(\mathcal{A}) &= \mathop{\mathbb{E}}_{\bar{S}\sim P^{2m}} \mathop{\mathbb{E}}_{\bar{C}_\ell^{(i)}|\bar{S}} \mathop{\mathbb{E}}_{\bar{\hat{W}}} \left[\hat{R}^{\text{test}}_{\bar{S}_{(m+1):2m}, \bar{C}_\ell^{(i)}}(\bar{\hat{W}}) \right] & \text{(C.1.11)} \\ &= \mathop{\mathbb{E}}_{\bar{Z}_{m+1}\sim P} \mathop{\mathbb{E}}_{\bar{C}_\ell^{(m+1)}|\bar{S}_{m+1}} \mathop{\mathbb{E}}_{\bar{\hat{W}}} \left[\left| \text{Tr}[E(\bar{Z}_{m+1})\rho_0(\bar{\hat{W}})] - \frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}} \bar{C}_\ell^{(m+1)} \right| \right], & \text{(C.1.12)} \end{aligned}$$

where $(\bar{Z}_{m+1}, \bar{C}_\ell^{(m+1)})$ has the same distribution as $(Z_{m+1}, C_\ell^{(m+1)})$, $\bar{\hat{W}}$ has the same distribution as $\hat{W}$ (induced via the random variables $\bar{B}_\ell^{(i)}$), but $(\bar{Z}_{m+1}, \bar{C}_\ell^{(m+1)})$ and $\bar{\hat{W}}$ are independent.

Next, we apply Corollary 24. As there is no quantum hypothesis and as the initial quantum data factorizes across the test-train bipartition, it suffices to verify the classical sub-gaussianity assumption. We can rewrite

$$\text{Tr}\left[L^{(i)}_{c_\ell^{(i)}}(Z_i, w)\rho_0^{\otimes m_{\text{test}}} \right] = \mathbb{E}_{C_\ell^{(i)}|Z_i} \left[\left| \text{Tr}[E(Z_i)\rho_0(w)] - \frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}} C_\ell^{(i)} \right| \right], \tag{C.1.13}$$

where, for any $m+1 \leq i \leq 2m$, conditioned on Z_i the random variables $C_1^{(i)}, \ldots, C_{m_{\text{test}}}^{(i)}$ are i.i.d., take values in $\{0,1\}$, and have mean $\text{Tr}[E(Z_i)\rho_0]$. So, Hoeffding's inequality (Hoeffding, 1963) implies that the random variable $\text{Tr}[E(Z_i)\rho_0] - \frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}} C_\ell^{(i)}$ is $\frac{C}{\sqrt{m_{\text{test}}}}$-sub-gaussian conditioned on Z_i. Here and below, we use C to denote a constant that may change with each occurrence. Next, using a triangle inequality and the equivalent formulation of sub-gaussianity in terms of L_p-norm bounds (compare (Vershynin, 2018, Proposition 2.5.2), we obtain the bound

$$\mathbb{E}_{C_\ell^{(i)}|Z_i}\left[\left|\mathrm{Tr}[E(Z_i)\rho_0(w)] - \frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}} C_\ell^{(i)}\right|\right] \tag{C.1.14}$$

$$\leq |\mathrm{Tr}[E(Z_i)\rho_0(w)] - \mathrm{Tr}[E(Z_i)\rho_0]| + \mathbb{E}_{C_\ell^{(i)}|Z_i}\left[\left|\mathrm{Tr}[E(Z_i)\rho_0] - \frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}} C_\ell^{(i)}\right|\right] \tag{C.1.15}$$

$$\leq 2 + \frac{C}{\sqrt{m_{\text{test}}}} \tag{C.1.16}$$

almost surely. So, the random variable $\mathrm{Tr}[L^{(i)}_{c_\ell^{(i)}}(Z_i, w)\rho_0^{\otimes m_{\text{test}}}]$, with $Z_i \sim P$, is $\left(C(1 + \frac{1}{\sqrt{m_{\text{test}}}})\right)$-sub-gaussian by Hoeffding's Lemma (Hoeffding, 1963). Notice also that this sub-gaussianity remains true if we further condition on $Z_1, \ldots, Z_m$, since $\mathrm{Tr}[L(Z,w)\rho_{\text{test}}]$ is independent of these random variables. Thus, first conditioning on $Z_1, \ldots, Z_m$ and then applying Corollary 24, we obtain the following expected generalization error bound:

$$|\mathrm{gen}_\rho(\mathcal{A})| = \left|\mathbb{E}_{Z_1,\ldots,Z_m}\left[\mathrm{gen}_\rho(\mathcal{A})|Z_1, \ldots, Z_m\right]\right| \tag{C.1.17}$$

$$\leq \mathbb{E}_{Z_1,\ldots,Z_m}\left[|\mathrm{gen}_\rho(\mathcal{A})|\,|Z_1, \ldots, Z_m\right] \tag{C.1.18}$$

$$\leq \mathbb{E}_{Z_1,\ldots,Z_m}\left[\sqrt{\left(\frac{C}{m}\left(1 + \frac{1}{\sqrt{m_{\text{test}}}}\right)^2\right) I(S;\hat{W}|Z_1, \ldots, Z_m)}\right]. \tag{C.1.19}$$

Next, we bound the conditional mutual information $I(S;\hat{W}|Z_1, \ldots, Z_m)$. By construction, conditioned on $Z_1, \ldots, Z_m$, the output hypothesis random variable $\hat{W}$ takes values in W_1. Thus, $I(S;\hat{W}|Z_1, \ldots, Z_m) \leq \log_2(|\mathsf{W}_1|)$. We can control $|\mathsf{W}_1|$ using bounds from classical learning theory. Notice that W_1 is an empirical $\tilde{\varepsilon}$-covering net for (a subset of) the function class $\mathcal{F}_{\mathcal{S}(\mathbb{C}^d)}$ of d-dimensional quantum states viewed as functionals on effect operators, that is,

$$\mathcal{F}_{\mathcal{S}(\mathbb{C}^d)} = \left\{\mathcal{E}(\mathbb{C}^d) \ni E \mapsto \mathrm{Tr}[E\rho] \in [0,1]\right\}_{\rho\in\mathcal{S}(\mathbb{C}^d)} \subseteq [0,1]^{\mathcal{E}(\mathbb{C}^d)}. \tag{C.1.20}$$

By (Mendelson and Vershynin, 2003, Theorem 1) (see also (Anthony and Bartlett, 1999, Sections 12 and 18), (Vidyasagar, 2003, Sections 4.2.2 and 4.2.4), or (Caro, 2022b, Section 3.3)), we can find such a covering net of cardinality $|\mathsf{W}_1| \leq (2/\tilde{\varepsilon})^{C\cdot\mathrm{fat}(\mathcal{F}_{\mathcal{S}(\mathbb{C}^d)}, c\tilde{\varepsilon})}$, where $c, C > 0$ are some constants and $\mathrm{fat}(\mathcal{F}, \alpha)$ denotes the α-fat-shattering dimension of a real-valued function class $\mathcal{F}$, introduced in (Kearns and Schapire, 1994). For our purposes, it suffices to know that the fat-shattering dimension of $\mathcal{F}_{\mathcal{S}(\mathbb{C}^d)}$ scales logarithmically in d: As shown in (Aaronson, 2007, Corollary 2.7), $\mathrm{fat}(\mathcal{F}_{\mathcal{S}(\mathbb{C}^d)}, \gamma) \leq {}^{C\log(d)}/_{\gamma^2}$ holds for all $\gamma > 0$, with $C > 0$ some constant. Therefore, we can take our covering net W_1 to have cardinality $|\mathsf{W}_1| \leq (2/\tilde{\varepsilon})^{C\log(d)/\tilde{\varepsilon}^2}$, for some constant $C > 0$. This gives the conditional mutual information bound

$$I(S;\hat{W}|Z_1, \ldots, Z_m) \leq \log_2(|\mathsf{W}_1|) \leq \frac{C\log(d)}{\tilde{\varepsilon}^2} \cdot \log\left(\frac{2}{\tilde{\varepsilon}}\right). \tag{C.1.21}$$

Plugging this back into our expected generalization error bound, we have shown:

$$\mathrm{gen}_\rho(\mathcal{A}) \leq \sqrt{\left(\frac{C}{m}\left(1 + \frac{1}{\sqrt{m_{\text{test}}}}\right)^2\right)\frac{\log(d)}{\tilde{\varepsilon}^2} \cdot \log\left(\frac{2}{\tilde{\varepsilon}}\right)}. \tag{C.1.22}$$

This shows that we can achieve good expected generalization performance with a training data size m scaling only logarithmically in the dimension d.

We now demonstrate the usefulness of this expected generalization error bound as a tool in bounding the expected excess prediction error of $\mathcal{A}$, which we denote by $\mathrm{excess}_\rho(\mathcal{A})$ and which is defined as the difference between the expected prediction error of $\mathcal{A}$, given by

$$\mathop{\mathbb{E}}_{\hat{\bar{W}}} \mathop{\mathbb{E}}_{\bar{Z}_{m+1}} \left[\left| \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(\hat{\bar{W}})] \right| \right], \tag{C.1.23}$$

and the optimal achievable expected prediction error, given by

$$\inf_{w \in \mathsf{W}} \mathop{\mathbb{E}}_{\bar{Z}_{m+1}} \left[\left| \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(w)] \right| \right]. \tag{C.1.24}$$

Namely, based on Equation (C.1.22), we show in Appendix D:

Corollary 26 *The quantum learning algorithm described above satisfies the excess prediction error bound*

$$\mathrm{excess}_\rho(\mathcal{A}) \leq \tilde{\varepsilon} + \tilde{\mathcal{O}}\left(\sqrt{\frac{\log(d)}{m\tilde{\varepsilon}^2}} + \frac{1}{\sqrt{m_{\mathrm{train}}}} + \frac{1}{\sqrt{m_{\mathrm{test}}}} \right). \tag{C.1.25}$$

In particular, picking $\tilde{\varepsilon} = \varepsilon/2$, our procedure achieves an expected excess prediction error of at most ε for $m \leq \tilde{\mathcal{O}}(\log(d)/\varepsilon^4)$ and $m_{\mathrm{train}}, m_{\mathrm{test}} \leq \tilde{\mathcal{O}}(1/\varepsilon^2)$. This way, our information-theoretic approach reproduces the essential feature of (Aaronson, 2007, Theorem 1.1), namely the favorable dimension-dependence, as well as the $(1/\varepsilon^4)$-scaling. Moreover, whereas (Aaronson, 2007) starts from classical training data obtained by measuring copies of the unknown state, our analysis begins with the quantum data and thereby simultaneously leads to bounds on m, m_{train}, and m_{test}. Here, m_{train} and m_{test} are d-independent. Note: If we consider m, m_{train}, and m_{test} as fixed, determining our resources, then we can achieve an excess prediction error of order $\max\{\sqrt[4]{\log(d)/m}, \sqrt{1/m_{\mathrm{train}}}, \sqrt{1/m_{\mathrm{test}}}\}$.

Remark 27 *From our reasoning leading to Corollary 26, one can extract a proof that extends the reasoning from (Xu and Raginsky, 2017, Section 4.2) beyond binary classification to regression with a continuous target space. This then shows how to recover in-expectation versions of known generalization bounds in terms of the fat-shattering dimension (Bartlett and Long, 1998; Anthony and Bartlett, 2000) via an information-theoretic approach to generalization and may be of independent interest.*

Extension to entangled quantum data. The above discussion of PAC learning quantum states assumed access to independent copies of the unknown state ρ_0. We now discuss how our framework and results can be applied if the copies of ρ_0 are correlated/entangled across the test-train bipartition. This should be viewed as a proof-of-principle demonstration, similar extensions beyond the case of independent quantum data are possible also for the applications discussed in the following subsections. Moreover, our framework can be modified to incorporate entanglement inside the test and train subsystems, respectively, upon suitably redefining the expected true risk.

Consider CQ data of the form

$$\rho = \mathop{\mathbb{E}}_{S=(Z_1,\ldots,Z_{2m})\sim P^{2m}} \left[\left(\bigotimes_{i=1}^{2m} |Z_i\rangle\langle Z_i| \right) \otimes \tilde{\rho} \right], \tag{C.1.26}$$

where $\tilde{\rho} \in \mathcal{S}(\mathcal{H}_{\text{data}})$ satisfies $\operatorname{Tr}_{\text{test}}[\tilde{\rho}] = (\rho_0^{\otimes m_{\text{train}}})^{\otimes m}$ and $\operatorname{Tr}_{\text{train}}[\tilde{\rho}] = (\rho_0^{\otimes m_{\text{train}}})^{\otimes m}$. Let us analyze the same learning strategy as discussed above with the same choice of loss observable. The expected empirical risk now becomes

$$\hat{R}_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{S\sim P^{2m}} \mathop{\mathbb{E}}_{D_\ell^{(i)}|S} \mathop{\mathbb{E}}_{\hat{W}} \left[\hat{R}^{\text{test}}_{S_{(m+1):2m}, D_\ell^{(i)}}(\hat{W}) \right], \tag{C.1.27}$$

where the $D_\ell^{(i)}$ are $\{0,1\}$-valued random variables that conditioned on s have the joint distribution

$$\mathbb{P}_{\{D_\ell^{(i)}\}|s}[(D_\ell^{(i)})_{\ell,i} = (d_\ell^{(i)})_{\ell,i}] = \operatorname{Tr}\left[\left(\bigotimes_{i=1}^{m} \bigotimes_{\ell=1}^{m_{\text{test}}} (d_\ell^{(i)} E(z_i) + (1 - d_\ell^{(i)})(\mathbb{1}_d - E(z_i)) \right) \rho^{\mathcal{A}}_{\text{test}}(s, \{b_\ell^{(i)}\}_{\ell,i}) \right], \tag{C.1.28}$$

where the $b_\ell^{(i)}$ are the measurement outcomes obtained by measuring for each $m+1 \leq i \leq 2m$, the 2-outcome POVM $\{E(z_i), \mathbb{1} - E(z_i)\}$ on the i^{th} set of m_{train} subsystems of $\tilde{\rho}$. Crucially, whereas in our previous analysis the expected empirical risk depended on random variables $C_\ell^{(i)}$ that, conditioned on s, were independent of the outcome random variables $B_\ell^{(i)}$ seen during training (and thus of the induced hypothesis $\hat{W}$), now it depends on random variables $D_\ell^{(i)}$ that may depend on the $B_\ell^{(i)}$. This occurs because, due to the initially present correlations and entanglement, the collapsing measurement performed by the learner on the training data subsystem may also influence the test data subsystem. Thus, using the "contaminated" test data for validation may lead to a worse risk estimate than in the i.i.d. case.

In our definition of expected true risk, we decoupled the test and training data subsystems before letting the learner act. This ensures that, even if correlations or entanglement are present across the test-train bipartition initially, our notion of expected true risk still reproduces the same quantity as in the case of independent quantum copies,

$$R_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{\bar{Z}_{m+1}\sim P} \mathop{\mathbb{E}}_{\bar{C}_\ell^{(m+1)}|\bar{S}_{m+1}} \mathop{\mathbb{E}}_{\bar{\hat{W}}} \left[\left| \operatorname{Tr}[E(\bar{Z}_{m+1})\rho_0(\bar{\hat{W}})] - \frac{1}{m_{\text{test}}} \sum_{\ell=1}^{m_{\text{test}}} \bar{C}_\ell^{(m+1)} \right| \right]. \tag{C.1.29}$$

The classical sub-gaussianity analysis is exactly the same as before. Now, we in addition have to determine the quantum sub-gaussianity behavior. To this end, note that $\rho_{\text{test}}(s) = (\rho_0^{\otimes m_{\text{train}}})^{\otimes m}$ factorizes by assumption. Moreover, our loss observable is $(2/m\cdot m_{\text{test}})$-Lipschitz w.r.t. the factorization into $m \cdot m_{\text{test}}$ subsystems. (This can be seen by a bounded differences argument: If two density matrices coincide after tracing out a single subsystem, then at most one of the $\bar{C}_\ell^{(i)}$ in $\hat{R}^{\text{test}}_{s_{(m+1):2m}, \bar{C}_\ell^{(i)}}(w)$ changes, leading to an overall change bounded by $2/m\cdot m_{\text{test}}$.) Thus, after conditioning on $Z_1, \ldots, Z_m$, we can apply Corollary 24 and, as there is no quantum hypothesis, obtain the following generalization bound:

$$\begin{aligned} |\operatorname{gen}_\rho(\mathcal{A})| \leq{}& \mathop{\mathbb{E}}_{Z_1,\ldots,Z_m} \left[\sqrt{ \frac{8}{m \cdot m_{\text{test}}} \mathop{\mathbb{E}}_{Z_{m+1},\ldots,Z_{2m}\sim P^m} \left[\chi \left(\left\{ P^{\mathcal{A}}_{B_\ell^{(i)}|S}(\{b_\ell^{(i)}\}_{\ell,i}), \rho^{\mathcal{A}}_{\text{test}}(S, \{b_\ell^{(i)}\}_{\ell,i}) \right\}_{b_\ell^{(i)}} \right) \right] } \right] \\ &+ \mathop{\mathbb{E}}_{Z_1,\ldots,Z_m} \left[\sqrt{ \left(\frac{C}{m} \left(1 + \frac{1}{\sqrt{m_{\text{test}}}} \right)^2 \right) I(S; \hat{W} | Z_1, \ldots, Z_m) } \right]. \end{aligned} \tag{C.1.30}$$

The second summand can be controlled as in the case of i.i.d. quantum copies. The first summand, which can be viewed as a proxy for the maximal information about the training outcomes $b_\ell^{(i)}$ accessible from the post-measurement state $\rho^{\mathcal{A}}_{\text{test}}(S, \{b_\ell^{(i)}\}_{\ell,i})$ on the test subsystem, requires a separate analysis. Obtaining bounds on this term via quantities measuring the initial correlations/entanglement between the $\mathtt{test}$ and $\mathtt{train}$ subsystems or via properties of the POVMs used by the learner is an interesting challenge that we leave open for future work.

C.2. Quantum PAC learning from entangled data

Next, we demonstrate that our framework allows us to prove information-theoretic generalization bounds for quantum PAC learning from entangled data, which can be viewed as a variation on the usual standard PAC learning framework (Bshouty and Jackson, 1998; Arunachalam and de Wolf, 2017). The classical framework of (Xu and Raginsky, 2017), as reviewed in Section 1, considers training data S consisting of i.i.d. examples Z_i drawn from P. Written in terms of states diagonal in the computational basis, this data corresponds to the mixed state $\left(\sum_{z\in\mathsf{Z}} P(z)\,|z\rangle\langle z|\right)^{\otimes m}$. Instead of this classical data, we consider entangled quantum data representing a purification of this probabilistic mixture. Namely, we consider a quantum data state $\rho = (|\phi\rangle\langle\phi|)^{\otimes m}$ with $|\phi\rangle = \sum_{z\in\mathsf{Z}} \sqrt{P(z)}\,|z\rangle_{\text{test}} \otimes |z\rangle_{\text{train}}$ and thus

$$|\phi\rangle^{\otimes m} = \sum_{z_1,\dots z_m\in\mathsf{Z}} \sqrt{P(z_1)\cdots P(z_m)}\,|z_1,\dots z_m\rangle_{\text{test}} \otimes |z_1,\dots z_m\rangle_{\text{train}} \tag{C.2.1}$$

$$= \sum_{s\in\mathsf{Z}^m} \sqrt{P^m(s)}\,|s\rangle_{\text{test}} \otimes |s\rangle_{\text{train}}\ , \tag{C.2.2}$$

where we identify the purifying system as the test data system. Here, the data is purely quantum, there is no classical part. As our focus is on learning a classical function, we take $\mathsf{Z} = \mathsf{X}\times\mathsf{Y}$, with $\mathsf{X} = \{0,1\}^n$ and $\mathsf{Y} = \{0,1\}$ and accordingly $\mathcal{H}_{\text{test}} = \mathcal{H}_{\text{train}} = ((\mathbb{C}^2)^{\otimes n}\otimes\mathbb{C}^2)^{\otimes m}$. We write $Z_i = (X_i, Y_i)$ and take $\mathsf{W}\subset\mathsf{Y}^{\mathsf{X}}$. We note that quantum data states as in Equation (C.2.2) can be obtained from the more established quantum superposition examples of Bshouty and Jackson (1998) by attaching an auxiliary register and applying CNOT gates, and the reverse conversion can be achieved by applying CNOTs and discarding the auxiliary system.

Before proceeding further, let us comment on how this formulation compares to the classical framework obtained by extending (Xu and Raginsky, 2017) to include test data, discussed in Section 1.1.1. Recall that this classical description involved perfectly correlated test and training data random variables; the entanglement between test and training subsystems in the pure state $|\phi\rangle^{\otimes m}$ can be viewed as a fully quantum analogue of this perfect correlation, with respect to the computational basis.

We are now ready to quantumly analyze a learner that acts according to a conditional probability distribution $P^{\mathcal{A}}(W|S)$. To this end, we consider a quantum learner $\mathcal{A}$ that measures the quantum data in the computational basis and processes the observed outcomes via $P^{\mathcal{A}}(W|S)$. To model this without introducing classical random variables, we take the hypothesis space $\mathcal{H}_{\text{hyp}} = \mathbb{C}^{|W|}$. The quantum learner $\mathcal{A}$, without performing any POVM with observed outcomes, implements the channel

$$\Lambda^{\mathcal{A}}(\rho) = \sum_{s\in\mathsf{Z}^m}\sum_{w\in\mathsf{W}} \langle s|\,\rho\,|s\rangle\, P^{\mathcal{A}}(w|s)\,|w\rangle\langle w|\ . \tag{C.2.3}$$

Thus, the state after the action of the learner is given by

$$\sigma^{\mathcal{A}} = \sum_{s\in \mathsf{Z}^m}\sum_{w\in \mathsf{W}} P^{\mathcal{A}}(s,w)\, |s\rangle\langle s|_{\text{test}} \otimes |w\rangle\langle w|_{\text{hyp}}\ . \tag{C.2.4}$$

To evaluate the performance of $\mathcal{A}$, we take the loss observable $L = \frac{1}{m}\sum_{i=1}^m L_i$ with

$$L_i = \sum_{z_i\in \mathsf{Z}}\sum_{w\in \mathsf{W}} \ell(w,z_i)\, |z_i\rangle\langle z_i|_{\text{test},i} \otimes |w\rangle\langle w|_{\text{hyp}}\ , \tag{C.2.5}$$

where $\ell : \mathsf{W}\times\mathsf{Z}\to\mathbb{R}_{\geq 0}$ is some classical loss function. As the relevant operators commute, it is easy to see that this choice reproduces the clasical notions of expected empirical risk

$$\operatorname{Tr}[L\sigma^{\mathcal{A}}] = \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[\hat{R}_S(W)\right] \tag{C.2.6}$$

and expected true risk

$$\operatorname{Tr}[L(\rho_{\text{test}}\otimes\sigma^{\mathcal{A}}_{\text{hyp}})] = \mathop{\mathbb{E}}_{(\bar{S},\bar{W})\sim P^m\otimes P^{\mathcal{A}}_{\mathsf{W}}}\left[\hat{R}_{\bar{S}}(\bar{W})\right] = \mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathsf{W}}}\left[R_P(W)\right]\ . \tag{C.2.7}$$

These are exactly the notions of risk familiar from the classical case.

Moreover, the QMGF bound for L w.r.t. $\rho_{\text{test}}\otimes\sigma^{\mathcal{A}}_{\text{hyp}}$ coincides with the classical MGF bound for $\frac{1}{m}\sum_{i=1}^m \ell(\bar{W},\bar{Z}_i)$. Also, as $\sigma^{\mathcal{A}}$ is diagonal, we see that $I(\text{test};\text{hyp})_{\sigma^{\mathcal{A}}} = I(S;W)$. Thus, Corollary 23 reproduces the main result of (Xu and Raginsky, 2017) via the QMI term[5]. Here, both the classical MI and the Holevo information terms vanish because there is no classical hypothesis.

Remark 28 *In this section, we have described learning from quantum data in the form of a pure entangled state. Recently, (Caro et al., 2024) proposed* mixture-of-superposition quantum examples *as an alternative to the more established superposition examples (Bshouty and Jackson, 1998; Arunachalam and de Wolf, 2017) for agnostic quantum learning. Similarly, one may change the model considered here and instead work with quantum data of the form* $\rho = (\mathbb{E}_{f\sim F_P}\left[(|\phi_f\rangle\langle\phi_f|)\right])^{\otimes m}$, *where*

$$|\phi_f\rangle = \sum_{x\in\mathsf{X}}\sqrt{P_{\mathsf{X}}(x)}\,|x,f(x)\rangle_{\text{test}}\otimes|x,f(x)\rangle_{\text{train}}\ , \tag{C.2.8}$$

and where F_P *is the probability distribution on the function space* $\{0,1\}^{\{0,1\}^n}$ *induced by* P *via*

$$F_P(f) = \prod_{x'\in\{0,1\}^n}\mathbb{P}_{(x,y)\sim P}\left[f(x') = y \mid x = x'\right]\ . \tag{C.2.9}$$

An analysis similar to the one presented above can also be carried out for this notion of quantum data and again reproduces the classical bound of (Xu and Raginsky, 2017).

[5] While the statement of (Xu and Raginsky, 2017, Theorem 1) is correct, the argument there was based on the claim that, if $\bar{X},\bar{Y}$ are independent random variables and if $f(x,\bar{Y})$ is β-sub-gaussian for every x, then also $f(\bar{X},\bar{Y})$ is β-sub-gaussian. This claim is in general not correct because of complications regarding centering, as pointed out, e.g., in Appendix C of the arXiv version of (Negrea et al., 2019). This issue can be circumvented by first conditioning on the hypothesis random variable (see, e.g., Raginsky, 2019, p. 22). Thus, our claim here is that we have reproduced the following version of (Xu and Raginsky, 2017, Theorem 1) without the improvement via conditioning: If $\frac{1}{m}\sum_{i=1}^m \ell(\bar{W},\bar{Z}_i)$ is $(\frac{\beta}{\sqrt{m}})$-sub-gaussian, then Eq. (10) of (Xu and Raginsky, 2017) holds.

C.3. Quantum parameter estimation

Next, we demonstrate how to incorporate quantum parameter estimation tasks, typically considered in quantum metrology (Giovannetti et al., 2006) and quantum sensing Degen et al. (2017), into our framework. Let $\mathsf{Z} = \Theta \subseteq \mathbb{R}^n$ be a parameter space, equipped with the induced Borel σ-algebra. Consider the data Hilbert space

$$\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{test}} \otimes \mathcal{H}_{\text{train}} = ((\mathbb{C}^d)^{\otimes m_{\text{test}}})^{\otimes m} \otimes ((\mathbb{C}^d)^{\otimes m_{\text{train}}})^{\otimes m}. \tag{C.3.1}$$

For an unknown probability measure P over Θ, let the quantum data state ρ be the CQ state

$$\rho = \mathbb{E}_{S=(Z_1,\dots,Z_m)\sim P^m}\left[\left(\bigotimes_{i=1}^m |Z_i\rangle\langle Z_i|\right) \otimes \left(\bigotimes_{i=1}^m \rho(Z_i)^{\otimes m_{\text{test}}}\right) \otimes \left(\bigotimes_{i=1}^m \rho(Z_i)^{\otimes m_{\text{train}}}\right)\right], \tag{C.3.2}$$

where the $\rho(Z_i)$ are parameter-dependent qudit states, with the mapping $z \mapsto \rho(z)$ known in advance. Note: Even if this mapping is known in principle, one may not be able to prepare copies of the respective state. Thus, when aiming to learn how to extract information about the unknown parameter from the quantum system, it nevertheless makes sense to work with a finite number of copies of each $\rho(Z_i)$.

The goal of a quantum learner here is to learn a POVM that, when performed on copies of $\rho(Z)$, produces an accurate estimate of the unknown parameter Z. Therefore, to model the learner, we let $\mathcal{H}_{\text{hyp}}$ be trivial, and we take W to be some measurable hypothesis space such that each $w \in \mathsf{W}$ is associated with a POVM $\{F_w(\hat{z})\}_{\hat{z}\in\mathsf{Z}} \subseteq \mathcal{E}((\mathbb{C}^d)^{\otimes m_{\text{test}}})$. The action of the learner is described by POVMs $\{E_s^{\mathcal{A}}(w)\}_{w\in\mathsf{W}} \subseteq \mathcal{E}(((\mathbb{C}^d)^{\otimes m_{\text{train}}})^{\otimes m})$, for $s = (z_i)_i \in \mathsf{Z}^m$. If we now define the loss observables as

$$L(s,w) = L((z_i)_i, w) = \sum_{\hat{z}\in\mathsf{Z}} \frac{1}{m}\sum_{i=1}^m \|z_i - \hat{z}\|_p \, (\mathbb{1}_d^{\otimes m_{\text{test}}})^{\otimes(i-1)} \otimes F_w(\hat{z}) \otimes (\mathbb{1}_d^{\otimes m_{\text{test}}})^{\otimes(m-i)}, \tag{C.3.3}$$

for some $p \geq 1$, then we can evaluate the expected empirical risk (Definition 11) as

$$\hat{R}_\rho(\mathcal{A}) = \mathbb{E}_{(S,\hat{Z})\sim P^{\mathcal{A}}}\left[\frac{1}{m}\sum_{i=1}^m \left\|Z_i - \hat{Z}\right\|_p\right], \tag{C.3.4}$$

where the classical data $S = (Z_i)_i$ and the estimated parameter $\hat{Z}$ have the joint probability distribution

$$P^{\mathcal{A}}((z_i)_i, \hat{z}) = \left(\prod_{i=1}^m P(z_i)\right) \cdot \sum_{w\in\mathsf{W}} \operatorname{Tr}\left[E^{\mathcal{A}}_{(z_i)_i}(w)\left(\bigotimes_{i=1}^m \rho(z_i)^{\otimes m_{\text{train}}}\right)\right] \cdot \operatorname{Tr}\left[F_w(\hat{z})\rho(z_i)^{\otimes m_{\text{test}}}\right]. \tag{C.3.5}$$

Similarly, the expected true risk (Definition 12) is

$$R_\rho(\mathcal{A}) = \mathbb{E}_{\bar{Z},\hat{Z}}\left[\left\|\bar{Z} - \hat{Z}\right\|_p\right], \tag{C.3.6}$$

where

$$P^{\mathcal{A}}(\bar{z},\hat{z}) = P(\bar{z}) \cdot \mathbb{E}_{\bar{W}}\left[\operatorname{Tr}\left[F_{\bar{W}}(\hat{z})\rho(\bar{z})^{\otimes m_{\text{test}}}\right]\right], \tag{C.3.7}$$

with the random variables $\bar{Z}$ and $\bar{W}$ being independent copies of Z and W. That is, the expected empirical risk measures the expected average norm error that estimates produced from the learned POVM make on the states that it has been learned from. Meanwhile, the expected true risk measures the expected average norm error that estimates produced from the learned POVM make on a new parameter setting drawn at random from the underlying distribution.

We next evaluate the guarantees of Appendix B for this setting. We are in the scenario of Corollary 24 without a quantum hypothesis and without initial test-train entanglement, so it suffices to study the sub-gaussianity parameter of the random variable $\sum_{\hat{z}\in\mathsf{Z}} \|Z_i - \hat{z}\|_p \operatorname{Tr}[F_w(\hat{z})\rho(Z_i)^{\otimes m_{\text{test}}}] = \mathbb{E}_{\hat{Z}|Z_i,w}\left[\left\|Z_i - \hat{Z}\right\|_p\right]$ for $Z_i \sim P$ and for fixed w. If we assume the parameter space Z to have a p-norm diameter $B_p < \infty$, then this random variable is bounded by B_p and thus $(\frac{B_p}{2})$-sub-gaussian by Hoeffding. Then, Corollary 24 implies

$$|R_\rho(\mathcal{A}) - \hat{R}_\rho(\mathcal{A})| \leq \sqrt{\frac{B_p}{2m} I((Z_i)_i; W)}. \tag{C.3.8}$$

Informally, this tells us: If the learned POVM performs well on the available classical-quantum data and does not depend too strongly on any specific sample parameter setting seen during training, then the POVM will also accurately extract the parameter of a previously unseen $\rho(Z)$. Similarly to Example 7, we may further bound the relevant mutual information in terms of the complexity of the admissible POVMs. To the best of our knowledge, this is the first generalization bound for quantum parameter estimation.

C.4. Variational quantum machine learning

In this subsection, we consider a task of classifying classical data via an embedding into quantum states, similarly to (Banchi et al., 2021). To formalize this task, consider the data Hilbert space

$$\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{data}} = ((\mathbb{C}^d)^{\otimes m_{\text{test}}})^{\otimes m} \otimes ((\mathbb{C}^d)^{\otimes m_{\text{train}}})^{\otimes m}. \tag{C.4.1}$$

Let P be an unknown probability measure over a measurable input space X, let $f : \mathsf{X} \to \{1, \ldots, k\}$ be an unknown labelling function, and consider the quantum data state

$$\rho = \mathbb{E}_{X_1,\ldots,X_m\sim P^m}\left[\left(\bigotimes_{i=1}^m |X_i, f(X_i)\rangle\langle X_i, f(X_i)|\right) \otimes \left(\bigotimes_{i=1}^m \rho(X_i)^{\otimes m_{\text{test}}}\right) \otimes \left(\bigotimes_{i=1}^m \rho(X_i)^{\otimes m_{\text{train}}}\right)\right], \tag{C.4.2}$$

where the $\rho(x_i)$ are quantum states into which the classical inputs x_i are embedded according to a mapping $x \mapsto \rho(x)$, which may be known or unknown. While the mapping $x \mapsto \rho(x)$ is typically in principle known in variational QML, since it is given by the parametrized circuit, it can nevertheless make sense to work with a restricted number of copies of output states, for example if running the quantum circuit itself is expensive. Importantly, while with a known mapping the output state and expectation values thereof could be computed classically, this will become infeasible for large system sizes. Then, using actual quantum circuits to prepare and measure states may be necessary.

The goal of a quantum learner in this scenario is to learn a POVM that, when performed on copies of $\rho(x)$, produces the correct label $f(x)$ with high probability. Accordingly, we model the learner by taking $\mathcal{H}_{\text{hyp}}$ to be trivial, and by taking W to be some hypothesis space such that each $w \in \mathsf{W}$ is associated with a k-outcome POVM $\{F_w(\ell)\}_{\ell=1}^k \subseteq \mathcal{E}((\mathbb{C}^d)^{\otimes m_{\text{test}}})$. We describe the action of the learner by POVMs $\{E^{\mathcal{A}}_{((x_i,f(x_i))_i}(w)\}_{w\in\mathsf{W}} \subseteq \mathcal{E}(((\mathbb{C}^d)^{\otimes m_{\mathrm{train}}})^{\otimes m})$, for $((x_i, f(x_i))_i \in (\mathsf{X} \times \{1,\dots,k\})^m$. We now consider the loss observables

$$L(s,w) = L((x_i,f(x_i))_i,w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{\ell\in\{1,\dots,k\}\setminus\{f(x_i)\}} (\mathbb{1}_d^{\otimes m_{\mathrm{test}}})^{\otimes(i-1)} \otimes F_w(\ell) \otimes (\mathbb{1}_d^{\otimes m_{\mathrm{test}}})^{\otimes(m-i)}. \tag{C.4.3}$$

According to Definition 11, this leads to the expected empirical risk

$$\hat{R}_\rho(\mathcal{A}) = \mathbb{E}_{(X_1,\dots,X_m;W)\sim P^{\mathcal{A}}}\left[\frac{1}{m}\sum_{i=1}^{m}\sum_{\ell\in\{1,\dots,k\}\setminus\{f(X_i)\}} \mathrm{Tr}[F_W(\ell)\rho(X_i)^{\otimes m_{\mathrm{test}}}]\right] \tag{C.4.4}$$

$$= \mathbb{E}_{(X_1,\dots,X_m;W)\sim P^{\mathcal{A}}}\left[1 - \frac{1}{m}\sum_{i=1}^{m}\mathrm{Tr}[F_W(f(X_i))\rho(X_i)^{\otimes m_{\mathrm{test}}}]\right], \tag{C.4.5}$$

Similarly, according to Definition 12 we obtain the expected true risk

$$R_\rho(\mathcal{A}) = \mathbb{E}_{\bar{X};\bar{W}}\left[\sum_{\ell\in\{1,\dots,k\}\setminus\{f(\bar{X})\}} \mathrm{Tr}[F_{\bar{W}}(\ell)\rho(\bar{X})^{\otimes m_{\mathrm{test}}}]\right] \tag{C.4.6}$$

$$= \mathbb{E}_{\bar{X};\bar{W}}\left[1 - \mathrm{Tr}[F_{\bar{W}}(f(\bar{X}))\rho(\bar{X})^{\otimes m_{\mathrm{test}}}]\right]. \tag{C.4.7}$$

In words, $\hat{R}_\rho(\mathcal{A})$ is the expected average misclassification probability on the available training data, and $R_\rho(\mathcal{A})$ is the expected msiclassification probability on a fresh test data point. Thus, our notions of risk are simply the expected version of those considered in (Banchi et al., 2021).

It remains to evaluate the guarantees proved in Appendix B for this scenario. According to Corollary 24, we can focus on determining the sub-gaussianity parameter of the random variable $1 - \mathrm{Tr}[F_w(f(X_i))\rho(X_i)^{\otimes m_{\mathrm{test}}}]$ for $X_i \sim P$ and for fixed w. As this random variable takes values in $[0,1]$, it is $(\frac{1}{2})$-sub-gaussian by Hoeffding. So, Corollary 24 yields the expected generalization error bound

$$|R_\rho(\mathcal{A}) - \hat{R}_\rho(\mathcal{A})| \leq \sqrt{\frac{1}{2m} I(((X_i, f(X_i)))_i; W)}. \tag{C.4.8}$$

We leave it as an open question whether this bound can be directly related and compared to (Banchi et al., 2021, Theorem 1) (see also the results in (Banchi et al., 2024)), which depends exponentially on the 2-Rényi mutual information between the classical input and the quantum register for a single copy. Moreover, it will be interesting to investigate whether recent quantum generalization bounds based on (quantum) Fisher information (Abbas et al., 2021b,a; Haug and Kim, 2023) can be reinterpreted in our information-theoretic framework. More generally, we envision that, similarly to how classical information-theoretic generalization guarantees help overcome the limitations of uniform generalization bounds pointed out in (Zhang et al., 2017, 2021), a quantum information-theoretic perspective will be an important tool in remedying the drawbacks (Gil-Fuster et al., 2024) of recently established uniform generalization bounds for variational quantum machine learning (Caro and Datta, 2020; Caro et al., 2021; Chen et al., 2021; Popescu, 2021; Cai, 2021; Du et al., 2022; Caro et al., 2022; Gyurik et al., 2023).

C.5. Approximate quantum membership learning

Next, we discuss a task of learning a POVM that approximately decides membership of quantum states in an a priori unknown set. To this end, consider the data Hilbert space

$$\mathcal{H}_{\mathrm{data}} = \mathcal{H}_{\mathrm{test}} \otimes \mathcal{H}_{\mathrm{train}} = ((\mathbb{C}^d)^{\otimes m_{\mathrm{test}}})^{\otimes m} \otimes ((\mathbb{C}^d)^{\otimes m_{\mathrm{train}}})^{\otimes m}. \tag{C.5.1}$$

Let P be an unknown probability measure over qudit states. Let $\varepsilon > 0$. Consider the CQ data state

$$\rho = \mathbb{E}_{\rho_1,\ldots,\rho_m \sim P^m}\left[\left(\bigotimes_{i=1}^m |f_{\mathcal{P},\varepsilon}(\rho_i)\rangle\langle f_{\mathcal{P},\varepsilon}(\rho_i)|\right) \otimes \left(\bigotimes_{i=1}^m \rho_i^{\otimes m_{\mathrm{test}}}\right) \otimes \left(\bigotimes_{i=1}^m \rho_i^{\otimes m_{\mathrm{train}}}\right)\right], \tag{C.5.2}$$

where $\mathcal{P} \subseteq \mathcal{S}(\mathbb{C}^d)$ is a subset of qudit states, and $f_{\mathcal{P},\varepsilon} : \mathcal{S}(\mathbb{C}^d) \to \{0, 1, \perp\}$ is defined as

$$f_{\mathcal{P},\varepsilon}(\rho) = \begin{cases} 1 & \text{if } \rho \in \mathcal{P} \\ 0 & \text{if } d_1(\rho, \mathcal{P}) \geq \varepsilon \\ \perp & \text{else} \end{cases}. \tag{C.5.3}$$

Here, we used the notation $d_1(\rho, \mathcal{P}) = \inf_{\sigma \in \mathcal{P}} \|\rho - \sigma\|_1$. Thus, given an input state ρ, the function value $f_{\mathcal{P},\varepsilon}(\rho)$ ε-approximately (and ambiguously for states with $d_1(\rho, \mathcal{P}) < \varepsilon$) decides whether ρ is in $\mathcal{P}$. If we let Q denote the probability measure over $\{0, 1, \perp\} \times \mathcal{S}(\mathbb{C}^d)$ induced by P via $Q(z_i, \rho_i) = P(\rho_i)\delta_{z_i, f_{\mathcal{P},\varepsilon}(\rho_i)}$, then we can rewrite ρ as

$$\rho = \mathbb{E}_{(Z_1,\rho_1),\ldots,(Z_m,\rho_m) \sim Q^m}\left[\left(\bigotimes_{i=1}^m |Z_i\rangle\langle Z_i|\right) \otimes \left(\bigotimes_{i=1}^m \rho_i^{\otimes m_{\mathrm{test}}}\right) \otimes \left(\bigotimes_{i=1}^m \rho_i^{\otimes m_{\mathrm{train}}}\right)\right] \tag{C.5.4}$$

$$= \mathbb{E}_{Z_1,\ldots,Z_m \sim Q_{\mathsf{Z}}^m}\left[\left(\bigotimes_{i=1}^m |Z_i\rangle\langle Z_i|\right) \otimes \left(\bigotimes_{i=1}^m \mathbb{E}_{\rho_i \sim Q_{\mathcal{S}(\mathbb{C}^d)}|Z_i}\left[\rho_i^{\otimes m_{\mathrm{test}}} \otimes \rho_i^{\otimes m_{\mathrm{train}}}\right]\right)\right]. \tag{C.5.5}$$

Thus, the data state has the form of Equation (27), with classical instance space $\mathsf{Z} = \{0, 1, \perp\}$. This rewriting also highlights a similarity to ambiguous state discrimination: The training data consists of a classical label (saying "far from $\mathcal{P}$", "in $\mathcal{P}$", or "marginal case") and a quantum part given by a conditioned average over copies of the corresponding quantum states. Given the data, the learner should essentially produce a 2-outcome POVM that distinguishes between " far from $\mathcal{P}$" and "in $\mathcal{P}$" well on average, where the marginal cases do not matter.

More precisely, the goal of a learner is to learn a 2-outcome POVM for deciding whether a state belongs to $\mathcal{P}$ or is ε-far from $\mathcal{P}$. For states that are not in $\mathcal{P}$ but less than ε-far from $\mathcal{P}$, any of the two outcomes is deemed acceptable. To model such a learner, we let $\mathcal{H}_{\mathrm{hyp}}$ be trivial, and we take W to be some measurable hypothesis space such that each $w \in \mathsf{W}$ is associated with a POVM $\{F_w, \mathbb{1}_d^{\otimes m_{\mathrm{test}}} - F_w\} \subseteq \mathcal{E}((\mathbb{C}^d)^{\otimes m_{\mathrm{test}}})$. The action of the learner is described by POVMs $\{E^{\mathcal{A}}_{(z_i)_i}(w)\}_{w \in \mathsf{W}} \subseteq \mathcal{E}(((\mathbb{C}^d)^{\otimes m_{\mathrm{train}}})^{\otimes m})$, for $(z_i)_i \in \{0, 1, \perp\}^m$. We define the loss observables as

$$L(s, w) = L((z_i)_i, w) \tag{C.5.6}$$

$$= \frac{1}{m}\sum_{i=1}^m (\mathbb{1}_d^{\otimes m_{\mathrm{test}}})^{\otimes(i-1)} \otimes \left(\delta_{z_i,0} F_w + \delta_{z_i,1}(\mathbb{1}_d^{\otimes m_{\mathrm{test}}} - F_w)\right) \otimes (\mathbb{1}_d^{\otimes m_{\mathrm{test}}})^{\otimes(m-i)}. \tag{C.5.7}$$

This leads to an expected empirical risk

$$\hat{R}_\rho(\mathcal{A}) = \mathbb{E}_{\rho_1,\ldots,\rho_m;W}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{1}_{\rho_i\in\mathcal{P}}\operatorname{Tr}[(\mathbb{1}_d^{\otimes m_{\text{test}}} - F_W)\rho_i^{\otimes m_{\text{test}}}] + \mathbf{1}_{d_1(\rho_i,\mathcal{P})\geq\varepsilon}\operatorname{Tr}[F_W\rho_i^{\otimes m_{\text{test}}}]\right)\right], \tag{C.5.8}$$

where the joint distribution of $((\rho_i)_i, W)$ is given by

$$P^{\mathcal{A}}((\rho_i)_i, W) = \left(\prod_{i=1}^{m} P(\rho_i)\right)\cdot\operatorname{Tr}\left[E^{\mathcal{A}}_{(f_{\mathcal{P},\varepsilon}(\rho_i))_i}(W)\left(\bigotimes_{i=1}^{m}\rho_i^{\otimes m_{\text{train}}}\right)\right]. \tag{C.5.9}$$

The expected true risk in this case becomes

$$R_\rho(\mathcal{A}) = \mathbb{E}_{\bar{\rho};\bar{W}}\left[\mathbf{1}_{\bar{\rho}\in\mathcal{P}}\operatorname{Tr}[(\mathbb{1}_d^{\otimes m_{\text{test}}} - F_{\bar{W}})\bar{\rho}^{\otimes m_{\text{test}}}] + \mathbf{1}_{d_1(\bar{\rho},\mathcal{P})\geq\varepsilon}\operatorname{Tr}[F_{\bar{W}}\bar{\rho}^{\otimes m_{\text{test}}}]\right], \tag{C.5.10}$$

where $\bar{\rho}$ and $\bar{W}$ are independent random variables with joint product distribution

$$P^{\mathcal{A}}(\bar{\rho},\bar{W}) = P(\bar{\rho})\cdot\mathbb{E}_{\bar{\rho}_1,\ldots,\bar{\rho}_m\sim P^m}\left[\operatorname{Tr}\left[E^{\mathcal{A}}_{(f_{\mathcal{P},\varepsilon}(\bar{\rho}_i))_i}(\bar{W})\left(\bigotimes_{i=1}^{m}\bar{\rho}_i^{\otimes m_{\text{train}}}\right)\right]\right]. \tag{C.5.11}$$

That is, the expected empirical risk is the expected average error that the learned POVM makes on the data that it was learned from. In contrast, the expected true risk is the expected average probability that the POVM makes a wrong prediction on a randomly drawn new state. (Again, the classification of marginal cases is irrelevant.)

To apply Corollary 24, since there is only a classical hypothesis here and since there are no initial correlations or entanglement across the test train bipartition, we study the sub-gaussianity parameter of the random variable $\operatorname{Tr}\left[\left(\delta_{Z_i,0}F_w + \delta_{Z_i,1}(\mathbb{1}_d^{\otimes m_{\text{test}}} - F_w)\right)\mathbb{E}_{\rho_i\sim Q_{\mathcal{S}(\mathbb{C}^d)|Z_i}}\left[\rho_i^{\otimes m_{\text{test}}}\right]\right]$. This random variable takes the value $0 \leq \operatorname{Tr}\left[(\mathbb{1}_d^{\otimes m_{\text{test}}} - F_w)\mathbb{E}_{\rho_i\sim Q_{\mathcal{S}(\mathbb{C}^d)|1}}\left[\rho_i^{\otimes m_{\text{test}}}\right]\right] \leq 1$ with probability $Q_{\mathsf{Z}}(1)$ and the value $0 \leq \operatorname{Tr}\left[F_w\mathbb{E}_{\rho_i\sim Q_{\mathcal{S}(\mathbb{C}^d)|0}}\left[\rho_i^{\otimes m_{\text{test}}}\right]\right] \leq 1$ with probability $Q_{\mathsf{Z}}(0)$. In particular, by Hoeffding's inequality, it is $(\frac{1}{2})$-sub-gaussian. Thus, Corollary 24 implies

$$|R_\rho(\mathcal{A}) - \hat{R}_\rho(\mathcal{A})| \leq \sqrt{\frac{1}{2m}I((Z_i)_i;W)}. \tag{C.5.12}$$

If the learner has prior knowledge indicating that membership in $\mathcal{P}$ can be (approximately) decided using only few-copy measurements and chooses the set of admissible POVMs $\{F_w, \mathbb{1}_d^{\otimes m_{\text{test}}} - F_w\}$ with a suitable locality structure, this is expected to lead to an improved generalization performance compared to a learner that considers general many-copy measurements as viable hypotheses (compare also the discussion in Example 7). To the best of our knowledge, we are the first to take this PAC perspective on quantum membership learning and to establish a generalization bound for it.

Remark 29 *The learning problem described in this section can also be interpreted as learning to solve an average-case version of quantum property testing for states, see (Montanaro and Wolf, 2016, Section 4). From this perspective, we are asking: Given data consisting of (copies of) quantum states correctly classified according to an unknown property $\mathcal{P}$ of states and a proximity parameter ε, learn a POVM that tests $\mathcal{P}$ w.r.t. proximity parameter ε well on average over states drawn from*

P. We note that formulating meaningful average-case property testing problems is subtle. For instance average-case property testing w.r.t. uniformly random bit strings becomes trivial because of the blow-up phenomenon for Hamming distance balls (Goldreich, 2011).

Complementary to the scenario discussed above, one might also consider testing for multiple properties, drawn from an unknown distribution, on a fixed (but unknown) quantum state. Here, the challenge would be to learn a mapping from a property $\mathcal{P}$ to an associated 2-outcome POVM that classifies the unknown state ρ_0 according to whether it has property $\mathcal{P}$ or is ε-far from it.

C.6. Learning quantum state-preparation channels from classical-quantum data

In this section, we discuss how our framework can incorporate recent work on learning classical-to-quantum mappings (Chung and Lin, 2021; Caro, 2021; Fanizza et al., 2022). Let Z be some measurable instance space. Let P be a probability measure over Z. Consider the data Hilbert space

$$\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{test}} \otimes \mathcal{H}_{\text{train}} = (\mathbb{C}^d)^{\otimes m} \otimes (\mathbb{C}^d)^{\otimes m}. \tag{C.6.1}$$

Take the CQ data state

$$\rho = \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\left(\bigotimes_{i=1}^m |Z_i\rangle\langle Z_i|\right) \otimes \left(\bigotimes_{i=1}^m \mathcal{N}(Z_i)\right) \otimes \left(\bigotimes_{i=1}^m \mathcal{N}(Z_i)\right)\right], \tag{C.6.2}$$

where $\mathcal{N} : \mathsf{X} \to \mathcal{S}(\mathbb{C}^d)$ is an unknown qudit state-preparation channel. The goal of a quantum learner with a hypothesis class $\{\mathcal{N}_w\}_{w\in\mathsf{W}}$ of classical descriptions of state preparation channels is to output w such that performing $\mathcal{N}_w$ on inputs drawn from P approximates the action of the unknown channel $\mathcal{N}$ on those inputs well in trace distance. Throughout, we assume that $\mathcal{N}(z)$ and $\mathcal{N}_w(z)$ are pure states for all $z \in \mathsf{Z}$ and $w \in \mathsf{W}$, and we therefore use notations like $\mathcal{N}(z) = |\mathcal{N}(z)\rangle\langle\mathcal{N}(z)|$ for these states.

With our framework, we now formalize this setting for a learner that produces only a classical hypothesis, by taking $\mathcal{H}_{\text{hyp}}$ to be trivial, and we define our loss observables as

$$L(s,w) = L((z_i)_i, w) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_d^{\otimes(i-1)} \otimes L_i(z_i,w) \otimes \mathbb{1}_d^{\otimes(m-i)}\,, \tag{C.6.3}$$

with local loss observables

$$L_i(z_i,w) = \mathbb{1}_d - \mathcal{N}_w(z_i)\,. \tag{C.6.4}$$

With these choices, Definitions 11 and 12 lead to the expected empirical risk

$$\hat{R}_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[1 - \frac{1}{m}\sum_{i=1}^m |\langle\mathcal{N}_W(Z_i)|\mathcal{N}(Z_i)\rangle|^2\right] \tag{C.6.5}$$

$$= \mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[\frac{1}{m}\sum_{i=1}^m \left(\frac{1}{2}\|\mathcal{N}_W(Z_i) - \mathcal{N}(Z_i)\|_1\right)^2\right], \tag{C.6.6}$$

and the expected true risk

$$R_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{\bar{Z},\bar{W}}[1 - |\langle\mathcal{N}_{\bar{W}}(\bar{Z})|\mathcal{N}(\bar{Z})\rangle|^2] = \mathop{\mathbb{E}}_{\bar{Z},\bar{W}}\left[\left(\frac{1}{2}\|\mathcal{N}_{\bar{W}}(\bar{Z}) - \mathcal{N}(\bar{Z})\|_1\right)^2\right]. \tag{C.6.7}$$

That is, the expected empirical risk is the expected squared trace distance between the output states of the true channel and the hypothesis channel averaged over the training data, whereas the expected true risk considers the average squared trace distance on a fresh input state.

In this scenario, we can apply Corollary 24. Namely, for every fixed $w \in \mathsf{W}$, the random variable $\operatorname{Tr}[L_i(Z_i, w)\mathcal{N}(Z_i)]$ with $Z_i \sim P$ takes values in $[0, 1]$ and thus is $(\frac{1}{2})$-sub-gaussian by Hoeffding. Hence, the generalization error can be bounded as

$$|\mathrm{gen}_\rho(\mathcal{A})| \leq \sqrt{\frac{1}{2m} I(S;W)}\,. \tag{C.6.8}$$

If W is finite, we can bound $I(S;W) \leq \log|\mathsf{W}|$, thus recovering an in-expectation version of the sample complexity bound of (Chung and Lin, 2021). If W is infinite, we can resort to empirical covering net arguments similarly to Example 7 and Appendix C.1. When the maps in W only ever output two possible quantum states, this approach, combined with standard bounds on the size of an empirical covering net via the VC-dimension (Vapnik and Chervonenkis, 1971) (compare for instance (Vershynin, 2018, Section 8.3.4) and (Caro, 2022b, Section 3.3)), leads to an in-expectation version of the guarantee proved in (Caro, 2021, Section 4.1). More generally, using covering nets w.r.t. empirical Schatten q-norms as in (Fanizza et al., 2022, Definition 1), we can obtain generalization bounds similar in spirit to (Fanizza et al., 2022, Theorem 4), which we may turn into bounds on the expected excess risk following the line of reasoning from Appendix C.1. Note, however, that these upper bounds on the mutual information via capacity measures are worst-case, we expect tighter data- and algorithm-dependent bounds to be possible.

Let us point out that the reasoning in this subsection was specific to state preparation channels outputting pure states, so that the overlap serves as a measurable quantity tightly related to the trace distance. For channels outputting mixed states, other loss observables would be required to obtain risks that accurately reflect the desired average trace distance approximation to the true output states. In the case of only two possible known output states, one may use the Holevo-Helstrom measurement as in (Caro, 2021). However, for the general case, the "right" choice is not immediate. We believe that measurements in a random orthonormal basis as used in (Chung and Lin, 2021) or the quantum data analysis approach of (Fanizza et al., 2022) may serve as inspiration for how to incorporate channels with mixed output states. Assuming purified access, an alternative route may proceed via combining the well known Fuchs-van de Graaf inequalities (Fuchs and Van De Graaf, 1999) with a recent quantum fidelity estimation procedure for low-rank states (Wang et al., 2023).

C.7. Generalization bounds for differentially private quantum learners

Differential privacy (Dwork et al., 2014) is a robust framework that ensures the privacy of individuals in a dataset by adding controlled noise to the data or to the output of data analyses, which becomes crucial when training machine learning models on sensitive information. In machine learning, integrating differential privacy helps in mitigating the risks of data leakage and model inversion attacks, ensuring that the model's predictions do not inadvertently reveal private information about any individual in the training data. With the advent of quantum machine learning, several works tried to quantize the basic concepts, definitions and results of differential privacy (Hirche et al., 2023; Angrisani et al., 2023; Angrisani and Kashefi, 2022; Nuradha et al., 2023; Aaronson and Rothblum, 2019; Zhou and Ying, 2017; Du et al., 2021). Classically, differentially private learners are known to satisfy mutual information stability (Feldman and Steinke, 2018), which can then be plugged into information-theoretic generalization bounds (see also (Hellström et al., 2023, Section 7.6)). Additionally, in the case of locally differentially private (LDP) classical learners, strong data processing inequalities (DPIs) have been established (see (Asoodeh and Zhang, 2022; Zamanlooy and Asoodeh, 2023; Angrisani et al., 2023) and the references therein), which also aid in controlling the entropic quantities appearing in our bounds. Here, we give proof-of-principle demonstrations for how recent quantum results on contraction properties of LDP channels and measurements (Hirche et al., 2023; Angrisani and Kashefi, 2022) can be used within our framework to analyze the generalization behavior of such quantum learners.

First, suppose that all channels $\Lambda^{\mathcal{A}}_{s,w}$ used by the learner $\mathcal{A}$ are ε-LDP (see (Hirche et al., 2023, Section V) or (Angrisani and Kashefi, 2022, Section 2.2) for a definition). Then, combining (Angrisani and Kashefi, 2022, Corollary 3.1) with the Pinsker inequality, we see that

$$I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}(s,w)} \leq 2\varepsilon(1-e^{-\varepsilon})\sqrt{2I(\mathrm{test};\mathrm{train})_{\rho^{\mathcal{A}}(s,w)}}\,. \tag{C.7.1}$$

Using Jensen's inequality, this means that the relevant expected QMI in our generalization bounds is upper bounded as

$$\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}\left[I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}(S,W)}\right] \leq 2\sqrt{2}\varepsilon(1-e^{-\varepsilon})\sqrt{\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}[I(\mathrm{test};\mathrm{train})_{\rho^{\mathcal{A}}(S,W)}]}\,. \tag{C.7.2}$$

To further upper bound the average QMI in the post-measurement states $\rho^{\mathcal{A}}(s,w)$, we can write $I(\mathrm{test};\mathrm{train})_{\rho^{\mathcal{A}}(S,W)}$ in terms of von Neumann entropies, and use concavity of the entropy as well as the definition of the Holevo information χ to arrive at

$$\mathop{\mathbb{E}}_{(S,W)\sim P^{\mathcal{A}}}[I(\mathrm{test};\mathrm{train})_{\rho^{\mathcal{A}}(S,W)}] \leq I(\mathrm{test};\mathrm{train})_{\mathbb{E}_{(S,W)\sim P^{\mathcal{A}}}[\rho^{\mathcal{A}}(S,W)]} + \chi\left(\{P^{\mathcal{A}}(s,w),\rho^{\mathcal{A}}(s,w)\}\right) \tag{C.7.3}$$

$$\leq I(\mathrm{test};\mathrm{train})_{\rho} + \chi\left(\{P^{\mathcal{A}}(s,w),\rho^{\mathcal{A}}(s,w)\}\right)\,, \tag{C.7.4}$$

where the last step used the data-processing inequality. Thus, we control the QMI contribution to the generalization error in terms of the initial QMI present in the data and a proxy for the maximum accessible information about the measurement outcomes accessible from the post-measurement ensemble. When performing a similar analysis for a learner using general (not ε-LDP) channels $\Lambda^{\mathcal{A}}_{s,w}$, a direct application of DPI yields the weaker $I(\mathrm{test};\mathrm{hyp})_{\sigma^{\mathcal{A}}(s,w)} \leq I(\mathrm{test};\mathrm{train})_{\rho^{\mathcal{A}}(s,w)}$ instead of Equation (C.7.1). Once we note that $1-e^{-\varepsilon} = \varepsilon + \mathcal{O}(\varepsilon^2)$, we obtain the following rule of thumb: We expect the QMI contribution to the generalization error to improve by a factor of $\mathcal{O}(\varepsilon^2)$ when using ε-LDP quantum channels.

Next, we turn our attention to the classical MI term in our generalization bounds. Here, we assume that the learner $\mathcal{A}$ uses an overall ε-LDP POVM. As the POVM $\{|s\rangle\langle s| \otimes E^{\mathcal{A}}_s(w)\}_{s,w}$ is not LDP even if every $\{E^{\mathcal{A}}_s(w)\}_w$ is, we make the simplifying assumption that the learner uses an s-independent ε-LDP POVM $\{E^{\mathcal{A}}(w)\}_w$. Then, we can write $I(S;W) = \mathbb{E}_{S\sim P^m}[D(P^{\mathcal{A}}_{\mathsf{W}|S}\|P^{\mathcal{A}}_{\mathsf{W}})]$, where $P^{\mathcal{A}}_{\mathsf{W}|S}$ is the outcome distribution when measuring $\{E^{\mathcal{A}}(w)\}_w$ on $\rho(S)$, and where $P^{\mathcal{A}}_{\mathsf{W}}$ is the outcome distribution when measuring $\{E^{\mathcal{A}}(w)\}_w$ on $\mathbb{E}_{\tilde{S}\sim P^m}[\rho(\tilde{S})]$. As we assume $\{E^{\mathcal{A}}(w)\}_w$ to be ε-LDP, (Angrisani and Kashefi, 2022, Lemma 3.1) now implies

$$I(S;W) \leq 2e^{\varepsilon}(1-e^{-\varepsilon})^2 \mathop{\mathbb{E}}_{S\sim P^m}\left[D\left(\rho(S)\middle\| \mathop{\mathbb{E}}_{\tilde{S}\sim P^m}[\rho(\tilde{S})]\right)\right] \tag{C.7.5}$$

$$= 2e^{\varepsilon}(1-e^{-\varepsilon})^2\,\chi\left(\{P^m(s),\rho(s)\}_{s\in\mathsf{Z}^m}\right)\,, \tag{C.7.6}$$

where the second step used Equation (22). So, the classical MI contribution to the generalization error is controlled by the Holevo information of the quantum data states. Again, compared to a general learner, we expect the classical MI contribution to the generalization error to be smaller by a factor of $\mathcal{O}(\varepsilon^2)$ when using an (s-independent) ε-LDP POVM.

In this subsection, we have used our generalization guarantees to show that requiring a quantum learner $\mathcal{A}$ to be ε-LDP – both in terms of the channels and the measurement used – is expected to be beneficial for generalization performance. While our discussion here already highlights the benefits of an LDP assumption for generalization in a broad sense, it would be interesting to instantiate this insight for specific quantum learning tasks of interest. Moreover, while our discussion focused on local differential privacy, it does not yet apply to differentially private quantum learners. The question of whether quantum differential privacy implies a version of mutual information stability useful for quantum generalization error bounds remains open. Finally, we have demonstrated how to use local differential privacy to control the classical and quantum mutual information terms in our generalization bounds. Investigating the effect of ε-LDP assumptions on the Holevo information term would give further insight into the relevance of ε-LDP to generalization when processing entangled quantum data.

C.8. Generalization bounds for inductive supervised quantum learning

(Monras et al., 2017) considered quantum learners described by multipartite quantum channels acting on quantum training data and on the input marginals of test states. Then, they defined the expected risk as the expectation value of a loss observable measured on the output of the learner and on the output marginals of the test states. We can formulate this in our framework as follows: We take a trivial classical instance space Z and consider the data Hilbert space

$$\begin{aligned}\mathcal{H}_{\text{data}} = \mathcal{H}_{\text{test}} \otimes \mathcal{H}_{\text{train}} &= \mathcal{H}_{\text{test,out}} \otimes (\mathcal{H}_{\text{test,in}} \otimes \mathcal{H}_{\text{train,in}}) &\text{(C.8.1)}\\ &= (\mathbb{C}^{d_{\text{out}}})^{\otimes m_{\text{test}}} \otimes ((\mathbb{C}^{d_{\text{in}}})^{\otimes m_{\text{test}}} \otimes (\mathbb{C}^{d})^{\otimes m_{\text{train}}}) . &\text{(C.8.2)}\end{aligned}$$

Then we take a quantum data state of the form

$$\rho = \rho_{\text{test}}^{\otimes m_{\text{test}}} \otimes \rho_{\text{train}} , \tag{C.8.3}$$

with $\rho_{\text{test}} \in \mathcal{S}(\mathbb{C}^{d_{\text{out}}} \otimes \mathbb{C}^{d_{\text{in}}})$. The goal of the learner is to use ρ_{train} to predict the mapping from the input to the output parts of the test systems.

We will consider quantum learners with hypothesis space $\mathcal{H}_{\text{hyp}} \cong \mathcal{H}_{\text{test,out}}$ that first perform a POVM $\{\mathbb{1}_{\text{test,in}} \otimes E^{\mathcal{A}}(w)\}_{w\in\mathsf{W}}$ that act non-trivially only on the $[\text{train}, \text{in}]$ subsystem, and, depending on the observed outcome, apply quantum processing of the form $(\Lambda_w^{\mathcal{A}})^{\otimes m_{\text{test}}} \otimes \text{id}_{\text{train,in}}$, with each $\Lambda_w^{\mathcal{A}} : \mathcal{T}_1(\mathbb{C}^{d_{\text{in}}}) \to \mathcal{T}_1(\mathbb{C}^{d_{\text{out}}})$ acting only on one of the $[\text{test}, \text{in}]$ subsystems. To measure the performance of such a learner, we use a local loss observable of the form

$$L(w) = \bar{L} = \frac{1}{m} \sum_{i=1}^{m_{\text{test}}} \mathbb{1}_{d_{\text{out}},d_{\text{out}}}^{\otimes(i-1)} \otimes L_0 \otimes \mathbb{1}_{d_{\text{out}},d_{\text{out}}}^{\otimes(m-i)} , \tag{C.8.4}$$

where $L_0 \in \mathcal{B}(\mathbb{C}^{d_{\text{out}}} \otimes \mathbb{C}^{d_{\text{out}}})$. With these choices, the expected empirical risk

$$\mathop{\mathbb{E}}_{W\sim P_{\mathsf{W}}^{\mathcal{A}}} \left[\text{Tr}\left[\bar{L}\sigma^{\mathcal{A}}(W)\right]\right] \tag{C.8.5}$$

reproduces what (Monras et al., 2017) simply call expected risk, whereas our expected true risk

$$\mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathsf{W}}}\left[\mathrm{Tr}\left[\bar{L}(\rho_{\mathrm{test,out}}^{\otimes m_{\mathrm{test}}}\otimes\sigma^{\mathcal{A}}(W)_{\mathrm{hyp}})\right]\right] \tag{C.8.6}$$

does not have a direct counterpart in (Monras et al., 2017). Note: While the inductive (i.e., "measure-then-process") learners that we consider here are not the most general form of quantum learner from ρ, we have a motivation for this focus. Namely, formulated in our language (Monras et al., 2017, Theorem 1) implies that, under a non-signalling assumption, quantum learners can approximately be assumed to be inductive. Here, the approximation is w.r.t. the expected empirical and true risks arising from a loss observable as in Equation (C.8.4) and improves with growing m_{test} because of a quantum de Finetti type behavior.

We can apply our generalization guarantees in this setting as follows: Notice that, since the POVM act trivially on $[\mathrm{test},\mathrm{in}]$ and since the quantum processing is a tensor power of single-system channels, both $\sigma^{\mathcal{A}}(w)$ and $\rho_{\mathrm{test,out}}\otimes\sigma^{\mathcal{A}}(w)_{\mathrm{hyp}}$ factorize according to the tensor product structure $\mathcal{H}_{\mathrm{test,out}}\otimes\mathcal{H}_{\mathrm{hyp}}\cong(\mathbb{C}^{d_{\mathrm{out}}}\otimes\mathbb{C}^{d_{\mathrm{out}}})^{\otimes m_{\mathrm{test}}}$. As the loss observable is local w.r.t. the same factorization, Corollary 24 applies and, simply using boundedness of L to get sub-gaussianity, yields the generalization bound

$$\left|\mathrm{gen}_{\rho}(\mathcal{A})\right|\leq\sqrt{\frac{C\|L\|^2}{m_{\mathrm{test}}}\mathop{\mathbb{E}}_{W\sim P^{\mathcal{A}}_{\mathsf{W}}}\left[\sum_{i=1}^{m_{\mathrm{test}}}I(\mathrm{test},\mathrm{out};\mathrm{hyp})_{\sigma_i^{\mathcal{A}}(W)}\right]}. \tag{C.8.7}$$

Thus, the framework of (Monras et al., 2017) fits naturally into our formulation, and this way our framework gives rise to a notion of generalization error that can be analyzed quantum information-theoretically. This, to the best of our knowledge, led us to the first generalization bound that applies to arbitrary inductive quantum learners.

Appendix D. Auxiliary Results and Proofs

Lemma 30 (Restatement of (De Palma and Trevisan, 2023, Theorem 8.1)) *Let H be a Hermitian m-qudit observable. Let $\rho=\bigotimes_{i=1}^m\rho_i$ be an m-fold tensor product of qudit states. Then, for any $\lambda\in\mathbb{R}$,*

$$\mathrm{Tr}[e^{\log(\rho)+\lambda H}]\leq e^{\frac{\lambda^2 m\|H\|^2_{\mathrm{Lip}}}{2}}. \tag{D.1}$$

Proof [Proof of Corollary 24] First, note that the sub-gaussianity assumption on the $L_i(z_i,w)$ implies

$$\log\mathrm{Tr}\left[\left(\rho_{\mathrm{test}}(s)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w)\right)\cdot e^{\lambda\left(L(s,w)-\mathrm{Tr}[L(s,w)\left(\rho_{\mathrm{test}}(s)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w)\right)]\mathbb{1}_{\mathrm{test,hyp}}\right)}\right] \tag{D.2}$$

$$=\sum_{i=1}^m\log\mathrm{Tr}\left[\left(\rho_{\mathrm{test},i}(Z_i)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp},i}(z_i,w)\right)\cdot e^{\frac{\lambda}{m}\left(L_i(z_i,w)-\mathrm{Tr}[L_i(z_i,w)\left(\rho_{\mathrm{test},i}(Z_i)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp},i}(z_i,w)\right)]\mathbb{1}_{\mathrm{test,hyp},i}\right)}\right] \tag{D.3}$$

$$\leq\sum_{i=1}^m\frac{\alpha_i^2\lambda^2}{2m^2}. \tag{D.4}$$

Therefore, $L(s,w)$ is α-sub-gaussian w.r.t. $\rho_{\mathrm{test}}(s)\otimes\sigma^{\mathcal{A}}_{\mathrm{hyp}}(s,w)$ with sub-gaussianity parameter $\alpha=m^{-1}\sqrt{\sum_{i=1}^m\alpha_i^2}$, for every $(s,w)\in\mathsf{Z}^m\times\mathsf{W}$.

Using the sub-gaussianity assumption on the $\mathrm{Tr}[L_i(Z_i,w)\left(\rho_{\text{test},i}(Z_i)\otimes\sigma^{\mathcal{A}}_{\text{hyp},i}(Z_i,w)\right)]$, we see that

$$\log \mathop{\mathbb{E}}_{S\sim P^m}\left[e^{\lambda(\mathrm{Tr}[L(S,w)\left(\rho_{\text{test}}(S)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(S,w)\right)]-\mathbb{E}_{S\sim P^m}[\mathrm{Tr}[L(S,w)\left(\rho_{\text{test}}(S)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(S,w)\right)]])}\right] \tag{D.5}$$

$$= \sum_{i=1}^{m}\log \mathop{\mathbb{E}}_{Z_i\sim P}\left[e^{\frac{\lambda}{m}(\mathrm{Tr}[L_i(Z_i,w)\left(\rho_{\text{test},i}(Z_i)\otimes\sigma^{\mathcal{A}}_{\text{hyp},i}(Z_i,w)\right)]-\mathbb{E}_{Z_i\sim P}[\mathrm{Tr}[L_i(Z_i,w)\left(\rho_{\text{test},i}(Z_i)\otimes\sigma^{\mathcal{A}}_{\text{hyp},i}(Z_i,w)\right)]])}\right] \tag{D.6}$$

$$\leq \sum_{i=1}^{m}\frac{\beta_i^2\lambda^2}{2m^2}. \tag{D.7}$$

In other words, $\mathrm{Tr}[L(S,w)\left(\rho_{\text{test}}(S)\otimes\sigma^{\mathcal{A}}_{\text{hyp}}(S,w)\right)]$, with $S\sim P^m$, is β-sub-gaussian with sub-gaussianity parameter $\beta = m^{-1}\sqrt{\sum_{i=1}^{m}\beta^2}$, for every $w\in\mathsf{W}$. Therefore, we can apply Corollary 23 and obtain the claimed bound, once we use that $\sigma^{\mathcal{A}}(s,w) = \bigotimes_{i=1}^{m}\sigma_i^{\mathcal{A}}(z_i,w)$ implies $I(\text{test};\text{hyp})_{\sigma^{\mathcal{A}}(s,w)} = \sum_{i=1}^{m} I(\text{test};\text{hyp})_{\sigma_i^{\mathcal{A}}(z_i,w)}$. ■

Proof [Proof of Corollary 26] On the one hand, we have

$$\left|\mathop{\mathbb{E}}_{\bar{w}}\mathop{\mathbb{E}}_{\bar{Z}_{m+1}}\left[\left|\mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0]-\mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(\bar{w})]\right|\right]-R_\rho(\mathcal{A})\right| \tag{D.8}$$

$$\leq \mathop{\mathbb{E}}_{\bar{Z}_{m+1}}\mathop{\mathbb{E}}_{\bar{C}_\ell^{(m+1)}|\bar{Z}_{m+1}}\left[\left|\mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0]-\frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}}\bar{C}_\ell^{(m+1)}\right|\right] \tag{D.9}$$

$$\leq \frac{C}{\sqrt{m_{\text{test}}}}, \tag{D.10}$$

where the first step is an application of the reverse triangle inequality and the second step is via first conditioning on $\bar{Z}_{m+1}$ and then using Hoeffding-based sub-gaussianity, as already argued in Appendix C.1. On the other hand, we have

$$\hat{R}_\rho(\mathcal{A}) = \mathop{\mathbb{E}}_{S\sim P^{2m}}\mathop{\mathbb{E}}_{C_\ell^{(i)}|S}\mathop{\mathbb{E}}_{\hat{W}}\left[\hat{R}^{\text{test}}_{S_{(m+1):2m},C_\ell^{(i)}}(\hat{W})\right] \tag{D.11}$$

$$= \mathop{\mathbb{E}}_{S\sim P^{2m}}\mathop{\mathbb{E}}_{C_\ell^{(i)}|S}\mathop{\mathbb{E}}_{\hat{W}}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\left|\mathrm{Tr}[E(Z_i)\rho_0(\hat{W})]-\frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}}C_\ell^{(i)}\right|\right] \tag{D.12}$$

$$\leq \mathop{\mathbb{E}}_{S\sim P^{2m}}\mathop{\mathbb{E}}_{B_\ell^{(i)}|S}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\left|\mathrm{Tr}[E(Z_i)\rho_0(\hat{W})]-\frac{1}{m_{\text{train}}}\sum_{\ell=1}^{m_{\text{train}}}B_\ell^{(i)}\right|\right] \tag{D.13}$$

$$+ \mathop{\mathbb{E}}_{S\sim P^{2m}}\mathop{\mathbb{E}}_{B_\ell^{(i)}|S}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\left|\mathrm{Tr}[E(Z_i)\rho_0]-\frac{1}{m_{\text{train}}}\sum_{\ell=1}^{m_{\text{train}}}B_\ell^{(i)}\right|\right] \tag{D.14}$$

$$+ \mathop{\mathbb{E}}_{S\sim P^{2m}}\mathop{\mathbb{E}}_{C_\ell^{(i)}|S}\left[\frac{1}{m}\sum_{i=m+1}^{2m}\left|\mathrm{Tr}[E(Z_i)\rho_0]-\frac{1}{m_{\text{test}}}\sum_{\ell=1}^{m_{\text{test}}}C_\ell^{(i)}\right|\right] \tag{D.15}$$

$$= \mathop{\mathbb{E}}_{S\sim P^{2m}}\mathop{\mathbb{E}}_{B_\ell^{(i)}|S}\left[\inf_{w\in\mathsf{W}_1}\frac{1}{m}\sum_{i=m+1}^{2m}\left|\mathrm{Tr}[E(Z_i)\rho_0(w)]-\frac{1}{m_{\text{train}}}\sum_{\ell=1}^{m_{\text{train}}}B_\ell^{(i)}\right|\right] \tag{D.16}$$

$$+ \mathop{\mathbb{E}}_{S\sim P^{2m}} \mathop{\mathbb{E}}_{B_\ell^{(i)}|S} \left[\frac{1}{m} \sum_{i=m+1}^{2m} \left| \mathrm{Tr}[E(Z_i)\rho_0] - \frac{1}{m_{\text{train}}} \sum_{\ell=1}^{m_{\text{train}}} B_\ell^{(i)} \right| \right] \tag{D.17}$$

$$+ \mathop{\mathbb{E}}_{S\sim P^{2m}} \mathop{\mathbb{E}}_{C_\ell^{(i)}|S} \left[\frac{1}{m} \sum_{i=m+1}^{2m} \left| \mathrm{Tr}[E(Z_i)\rho_0] - \frac{1}{m_{\text{test}}} \sum_{\ell=1}^{m_{\text{test}}} C_\ell^{(i)} \right| \right] \tag{D.18}$$

$$\leq \mathop{\mathbb{E}}_{S\sim P^{2m}} \left[\inf_{w\in\mathsf{W}_1} \frac{1}{m} \sum_{i=m+1}^{2m} |\mathrm{Tr}[E(Z_i)\rho_0(w)] - \mathrm{Tr}[E(Z_i)\rho_0]| \right] \tag{D.19}$$

$$+ 2 \mathop{\mathbb{E}}_{S\sim P^{2m}} \mathop{\mathbb{E}}_{B_\ell^{(i)}|S} \left[\frac{1}{m} \sum_{i=m+1}^{2m} \left| \mathrm{Tr}[E(Z_i)\rho_0] - \frac{1}{m_{\text{train}}} \sum_{\ell=1}^{m_{\text{train}}} B_\ell^{(i)} \right| \right] \tag{D.20}$$

$$+ \mathop{\mathbb{E}}_{S\sim P^{2m}} \mathop{\mathbb{E}}_{C_\ell^{(i)}|S} \left[\frac{1}{m} \sum_{i=m+1}^{2m} \left| \mathrm{Tr}[E(Z_i)\rho_0] - \frac{1}{m_{\text{test}}} \sum_{\ell=1}^{m_{\text{test}}} C_\ell^{(i)} \right| \right] \tag{D.21}$$

$$\leq \mathop{\mathbb{E}}_{S\sim P^{2m}} \left[\inf_{w\in\mathsf{W}_1} \frac{1}{m} \sum_{i=m+1}^{2m} |\mathrm{Tr}[E(Z_i)\rho_0(w)] - \mathrm{Tr}[E(Z_i)\rho_0]| \right] \tag{D.22}$$

$$+ C \left(\frac{1}{\sqrt{m_{\text{train}}}} + \frac{1}{\sqrt{m_{\text{test}}}} \right). \tag{D.23}$$

Here, the first step is plugging in the definition of $\hat{R}_{S_{(m+1):2m}, C_\ell^{(i)}}(\cdot)$, the second step holds by applying the triangle inequality twice, the third step uses the definition of $\hat{W}$, the fourth step is one more triangle inequality, and the final step follows from Hoeffding-type sub-gaussianity bounds.

To finish the proof, we need the following fact:

Claim 31 *With the notation introduced above,*

$$\mathop{\mathbb{E}}_{S\sim P^{2m}} \left[\inf_{w\in\mathsf{W}_1} \frac{1}{m} \sum_{i=m+1}^{2m} |\mathrm{Tr}[E(Z_i)\rho_0(w)] - \mathrm{Tr}[E(Z_i)\rho_0]| \right] \tag{D.24}$$

$$\leq \inf_{w\in\mathsf{W}} \mathop{\mathbb{E}}_{\bar{Z}_{m+1}} \left[\left| \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(w)] \right| \right] + \tilde{\varepsilon} + C\sqrt{\frac{\log(|\mathsf{W}_1|)}{m}}. \tag{D.25}$$

where $C > 0$ is some positive constant.

Proof See below. ■

Combining Theorem 31 with our previous upper bound on $\hat{R}_\rho(\mathcal{A})$, we have shown

$$\hat{R}_\rho(\mathcal{A}) \leq \inf_{w\in\mathsf{W}} \mathop{\mathbb{E}}_{\bar{Z}_{m+1}} \left[\left| \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(w)] \right| \right] + \tilde{\varepsilon} \tag{D.26}$$

$$+ C \left(\sqrt{\frac{\log(|\mathsf{W}_1|)}{m}} + \frac{1}{\sqrt{m_{\text{train}}}} + \frac{1}{\sqrt{m_{\text{test}}}} \right). \tag{D.27}$$

Finally, once we recall the bound $|\mathsf{W}_1| \leq (2/\tilde{\varepsilon})^{C\log(d)/\tilde{\varepsilon}^2}$ on the size of the covering net, we can bring together our upper bound on $R_\rho(\mathcal{A})$, our upper bound on $\hat{R}_\rho(\mathcal{A})$, and our generalization error bound to obtain

$$\mathrm{excess}_\rho(\mathcal{A}) \leq \mathrm{gen}_\rho(\mathcal{A}) + \tilde{\varepsilon} + \mathcal{O}\left(\sqrt{\frac{\log(|\mathsf{W}_1|)}{m}} + \sqrt{\frac{\log(|\mathsf{W}_1|)}{m_{\mathrm{train}}}} + \frac{1}{\sqrt{m_{\mathrm{test}}}}\right) \tag{D.28}$$

$$\leq \tilde{\varepsilon} + \tilde{\mathcal{O}}\left(\sqrt{\frac{\log(d)}{m\tilde{\varepsilon}^2}} + \frac{1}{\sqrt{m_{\mathrm{train}}}} + \frac{1}{\sqrt{m_{\mathrm{test}}}}\right), \tag{D.29}$$

as claimed. ■

Proof [Proof of Theorem 31] First recall that the chosen covering net W_1 depends only on $Z_1, \dots, Z_m$, so that we can exchange $\mathbb{E}_{Z_{m+1},\dots,Z_{2m}\sim P^m}$ and $\inf_{w\in\mathsf{W}_1}$ to obtain the bound

$$\mathop{\mathbb{E}}_{S\sim P^{2m}}\left[\inf_{w\in\mathsf{W}_1}\frac{1}{m}\sum_{i=m+1}^{2m}|\mathrm{Tr}[E(Z_i)\rho_0(w)] - \mathrm{Tr}[E(Z_i)\rho_0]|\right] \tag{D.30}$$

$$\leq \mathop{\mathbb{E}}_{(Z_1,\dots,Z_m)\sim P^m}\left[\inf_{w\in\mathsf{W}_1}\mathop{\mathbb{E}}_{(Z_{m+1},\dots,Z_{2m})\sim P^m}\left[\frac{1}{m}\sum_{i=m+1}^{2m}|\mathrm{Tr}[E(Z_i)\rho_0(w)] - \mathrm{Tr}[E(Z_i)\rho_0]|\right]\right] \tag{D.31}$$

$$= \mathop{\mathbb{E}}_{(Z_1,\dots,Z_m)\sim P^m}\left[\inf_{w\in\mathsf{W}_1}\mathop{\mathbb{E}}_{\bar{Z}_{m+1}\sim P}\left[\left|\mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(w)] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0]\right|\right]\right]. \tag{D.32}$$

Next, we define the following pieces of notation for the true risk with perfectly accurately evaluated quantum expectation values

$$R(w) := \mathop{\mathbb{E}}_{\bar{Z}_{m+1}\sim P}\left[\left|\mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(w)] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0]\right|\right], \tag{D.33}$$

the empirical risk with perfectly accurately evaluated quantum expectation values

$$\hat{R}_s(w) := \frac{1}{m}\sum_{j=1}^{m}|\mathrm{Tr}[E(z_j)\rho_0(w)] - \mathrm{Tr}[E(z_j)\rho_0]|, \tag{D.34}$$

and the corresponding true risk minimizers

$$w_{\mathsf{W}_1} \in \mathrm{argmin}_{w\in\mathsf{W}_1} R(w),\ w_{\mathsf{W}} \in \mathrm{argmin}_{w\in\mathsf{W}} R(w), \tag{D.35}$$

and empirical risk minimizers

$$\hat{w}_{\mathsf{W}_1} \in \mathrm{argmin}_{w\in\mathsf{W}_1} \hat{R}_s(w),\ \hat{w}_{\mathsf{W}} \in \mathrm{argmin}_{w\in\mathsf{W}} \hat{R}_s(w). \tag{D.36}$$

With this, we can rewrite and bound

$$\mathop{\mathbb{E}}_{(Z_1,\dots,Z_m)\sim P^m}\left[\inf_{w\in\mathsf{W}_1}\mathop{\mathbb{E}}_{\bar{Z}_{m+1}\sim P}\left[\left|\mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(w)] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0]\right|\right]\right] \tag{D.37}$$

$$- \inf_{w\in\mathsf{W}}\mathop{\mathbb{E}}_{\bar{Z}_{m+1}}\left[\left|\mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0] - \mathrm{Tr}[E(\bar{Z}_{m+1})\rho_0(w)]\right|\right] \tag{D.38}$$

$$= \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}[R(w_{\mathsf{W}_1})] - R(w_{\mathsf{W}}) \tag{D.39}$$

$$= \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}[R(w_{\mathsf{W}_1}) - R(w_{\mathsf{W}})] \tag{D.40}$$

$$= \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}[R(w_{\mathsf{W}_1}) - R(w_{\mathsf{W}})] + \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\hat{R}(\hat{w}_{\mathsf{W}_1}) - \hat{R}(w_{\mathsf{W}})\right] \tag{D.41}$$

$$- \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\hat{R}(\hat{w}_{\mathsf{W}_1}) - \hat{R}(w_{\mathsf{W}})\right] \tag{D.42}$$

$$= \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[R(w_{\mathsf{W}_1}) - \hat{R}(\hat{w}_{\mathsf{W}_1})\right] + \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\hat{R}(\hat{w}_{\mathsf{W}_1}) - \hat{R}(w_{\mathsf{W}})\right] \tag{D.43}$$

$$+ \underbrace{\mathop{\mathbb{E}}_{S\sim P^m}\left[\hat{R}(w_{\mathsf{W}}) - R(w_{\mathsf{W}})\right]}_{=0} \tag{D.44}$$

$$= \underbrace{\mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}[R(w_{\mathsf{W}_1}) - R(\hat{w}_{\mathsf{W}_1})]}_{\leq 0} + \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[R(\hat{w}_{\mathsf{W}_1}) - \hat{R}(\hat{w}_{\mathsf{W}_1})\right] \tag{D.45}$$

$$+ \underbrace{\mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\hat{R}(\hat{w}_{\mathsf{W}}) - \hat{R}(w_{\mathsf{W}})\right]}_{\leq 0} + \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\hat{R}(\hat{w}_{\mathsf{W}_1}) - \hat{R}(\hat{w}_{\mathsf{W}})\right] \tag{D.46}$$

$$\leq \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\sup_{w\in\mathsf{W}_1} R(w) - \hat{R}(w)\right] + \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\|\hat{w}_{\mathsf{W}_1} - \hat{w}_{\mathsf{W}}\|_{1,\{Z_j\}_{j=1}^m}\right] \tag{D.47}$$

$$\leq \mathop{\mathbb{E}}_{(Z_1,\ldots,Z_m)\sim P^m}\left[\sup_{w\in\mathsf{W}_1} R(w) - \hat{R}(w)\right] + \tilde{\varepsilon}. \tag{D.48}$$

Here, the second-to-last step used a reverse triangle inequality, and the final step holds because W_1 is by definition a $\tilde{\varepsilon}$-covering net for W w.r.t. $\|\cdot\|_{2,\{Z_j\}_{j=1}^m}$ and thus also w.r.t. $\|\cdot\|_{1,\{Z_j\}_{j=1}^m}$. Next, observe that, for any fixed $w \in \mathsf{W}_1$, the random variable $R(w) - \hat{R}(w)$ is an average of m i.i.d. centered 2-bounded random variables and thus is $(\frac{C}{\sqrt{m}})$-sub-gaussian by Hoeffding's Lemma. Using the equivalence of sub-gaussianity in terms of MGF bounds and tail bounds (Vershynin, 2018, Proposition 2.5.2), this can now be combined with a union bound over W_1 to see that the random variable $\sup_{w\in\mathsf{W}_1} R(w) - \hat{R}(w)$ is $(C\sqrt{\frac{\log(|\mathsf{W}_1|)}{m}})$-sub-gaussian. Therefore, using again the L_p bound version of sub-gaussianity (Vershynin, 2018, Proposition 2.5.2), we conclude

$$\mathop{\mathbb{E}}_{S\sim P^m}\left[\sup_{w\in\mathsf{W}_1} R(w) - \hat{R}(w)\right] \leq C\sqrt{\frac{\log(|\mathsf{W}_1|)}{m}}. \tag{D.49}$$

Plugging this into our previous bound and rearranging, we get the claimed inequality. ■