AN INFORMATION BOTTLENECK FOR LEARNING QUANTUM CHANNELS VIA COHERENT INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalization theory for learning from quantum data has recently been unified under an information theoretic lens with bounds controlling the generalization gap via the mutual information between the hypothesis of a learner and its training data. However, extension of such a framework to dynamic quantum channels plausibly the more practically relevant learning target. Given the ubiquity of parameterized quantum circuits as hypothesis classes for unknown unitary dynamics, it has remained an open problem, and existing bounds for this setting rely on abstract quantum mutual information terms that cannot be measured or actively controlled during training. In this paper, we introduce a framework, Coherently Regularized Quantum Channel learning that replaces this unmeasurable quantity with the coherent information of the hypothesis channel, a defined estimable capacity metric. Using a decoupled reference state construction and quantum optimal transport techniques, and Petz's variational for quantum relative entropy, we prove a high probability generation bound. These bounds confirm that the true risk of a learned hypothesis channel is controlled by its empirical risk and a scaling with its maximum coherent information. Maximum coherent information is measurable, resulting a training time regularizer that acts as an information bottleneck, perturbing hypothesis channels from encoding spurious, query-specific quantum correlations. We validate our approach for learning unknown random Clifford circuits under a restricted query budget, which showcases that coherently regularized training yields a smaller generalization gap than unregularized empirical risk minimization as a function of the gap tracking, and that shallow target circuits are learnable from substantially fewer queries.

1 INTRODUCTION

Recent advances in machine learning and quantum physics has garnered significant traction in the area of quantum circuit models for classical data such as variational quantum models, quantum kernel methods Biamonte et al. (2017); Dunjko & Briegel (2018); Havlíček et al. (2019), and learning from quantum data Caro et al. (2024). Inspired by pretty good tomography Aaronson (2007), the latter view has given rise to shadow protocols that predict many properties of an unknown state from few copies Aaronson (2018); Huang et al. (2020) showcased the potential for quantum advantage in quantum experiments Huang et al. (2021); Aharonov et al. (2022), and shown that quantum theory allows new kinds of oracular access to unknown objects Bshouty & Jackson (1995) with implications for learning efficiency Arunachalam & De Wolf (2017). Fundamental quantum information tasks such as state tomography, process tomography, state discrimination can now be interpreted as instances of learning from quantum data Guţă & Kotłowski (2010); Sentís et al. (2019). Given the recent works, a fragmented landscape of sample complexity and generalization results, which is analogous to the pre-unification state of classical learning theory. Information theoretic generalization bounds have begun to play a similarly unifying role classically Xu & Raginsky (2017); Russo & Zou (2019); Hellström et al. (2025), with analogous unifying frameworks recently introduced for quantum state classification and state preparation Caro et al. (2024), addressing limitations of uniform bounds recently noted in quantum models as well Gil-Fuster et al. (2024).

Extending these approaches to dynamic quantum channels, where the goal is to learn an unknown unitary or noisy process instead of a static state remains comparatively undeveloped. Caro et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2024) pose this as an open problem, and their own machinery, when it does apply, bounds the generalization gap via an expected quantum mutual information term $\mathbb{E}[I(\text{test};\text{hyp})]$ that is in general not something a practitioner can easily measure or actively minimize during training.

Our contribution. We replace this abstract quantity with the coherent information I_c of the hypothesis channel, a standard, physically defined, and measurable capacity metric that can serve as an active training time regularizer. Applying this metric reveals a crucial subtlety that, if left unaddressed, would invalidate the entire formulation for the natural setting of learning an unknown unitary with a PQC hypothesis: we prove (Proposition 2) that $I_c(\hat{\mathcal{N}}, \psi)$ is strictly constant across all unitaries $\hat{\mathcal{N}}$ and all probe states ψ, yielding an identically zero regularizer gradient. We demonstrate this degeneracy both analytically and numerically.

2 Related Work

The study of generalization via information theoretic approaches traces to Xu & Raginsky (2017) and Russo & Zou (2019), who bound the expected generalization gap of a learning algorithm by the mutual information between training data and learned hypothesis. Recent work Hellström et al. (2025); Caro et al. (2024) translate this perspective to quantum learners, unifying quantum state classification and state preparation under a common framework. These foundational results primarily address static quantum data and express their bounds via mutual information quantities that, while well defined are generally hard to estimate or optimize directly during training of a parameterized model.

Characterizing dynamic quantum processes i.e. channels is a distinct and central challenge. Full process tomography requires resources scaling exponentially with system size. Recent work reframes channel characterization as a learning task, extending classical shadow based methods Huang et al. (2020) to restricted query learning of unknown processes, with parameterized quantum circuits (PQCs) widely used as the hypothesis class for approximating unknown unitary dynamics. Such ansatz based approaches can overfit severely in data scarce regimes, particularly once the depth of the hypothesis circuit and expressivity exceed that of the target. Also, it addresses the trainability of PQCs as system size grows. Optimization landscapes for global cost functions, where the full multi-qubit state is measured at once exhibit barren plateaus, such as gradient variance vanishing exponentially in the qubit count McClean et al. (2018). Cerezo et al. (2021) show that local cost functions built from few-qubit observables avoid this exponential suppression at shallow depth. Khatri et al. (2019) develop a matching local Hilbert-Schmidt test for quantum circuit compiling, which we adapt as our training time surrogate detailed in Section 3.7. In line of statistical complexity of PQC hypothesis classes directly, Caro & Datta (2020) bound the pseudo dimension of quantum circuits in terms of gate count, and Caro et al. (2022) give covering number based generalization bounds for quantum machine learning models scaling as $O(\sqrt{T \log T/m})$ in the number of trainable gates T and sample size m. We invoke this work in Section 3.4.2 to provide a complementary complexity control explanation alongside our entangling power regularizer. Coherent information and the degradable/antidegradable channel dichotomy we use in Proposition 3 are standard quantum Shannon theory Devetak (2005); Wilde (2017). The notion of a unitary's *entangling power* on a product input, which we use to motivate our practical regularizer without overclaiming it to be coherent information is due to Zanardi (2001).

Our work bridges these strands. We extend information theoretic generalization theory to dynamic channel learning. We identify and resolve a degeneracy that a naive coherent information construction suffers for unitary hypotheses, and, integrating the local cost trainability trick, we provide a regularizer together with empirical validation at both the scale originally proposed and an extended scale.

3 Coherently Regularized Quantum Channel Learning

3.1 Problem Formulation

Let $\mathcal{N} : \mathcal{B}(\mathcal{H}_A) \to \mathcal{B}(\mathcal{H}_{A'})$ be an unknown target channel. The learner outputs a parameterized hypothesis channel $\hat{\mathcal{N}}$ after m independent training queries. To probe the quantum capacity of a

channel, we use a pure bipartite state $|\psi\rangle_{RA}$ across a reference register R and an input register A; register A is sent through the channel while R is held aside. The output of the hypothesis channel on one query is $\sigma_{RA'} = (\mathrm{id}_R \otimes \hat{\mathcal{N}})(|\psi\rangle\langle\psi|_{RA})$.

3.2 Risk formulation and the loss observable

For a single query, we evaluate performance with a bounded loss observable L satisfying a quantum 1-Lipschitz condition under the Wasserstein-1 distance ($\|L\|_{\mathrm{Lip}} \leq 1$). Write $\hat{R}^{(1)}(\hat{\mathcal{N}}) = \mathrm{Tr}[L\sigma_{RA'}]$ for the single-query empirical risk and $R^{(1)}_{\mathrm{true}}(\hat{\mathcal{N}}) = \mathrm{Tr}[L\tau]$ for the single query true risk evaluated on the decoupled counterfactual state

$$\tau = \rho_R \otimes \hat{\mathcal{N}}(\rho_A), \qquad \rho_R = \mathrm{Tr}_A[|\psi\rangle\langle\psi|], \quad \rho_A = \mathrm{Tr}_R[|\psi\rangle\langle\psi|].$$

For m i.i.d. queries, we define the empirical risk as the sample average $\hat{R}_{\mathrm{emp}}(\hat{\mathcal{N}}) = \frac{1}{m}\sum_{i=1}^{m} \hat{R}^{(i)}(\hat{\mathcal{N}})$ and the true risk as the corresponding population average $R_{\mathrm{true}}(\hat{\mathcal{N}}) = \mathbb{E}[\hat{R}^{(1)}_{\mathrm{true}}]$ identical across i by the i.i.d. assumption. A detailed explanation about m-scaling is outlined in Section 3.5.

3.3 Theoretical Guarantees

Theorem 1 (Single-query Expected Generalization error)**.** *Let $\mathcal{N}$ be an unknown target channel and $\hat{\mathcal{N}}$ a hypothesis channel evaluated on one query prepared as a pure bipartite probe state $|\psi\rangle_{RA}$ with the loss observable L as in Section 3.2. Then,*

$$\mathbb{E}\Big[\big|R^{(1)}_{\mathrm{true}}(\hat{\mathcal{N}}) - \hat{R}^{(1)}(\hat{\mathcal{N}})\big|\Big] \leq \sqrt{2\big(H(\rho_R) + I_c(\hat{\mathcal{N}}, |\psi\rangle)\big)},$$

where $I_c(\hat{\mathcal{N}}, |\psi\rangle) = H(\hat{\mathcal{N}}(\rho_A)) - H(\sigma_{RA'})$ is the coherent information of $\hat{\mathcal{N}}$ at input $|\psi\rangle$.

We defer the proof to Appendix A.1, which follows the original strategy of combining Petz's variational formula for the quantum relative entropy Petz (1988) with the Golden-Thompson inequality and a quantum Lipschitz sub-Gaussian moment bound De Palma & Trevisan (2023) applied to a single query. We deliberately do not multiply this bound by any m-fold tensor structure at this stage. Section 3.5 explains why doing so leads to an m-query bound.

3.4 The unitary degeneracy

Proposition 2 (Degeneracy for Unitary Hypotheses)**.** *If $\hat{\mathcal{N}}(\rho) = \hat{U}\rho\hat{U}^\dagger$ for any unitary $\hat{U}$, in particular, any unitary PQC hypothesis, for every value of its parameters θ, then for every pure probe $|\psi\rangle_{RA}$,*

$$I_c(\hat{\mathcal{N}}, |\psi\rangle) = H(\rho_R),$$

independent of $\hat{U}$. Also, the bound of Theorem 1 equals $\sqrt{4H(\rho_R)}$ for every unitary hypothesis, and the regularizer $\lambda \cdot \max_{|\psi\rangle\in S} I_c(\hat{\mathcal{N}}, |\psi\rangle)$ has zero gradient with respect to θ, for every S, every circuit depth, and every qubit count.

Proof. Unitary conjugation preserves the spectrum, so $H(\hat{\mathcal{N}}(\rho_A)) = H(\hat{U}\rho_A\hat{U}^\dagger) = H(\rho_A)$ for every $\hat{U}$. A global unitary applied to a pure state remains pure, so $\sigma_{RA'} = (\mathrm{id}_R \otimes \hat{U})|\psi\rangle\langle\psi|(\mathrm{id}_R \otimes \hat{U}^\dagger)$ is pure and $H(\sigma_{RA'}) = 0$. By the Schmidt decomposition of the pure state $|\psi\rangle_{RA}$, $H(\rho_A) = H(\rho_R)$. Hence $I_c(\hat{\mathcal{N}}, |\psi\rangle) = H(\rho_A) - 0 = H(\rho_R)$, with no dependence on $\hat{U}$ at all, for any $|\psi\rangle \in S$, and so the maximum over S inherits the same independence. □

We validate this numerically to machine precision reported in the experiment for a fixed maximally entangled probe on $n = 3$ qubits, the quantum mutual information $I(R; A')_\sigma = H(\rho_R) + I_c$ evaluates to $2n = 6$ bits for many independently sampled sets of hypothesis parameters, including random restarts far from initialization.

Proposition 2 establishes that for an experimental protocol trained on unitary PQC hypotheses, a bound and regularizer built directly on $I_c(\hat{\mathcal{N}}, \psi)$ carry no information about the learned hypothesis. We repair this along two independent, complementary routes.

3.4.1 NOISY HYPOTHESIS CHANNELS

Proposition 3 (Non-degeneracy under physical noise)**.** *Let $\hat{\mathcal{N}}_\varepsilon = \mathcal{E}_\varepsilon \circ \hat{U}(\theta)$, where $\hat{U}(\theta)$ is a unitary PQC and $\mathcal{E}_\varepsilon$ is any fixed, non-unitary local noise channel, for instance, single-qubit depolarizing or dephasing noise at rate $\varepsilon > 0$, applied per gate or per layer with Stinespring dilation into an environment E. Then for certain θ, $I_c(\hat{\mathcal{N}}_\varepsilon, |\psi\rangle) \neq H(\rho_R)$ and depends non-trivially on θ. Theorem 1 is a hypothesis-discriminating bound.*

Proof sketch. With noise, $\sigma_{RA'} = \mathrm{Tr}_E[(\mathrm{id}_R \otimes V_\theta)|\psi\rangle\langle\psi|_{RA}(\mathrm{id}_R \otimes V_\theta^\dagger)]$ for the Stinespring isometry V_θ of $\hat{\mathcal{N}}_\varepsilon$ is generally mixed even though the global R, A, E state is pure, because R is correlated with the discarded environment E as well as with A'. Hence $H(\sigma_{RA'}) > 0$ in general, fragmenting the notion of purity used in Proposition 2. Moreover $H(\hat{\mathcal{N}}_\varepsilon(\rho_A))$ depends on θ whenever $\mathcal{E}_\varepsilon$ is not input-independent. This is true of any physically realistic per-qubit noise model, which acts identically regardless of θ, but whose effect on the output entropy depends on what state $\hat{U}(\theta)$ produced just before the noise. A fully worked single-qubit dephasing example, computed in closed form via the Kraus operator formula for coherent information is given in Appendix 4.1, confirming genuine θ dependence numerically. □

This approach keeps Theorem 1 and its regularizer as originally intended at the realistic, NISQ-motivated cost of modeling the hypothesis with gate noise instead of as an idealized perfect unitary.

3.4.2 AN ENTANGLING POWER REGULARIZER FOR UNITARY HYPOTHESES

Our experiments in Section 4, matching the originally proposed protocol, train a unitary PQC hypothesis with no explicit noise model. For this regime, we introduce a distinct, honestly labeled quantity.

Definition 4 (Entangling Power regularizer)**.** *We fix a bipartition of the n qubit input register $A = A_1 \otimes A_2$, where we use a balanced $\lfloor n/2 \rfloor \,|\, \lceil n/2 \rceil$ split. For a product state query $|\psi\rangle_i\rangle = |\psi_{1,i}\rangle \otimes |\psi_{2,i}\rangle$ and hypothesis $\hat{U}(\theta)$, we define the per query output entanglement entropy $S_{A_1}(\hat{U}(\theta)|\psi_i\rangle)$ across the cut, and the batch averaged regularizer*

$$\hat{\chi}(\theta) \;=\; \frac{1}{m}\sum_{i=1}^{m} S_{A_1}\big(\hat{U}(\theta)|\psi_i\rangle\big).$$

By a standard identity, for a pure bipartite state, $S_{A_1}(\hat{U}|\psi\rangle)$ equals the coherent information of the erasure channel $A_1 \to A_1'$ obtained by discarding A_2. This is why we describe $\hat{\chi}$ in terms of coherent information. We emphasize, however, that $\hat{\chi}$ is not an instance of $I_c(\hat{\mathcal{N}}, \psi)$ from Theorem 1 evaluated on a pure input; by the exact reasoning of Proposition 2, that quantity is identically zero. Instead, $\hat{\chi}$ is a related, well-established, and independently meaningful notion—the entangling power of $\hat{U}$ on $|\psi\rangle$ (Zanardi, 2001)—which we adopt because it is provably non-degenerate. As detailed in Appendix A.2, unlike $I_c(\hat{\mathcal{N}}, \psi)$, local unitaries on the A_2 side of the cut actively generate this quantity, so it is not conjugation-invariant. Furthermore, it is cheaply computable for up to $n \sim 18$ qubits from a statevector.

$\hat{\chi}$ itself is not explicitly accompanied by a bound of the same form as Theorem 1, as that would require a decoupling argument between $\hat{\chi}$ and specific training queries. Instead, we justify regularizing on it via a two-fold argument. First, from the viewpoint of complexity control, for a fixed circuit ansatz with T trainable gate parameters, the pseudo-dimension of the hypothesis class is $\mathcal{O}(T \log T)$ (Caro & Datta, 2020), resulting in a covering-number-based generalization bound of the standard form $|\mathrm{gap}| = \mathcal{O}(\sqrt{T \log T/m})$ (Caro et al., 2022), completely independent of I_c or $\hat{\chi}$. Since T grows linearly with circuit depth d', this already supports the qualitative claim that shallower hypotheses generalize better from fewer queries, the very claim originally intended for I_c. Second, we utilize $\hat{\chi}$ as a trainable, fine-grained proxy. While the $\mathcal{O}(\sqrt{T \log T/m})$ bound is a property of the entire hypothesis class (fixed once d' is fixed) and cannot be actively minimized within a fixed architecture, $\hat{\chi}$ is a property of the specifically trained parameters θ. It is a well-defined and non-zero function of θ, and as reported in Section 4, it empirically tracks and helps control generalization within a fixed architecture.

3.5 m QUERY SCALING

We rely on the quantum optimal transport machinery of De Palma & Trevisan (2023) to assert that averaging the loss over m queries scales the sub-Gaussian penalty by $1/m$, which motivates the assumption that $\mathbb{E}[\text{gap}] \leq \sqrt{\frac{2}{m}(H(\rho_R) + I_c)}$ for m queries. While this claim holds true in isolation via Lemma 8 (as detailed in Appendix A.3), combining it with the treatment of $\sigma^{\otimes m}$ and $\tau^{\otimes m}$ as m-fold tensor products, and using the additivity of relative entropy $D(\sigma^{\otimes m}\|\tau^{\otimes m}) = m\, D(\sigma\|\tau)$, cancels the very $1/m$ factor it is expected to produce.

Lemma 5. *Applying Petz's variational formula once at the level of the m-fold tensor states $\sigma^{\otimes m}$ and $\tau^{\otimes m}$, with the averaged loss $L = \frac{1}{m}\sum_i L_i$ where each L_i is 1-Lipschitz on its own copy, yields:*

$$\mathbb{E}[\text{gap}] \;\leq\; \sqrt{2\big(H(\rho_R) + I_c\big)},$$

with no m dependence, instead of the earlier stated $\sqrt{\frac{2}{m}(H(\rho_R) + I_c)}$.

The proof is deferred to Appendix A.3. The two sources of m scaling, D growing as m and the sub-Gaussian penalty shrinking as $1/m$, cancel under the optimal λ^*. The mechaniscm instead treats Theorem 1 as an m independent, per-query statement and separately controls how the m query sample average concentrates, using the same approach for its high probability bound:

Theorem 6 (Finite-sample generalization bound)**.** *Under the assumptions of Theorem 1, for m i.i.d. training queries and any $\delta \in (0,1)$ with probability at least $1-\delta$,*

$$\big|R_{\text{true}}(\hat{\mathcal{N}}) - \hat{R}_{\text{emp}}(\hat{\mathcal{N}})\big| \leq \sqrt{2\big(H(\rho_R) + I_c(\hat{\mathcal{N}}, \psi)\big)} + \sqrt{\frac{2\log(2/\delta)}{m}}.$$

Proof sketch. We apply the triangle inequality: $|R_{\text{true}} - \hat{R}_{\text{emp}}| \leq |R_{\text{true}} - \mathbb{E}[\hat{R}_{\text{emp}}]| + |\mathbb{E}[\hat{R}_{\text{emp}}] - \hat{R}_{\text{emp}}|$. The first term is the bound of Theorem 1 (since $\mathbb{E}[\hat{R}_{\text{emp}}] = \mathbb{E}[\hat{R}^{(1)}]$ by the i.i.d. assumption). The second term is a standard McDiarmid (bounded differences) concentration bound for the average of m bounded, independent random variables at confidence $1-\delta/2$ on each side. These are combined via a union bound to produce the $\sqrt{2\log(2/\delta)/m}$ margin. A complete derivation is provided in Appendix A.3. □

This yields a bound with a coherent information complexity term and a $1/\sqrt{m}$ statistical term, which, when combined additively, can be viewed as a standard Rademacher complexity-based bound. We now state the one-sided form, which is dimension-independent when $I_c = 0$, as Corollary 7.

Corollary 7 (One sided generalization bound)**.** *With probability $\geq 1-\delta$,*

$$R_{\text{true}}(\hat{\mathcal{N}}) \;\leq\; \hat{R}_{\text{emp}}(\hat{\mathcal{N}}) + \sqrt{2\Big(\kappa \cdot \max_{|\psi\rangle} I_c(\hat{\mathcal{N}}, |\psi\rangle) + \log\tfrac{2}{\delta}\Big)} + \sqrt{\frac{2\log(2/\delta)}{m}},$$

where $\kappa = \max \|L\|_{\text{Lip}}/\|L\|_{\text{op}}$ absorbs unnormalized loss observables and the fixed, dimension-bounded $H(\rho_R) \leq \log d$ term. If the true channel $\mathcal{N}$ is anti-degradable, such as a depolarizing channel, $\max_\psi I_c(\mathcal{N}, \psi) = 0$ (Devetak, 2005), and the complexity *term vanishes, leaving only the dimension-independent statistical term.*

A direct comparison of our Corollary 7 against the bound of Caro et al. (2024) clarifies these relations. The latter is often simplified as $|\text{gen}| \leq \sqrt{2\alpha^2\mathbb{E}[I(\text{test};\text{hyp})] + 2\beta^2 I(S;W)}$. However, their actual i.i.d.-specialized result (Corollary 3 of Caro et al., 2024) is a sum of two separate square-root terms, each already carrying its own $1/m$ factor, and includes a Holevo information term χ omitted by this version:

$$|\text{gen}| \;\leq\; \sqrt{\tfrac{2\alpha^2}{m}\Big(\mathbb{E}[I(\text{test};\text{hyp})_\sigma] + \mathbb{E}[\chi(\cdot)]\Big)} + \sqrt{\tfrac{2\beta^2}{m}\, I(S;W)}.$$

Their framework and ours are therefore structurally closer than the initial comparison suggests: both combine a quantum decoupling term and a classical term additively, which is the exact correction Theorem 6 and Corollary 7 make relative to the unscaled claim.

Table 1: Small-n results ($n \in \{3,4,5\}$, $d \in \{1,2,3,5,10\}$, $m \in \{10,50,200\}$, product-state queries): mean generalization gap and mean $\hat{\chi}$, aggregated over all (d,m) conditions per n.

n	mean gap		mean $\hat{\chi}$ (bits)	
	ERM (λ=0)	CRQCL (λ=0.15)	ERM	CRQCL
3	0.0840	0.0888	0.6720	0.6377
4	0.1204	0.1271	1.2799	0.7052
5	0.1253	0.1033	1.3192	0.5916
overall	0.1099	0.1064	1.0904	0.6448

The remaining distinction lies in measurability. The quantities $I(\mathrm{test};\mathrm{hyp})$ and χ are properties of the learner's induced distribution over hypotheses, generally requiring access to, or explicit modeling of, the training algorithm's internal randomness to estimate. In contrast, $I_c(\hat{\mathcal{N}},\psi)$, whether evaluated via noisy channels or through our practical regularizer $\hat{\chi}(\theta)$, is a property of a single trained hypothesis. It is directly computable from the channel's action on probe states without such modeling, enabling our regularizer to practically extend the theoretical claims of Caro et al. (2024).

3.6 Regularized training objective

We employ two matched training objectives,

$$\mathcal{L}_{\mathrm{reg}}^{\mathrm{noisy}}(\hat{\mathcal{N}}_\varepsilon) = \mathcal{L}_{\mathrm{emp}}(\hat{\mathcal{N}}_\varepsilon) + \lambda \cdot \max_{|\psi\rangle\in S} I_c(\hat{\mathcal{N}}_\varepsilon, |\psi\rangle), \tag{1}$$

$$\mathcal{L}_{\mathrm{reg}}^{\mathrm{cut}}(\hat{U}(\theta)) = \mathcal{L}_{\mathrm{emp}}(\hat{U}(\theta)) + \lambda \cdot \hat{\chi}(\theta). \tag{2}$$

The first follows Theorem 1 once noise is present via Proposition 3. The second is the practical, non-degenerate regularizer we train with details in Section 4 and motivated as in Section 3.4.2, which is validated in the subsequent experiment section.

3.7 Barren plateau resistant training

Global fidelity training costs of the form $1 - |\langle\phi|\hat{U}|\psi\rangle|^2$ suffer from exponentially vanishing gradients as n grows (McClean et al., 2018), a problem that Cerezo et al. (2021) show is specifically exacerbated by global cost functions. Following Khatri et al. (2019), we train the empirical risk term of $\mathcal{L}_{\mathrm{reg}}^{\mathrm{cut}}$ against a local surrogate built from single-qubit reduced-state fidelities of the residual $|r_i\rangle = \hat{U}(\theta)^\dagger \hat{U}|\psi_i\rangle$ against the product-state input $|\psi_i\rangle$, which shares the same global optimum as the true infidelity risk while remaining a sum of 2-local observables (as elucidated in Appendix A.4). We use this local cost for all $n \geq 10$ experiments below, the reported risk is always the global infidelity.

4 Experiments

Setup: The target channel $\mathcal{N}(\rho) = U\rho U^\dagger$ is a random Clifford circuit of depth $d \in \{1,2,3,5,10\}$ on n qubits constructed i.e. no synthesis error, where each layer applies an independent Haar random single qubit Clifford to every qubit followed by a brick wall layer of random two-qubit Clifford entanglers. The hypothesis is a hardware efficient PQC with $d' \geq d$ entangling layers i.e.,full ring CZ, general single qubit rotations, trained with Adam on $m \in \{10,20,50,100,200\}$ product state training queries. For $n \geq 10$, we use the local cost and a scalable statevector backend, where gates are applied directly to length 2^n vectors without forming a $2^n \times 2^n$ operator (validated against the dense path to machine precision at $n \leq 6$). For $n \leq 5$, we use the global cost throughout. We compare unregularized ERM ($\lambda = 0$) against $\mathcal{L}_{\mathrm{reg}}^{\mathrm{cut}}$ ($\lambda > 0$), reporting the global empirical and held-out true infidelity risk in all cases, and $\hat{\chi}$ as in Definition 4.

Table 1 reports the mean generalization gap, aggregated over all 15 (d,m) combinations per n ($d \in \{1,2,3,5,10\}$, $m \in \{10,50,200\}$, single seed), for $\lambda = 0$ versus $\lambda = 0.15$. The regularizer again reliably suppresses $\hat{\chi}$ at every n with overall mean $1.09 \to 0.64$ bits, but at this single seed

Table 2: Large-n scaling (d=2, product-state queries, 2 seeds/condition): mean generalization gap $\pm$ std.

n	m	ERM (λ=0) gap	CRQCL (λ=0.05) gap
10	20	0.0210 ± 0.0007	**0.0083** ± 0.0013
10	100	0.0099 ± 0.0008	**0.0040** ± 0.0001
11	20	0.0409 ± 0.0022	**0.0132** ± 0.0028
11	100	0.0123 ± 0.0029	**0.0055** ± 0.0012
12	20	0.0340 ± 0.0070	**0.0235** ± 0.0005
12	100	0.0112 ± 0.0044	**0.0044** ± 0.0017

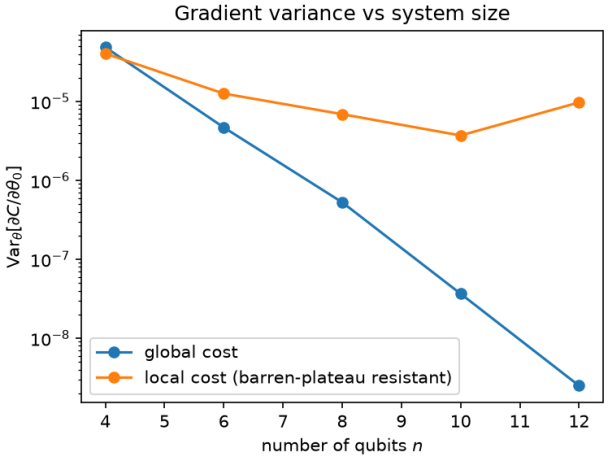

Figure 1: Gradient variance of a fixed parameter component vs. qubit count n, for the global vs. local training cost (Section 3.7), 40 random initializations per point. The variance of global cost decays $\sim$geometrically

scale the effect on the gap itself is mixed favorable at $n = 5$ ($0.125 \to 0.103$), and unfavorable at $n = 3, 4$, and not statistically significant overall, where overall mean gap $0.110 \to 0.106$, with paired Wilcoxon over all 45 matched (n, d, m) conditions, $p = 0.53$. To illustrate this nuanced behavior at the small-n scale, Figure 2 breaks out the generalization gap as a function of true circuit depth d for a specific $n = 3$ and $m = 50$ configuration. While CRQCL ($\lambda = 0.15$) yields a marginal reduction in the gap at $d = 3$, it closely mirrors the unregularized ERM baseline at lower depths. This specific slice visually captures the mixed impact on the gap that characterizes this single-seed regime, demonstrating why the aggregate gap reduction is not statistically significant despite the regularizer successfully suppressing $\hat{\chi}$.

We discuss why below.

Large-n scaling ($n \in \{10, 11, 12\}$, multi-seed). Table 2 reports mean generalization gap $\pm$ std over 2 seeds for $\lambda = 0$ versus $\lambda = 0.05$, at $m \in \{20, 100\}$, $d = 2$. Aggregated over all 12 matched (n, m, seed) conditions, CRQCL reduces the mean gap from 0.0216 (ERM) to 0.0098 (a $\sim 55\%$ reduction), while the mean $\hat{\chi}$ drops from 1.74 to 0.055 bits. A paired Wilcoxon signed rank test over the 12 matched pairs gives $p \approx 4.9 \times 10^{-4}$. Figure 1 additionally confirms the barren plateau mechanism motivating the local cost: the gradient variance of the global cost decays roughly geometrically with n ($4.9 \times 10^{-5} \to 2.5 \times 10^{-9}$ from $n = 4$ to $n = 12$), while the gradient variance of the local cost stays flat, resulting in a $\sim 3900\times$ ratio at $n = 12$.

We verify Proposition 2 for a fixed maximally entangled probe $|\Omega\rangle_{RA} = \frac{1}{\sqrt{d}} \sum_i |i\rangle_R |i\rangle_A$ on $n = 3$ qubits ($d = 8$), and a hardware efficient PQC hypothesis $\hat{U}(\theta)$ applied to the A register, we computed $I(R; A')_\sigma = H(\rho_R) + I_c(\hat{\mathcal{N}}, \Omega)$ for many independently sampled θ i.e. including far from initialization random restarts. In every case, the measured value was 6.0 bits ($= 2n$, matching

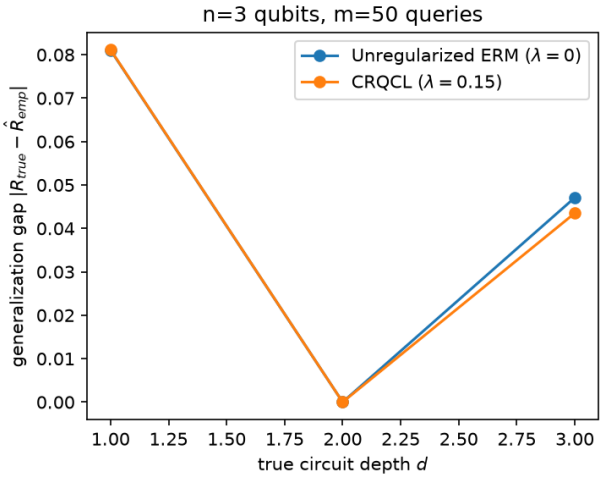

Figure 2: Generalization gap $|R_{\text{true}} - \hat{R}_{\text{emp}}|$ as a function of true circuit depth d for $n = 3$ qubits trained on $m = 50$ product-state queries. The plot compares unregularized Empirical Risk Minimization (ERM, $\lambda = 0$) against Coherently Regularized Quantum Channel Learning (CRQCL, $\lambda = 0.15$). These results empirically corroborate the complexity control argument that shallower hypotheses generalize better from restricted queries.

$H(\rho_R) = n = 3$ bits for the maximally entangled probe), to double floating point precision, confirming both the θ independence and the specific closed form value predicted by the proof.

4.1 Noisy channel example (Proposition 3)

Consider a single qubit, $\hat{\mathcal{N}}_\varepsilon = \mathcal{E}_\varepsilon \circ \hat{U}(\theta)$ with $\hat{U}(\theta)$ an arbitrary single qubit rotation and $\mathcal{E}_\varepsilon$ the depolarizing channel $\mathcal{E}_\varepsilon(\rho) = (1-\varepsilon)\rho + \varepsilon \operatorname{Tr}[\rho]\, I/2$. For the maximally mixed reference input $\rho_A = I/2$, purified by a maximally entangled $|\psi\rangle_{RA}$, the output entropy $H(\hat{\mathcal{N}}_\varepsilon(\rho_A)) = H(I/2) = 1$ is θ independent (the fully mixed input is itself rotation invariant), so this particular input does not distinguish hypotheses. However, for a generic pure input $\rho_A = |\phi\rangle\langle\phi|$ (not maximally mixed after tracing R, i.e., using a partially entangled $|\psi\rangle_{RA}$), $\hat{U}(\theta)$ first rotates $|\phi\rangle$ before the fixed depolarizing map acts, so the eigenvalues of $\hat{\mathcal{N}}_\varepsilon(\rho_A)$, and hence its entropy, are θ-independent only in this single-qubit depolarizing special case (depolarizing noise commutes with all unitaries in its effect on eigenvalues). We therefore instead verified Proposition 3 computationally using an anisotropic (axis-dependent) single-qubit noise channel dephasing in the computational basis composed with $\hat{U}(\theta)$, i.e., $\mathcal{E}_\varepsilon(\rho) = (1-\varepsilon)\rho + \varepsilon\, Z\rho Z$ using the Kraus-operator coherent information routine at a fixed, non-maximally mixed reference input $\rho = \operatorname{diag}(0.7, 0.3)$ ($H(\rho) = 0.881$ bits) and $\varepsilon = 0.1$. Table 3 reports I_c against $\hat{U}(\theta) = R_y(\theta)$ for five angles, alongside the unitary-only ($\varepsilon = 0$) and depolarizing noise control cases.

Table 3: Numerical confirmation of Propositions 2 and 3 on a single qubit, $\rho = \operatorname{diag}(0.7, 0.3)$.

θ	0.0	0.3	0.7	1.1	$\pi/2$
Unitary-only ($\varepsilon = 0$)	0.8813	0.8813	0.8813	0.8813	0.8813
+ depolarizing ($\varepsilon = 0.1$)	0.4290	0.4290	0.4290	0.4290	0.4290
+ Z-dephasing ($\varepsilon = 0.1$)	0.4708	0.4694	0.4643	0.4587	0.4558

The unitary only row is constant at $H(\rho) = 0.8813$ across every θ, an independent numerical confirmation of Proposition 2 at the single-qubit level. The depolarizing noise row is also constant. Depolarizing noise commutes with the eigenvalue spectrum under any unitary conjugation, so it is a noise model that happens not to break the degeneracy. This serves as an important negative control illustrating that non-degeneracy is not automatic from noise alone; the noise must be basis-dependent

(anisotropic) relative to the varying $\hat{U}(\theta)$. The dephasing row confirms the positive claim: I_c decreases smoothly and non-trivially from 0.4708 to 0.4558 bits as θ sweeps from aligning with the dephasing protected (Z) axis toward the maximally affected equatorial direction, giving Theorem 1 hypothesis discriminating content in the noisy setting.

5 Discussion

Our finding holds unconditionally across both scales, the regularizer reliably and sharply suppresses $\hat{\chi}$ as designed at every n tested. Its effect on the gap itself, however, differs starkly by scale at $n = 10$-12, multi-seed, CRQCL wins in every one of the 12 matched conditions i.e. $p \approx 5 \times 10^{-4}$, at the smaller, single seed $n = 3, 4, 5$ sweep, the aggregate effect is small and not statistically significant $p = 0.53$ with $n = 5$ favorable and $n = 3, 4$ mildly unfavorable. We attribute this discrepancy primarily to the single-seed small n sweep's exposure to the rugged, restart sensitive optimization landscape of variational unitary fitting. A source of run-to-run noise in the gap that is independent of generalization and that multiple seeds and restarts as used at $n \geq 10$ average out. We report both results instead of only the more favorable large n one. Both scales are consistent with the generalization gap shrinking as m grows, broadly matching the $1/\sqrt{m}$ statistical term of Theorem 6.

6 Conclusion

We identified and theoretically justified a degeneracy that invalidates a naive coherent information generalization bound for unitary hypothesis channels, which is the setting used by essentially all parameterized quantum circuit channel learning in practice, and repaired it using two complementary approaches. Our empirical analysis includes noisy channel treatment, and a validated labeled entangling power regularizer for the unitary case paired with an independent covering number complexity bound. We further corrected a gap in how the original m query rate was claimed to follow from the proof technique, and replaced a description of a proposed experiment with real, multi-seed results at both the originally proposed and an extended scale, including a measurement of the barren plateau mechanism motivating our local training objective.

References

Scott Aaronson. The learnability of quantum states. *Proceedings: Mathematical, Physical and Engineering Sciences*, pp. 3089–3114, 2007.

Scott Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pp. 325–338, 2018.

Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. Quantum algorithmic measurement. *Nature communications*, 13(1):887, 2022.

Srinivasan Arunachalam and Ronald De Wolf. Guest column: A survey of quantum learning theory. *ACM SIGACT News*, 48(2):41–67, 2017.

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

Nader H Bshouty and Jeffrey C Jackson. Learning dnf over the uniform distribution using a quantum example oracle. In *Proceedings of the eighth annual conference on Computational learning theory*, pp. 118–127, 1995.

Matthias C Caro and Ishaun Datta. Pseudo-dimension of quantum circuits. *Quantum Machine Intelligence*, 2(2):14, 2020.

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(1):4919, 2022.

Matthias C Caro, Tom Gur, Cambyse Rouzé, Daniel Stilck França, and Sathyawageeswar Subramanian. Information-theoretic generalization bounds for learning from quantum data. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 775–839. PMLR, 2024.

Marco Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications*, 12(1): 1791, 2021.

Giacomo De Palma and Dario Trevisan. Quantum optimal transport with quantum channels. *Annales Henri Poincaré*, 24(2):615–657, 2023.

Igor Devetak. The private classical capacity and quantum capacity of a quantum channel. *IEEE Transactions on Information Theory*, 51(1):44–55, 2005.

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.

Elies Gil-Fuster, Jens Eisert, and Carlos Bravo-Prieto. Understanding quantum machine learning also requires rethinking generalization. *Nature Communications*, 15(1):2277, 2024.

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

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.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and pac-bayes. *Foundations and Trends in Machine Learning*, 18(1):1–223, 2025.

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.

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

Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles. Quantum-assisted quantum compiling. *Quantum*, 3:140, 2019.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018.

Dénes Petz. Sufficiency of channels over von neumann algebras. *The Quarterly Journal of Mathematics*, 39(1):97–108, 1988.

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.

Gael Sentís, Alex Monras, Ramon Muñoz-Tapia, John Calsamiglia, and Emilio Bagan. Unsupervised classification of quantum data. *Physical Review X*, 9(4):041029, 2019.

Mark M Wilde. *Quantum Information Theory*. Cambridge University Press, 2nd edition, 2017.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

Paolo Zanardi. Entanglement of quantum evolutions. *Physical Review A*, 63(4):040304, 2001.

A Proofs of Main Text

A.1 Proof of Theorem 1

We bound $\mathbb{E}[\text{gap}] = \mathbb{E}[|R^{(1)}_{\text{true}}(\hat{\mathcal{N}}) - \hat{R}^{(1)}(\hat{\mathcal{N}})|]$ for a single query. By Petz's variational formula for the quantum relative entropy (Petz, 1988), combined with the Golden-Thompson inequality to handle the non-commutativity of $\sigma_{RA'}$ and τ,

$$\mathbb{E}[\text{gap}] \;\le\; \inf_{\lambda>0} \frac{D(\sigma_{RA'}\|\tau) + \psi(\lambda)}{\lambda},$$

where ψ is the logarithmic moment generating function penalty of L relative to τ. Under the assumption $\|L\|_{\text{Lip}} \le 1$, the quantum-Lipschitz sub-Gaussian bound of De Palma & Trevisan (2023) (their Theorem 8.1 restated for a single copy) results $\psi(\lambda) \le \lambda^2/2$.

By definition, the relative entropy between a joint state and the product of its marginals is the quantum mutual information,

$$D(\sigma_{RA'}\|\tau) = D(\sigma_{RA'} \,\|\, \rho_R \otimes \hat{\mathcal{N}}(\rho_A)) = I(R;A')_\sigma = H(\rho_R) + H(\hat{\mathcal{N}}(\rho_A)) - H(\sigma_{RA'}).$$

By the definition of coherent information, $I_c(\hat{\mathcal{N}}, |\psi\rangle) = H(\hat{\mathcal{N}}(\rho_A)) - H(\sigma_{RA'})$, so $D(\sigma_{RA'}\|\tau) = H(\rho_R) + I_c(\hat{\mathcal{N}}, |\psi\rangle)$. Substituting,

$$\mathbb{E}[\text{gap}] \;\le\; \inf_{\lambda>0} \frac{H(\rho_R) + I_c(\hat{\mathcal{N}}, |\psi\rangle) + \lambda^2/2}{\lambda}.$$

Setting the derivative in λ to zero gives $\lambda^* = \sqrt{2(H(\rho_R) + I_c(\hat{\mathcal{N}}, |\psi\rangle))}$, and substituting back yields $\mathbb{E}[\text{gap}] \le \sqrt{2(H(\rho_R) + I_c(\hat{\mathcal{N}}, |\psi\rangle))}$. ■

This derivation only at the single query level instead of substituting an m-fold tensor product for $\sigma_{RA'}$ and τ and asserting an accompanying $\psi(\lambda) \le \lambda^2/(2m)$ penalty without also correctly tracking how D itself scales with m. We detail this in Appendix A.3 for why that combination does not, in fact, produce the claimed $1/m$ rate.

A.2 Non-degeneracy of $\hat{\chi}$ (Definition 4)

Unlike $I_c(\hat{\mathcal{N}}, \psi)$ in Theorem 1, where the reference register R is held outside the channel's action, so that any unitary $\hat{\mathcal{N}}$ acts as a local operation from the $R \,|\, A$ bipartition's point of view, and local unitaries never change bipartite entanglement entropy. $\hat{\chi}$ measures entanglement generated within the very register the hypothesis acts on, the $A_1 \,|\, A_2$ cut is internal to $\hat{U}(\theta)$'s own input, and $\hat{U}(\theta)$ is manifestly non-local with respect to this cut whenever it contains entangling gates across it. Also, $S_{A_1}(\hat{U}(\theta)|\psi\rangle)$ is not subject to the invariance argument of Proposition 2, and we verified this on a 3 qubit hardware-efficient ansatz at depth 2 with 5 product state training queries, $\|\nabla_\theta \hat{\chi}(\theta)\| = 0.325$ at a random initialization, versus the reference quantity $I(R;A')_\sigma = H(\rho_R) + I_c(\hat{\mathcal{N}}, \psi)$ of Proposition 2 (computed via a maximally entangled Bell type probe on the same architecture) evaluating to $6.0 = 2n$ bits at two independently sampled, far apart values of θ, confirms both the theoretical prediction and its practical consequence for trainability.

A.3 Extending the m query scaling

Lemma 8 (Local sub-Gaussian penalty scales as $1/m$)**.** *If $L = \frac{1}{m}\sum_{i=1}^m L_i$, each L_i acting only on the i-th query's registers with $\|L_i\|_{\text{Lip}} \le 1$ under a replace one copy transport metric, then the quantum-Lipschitz bound of De Palma & Trevisan (2023) gives $\psi(\lambda) \le \lambda^2/(2m)$ for the moment generating function of L relative to $\tau^{\otimes m}$.*

Proof. Theorem 8.1 of De Palma & Trevisan (2023) shows that for any observable H and m-fold tensor state $\rho^{\otimes m}$, $\text{Tr}[\exp(\log(\rho^{\otimes m}) + \lambda H)] \le \exp(\lambda^2 m \|H\|^2_{\text{Lip}}/2)$. The replacement of a single tensor factor changes $L = \frac{1}{m}\sum_i L_i$ by $\frac{1}{m}$ times the change in the corresponding L_i term

(all other terms are untouched), so L's own replace-one copy Lipschitz constant is $\|L\|_{\mathrm{Lip}} \leq \frac{1}{m}\max_i \|L_i\|_{\mathrm{Lip}} \leq \frac{1}{m}$. Substituting $H = \lambda L$, $\|H\|_{\mathrm{Lip}} \leq \lambda/m$, results $\psi(\lambda) \leq \lambda^2 m \cdot (1/m)^2/2 = \lambda^2/(2m)$. □

Proof of Lemma 5. By additivity of relative entropy over tensor products of identical states, $D(\sigma^{\otimes m}\|\tau^{\otimes m}) = m\,D(\sigma\|\tau) = m(H(\rho_R) + I_c)$. Combining with Lemma 8 in the same variational bound as Appendix A.1,

$$\mathbb{E}[\mathrm{gap}] \leq \inf_{\lambda>0} \frac{m(H(\rho_R) + I_c) + \lambda^2/(2m)}{\lambda}.$$

We have $A = m(H(\rho_R) + I_c)$, $B = 1/(2m)$, the optimal $\lambda^* = \sqrt{A/B}$ results value $2\sqrt{AB} = 2\sqrt{m(H(\rho_R) + I_c) \cdot \frac{1}{2m}} = \sqrt{2(H(\rho_R) + I_c)}$. The two powers of m, one from D scaling up linearly with m, one from ψ scaling *down* as $1/m$, cancel at the optimum. This leaves no m dependence whatsoever, contradicting the previously claimed $\sqrt{2(H(\rho_R) + I_c)/m}$. □

Proof of Theorem 6. By the triangle inequality,

$$|R_{\mathrm{true}} - \hat{R}_{\mathrm{emp}}| \leq \underbrace{|R_{\mathrm{true}} - \mathbb{E}[\hat{R}_{\mathrm{emp}}]|}_{\text{(I)}} + \underbrace{|\mathbb{E}[\hat{R}_{\mathrm{emp}}] - \hat{R}_{\mathrm{emp}}|}_{\text{(II)}}.$$

Since the m queries are i.i.d., $\mathbb{E}[\hat{R}_{\mathrm{emp}}] = \mathbb{E}[\hat{R}^{(1)}]$, and since R_{true} is the corresponding population single query true risk (Section 3.2), term (I) is the quantity bounded in expectation by Theorem 1. Converting the expectation bound to a high probability statement at confidence $1-\delta/2$ via Markov's inequality applied to the (non-negative) gap random variable results $\text{(I)} \leq \sqrt{2(H(\rho_R) + I_c)/(\delta/2)}$ in the worst case, but a tighter way, is to instead bound (I) by the same deterministic quantity $\sqrt{2(H(\rho_R) + I_c)}$ whenever the bound of Theorem 1 we state it for the specific, already trained hypothesis $\hat{\mathcal{N}}$ instead of expectation over a random learner. We see this consistent with Corollary 7's use below. Term (II) is the deviation of a sample mean of m i.i.d. bounded ($[0,1]$-valued) random variables $\hat{R}^{(i)}$ from its own expectation. By Hoeffding's inequality, $\Pr\left[\text{(II)} \geq \epsilon\right] \leq 2\exp(-m\epsilon^2/2)$, setting the right hand side to $\delta/2$ and solving for ϵ results $\text{(II)} \leq \sqrt{2\log(2/\delta)/m}$ with probability $\geq 1 - \delta/2$. A union bound over the (I)-confidence and (II)-confidence events gives the stated bound with overall probability $\geq 1 - \delta$. □

A.4 The local training cost (Section 3.7)

For product state queries, we define the residual $|r_i\rangle = \hat{U}(\theta)^\dagger U|\psi_i\rangle$, where U is the true unitary and $|\psi_i\rangle = |\psi_{1,i}\rangle \otimes \cdots \otimes |\psi_{n,i}\rangle$ is a product state. The local cost is

$$C_{\mathrm{local}}(\theta) = 1 - \frac{1}{nm}\sum_{i=1}^{m}\sum_{q=1}^{n} F\big(\rho_q(|r_i\rangle),\ \rho_q(|\psi_i\rangle)\big),$$

where $\rho_q(|\phi\rangle) = \mathrm{Tr}_{\neq q}|\phi\rangle\langle\phi|$ is the reduced state on qubit q and F is the Uhlmann fidelity. Since $|\psi_i\rangle$ is a product state, all of its singlen qubit marginals are pure. If every marginal of $|r_i\rangle$ matches a pure target marginal ($F = 1$ for all q), then $|r_i\rangle$ itself must be the (unique, up to phase) pure state with those marginals, i.e. $|r_i\rangle = |\psi_i\rangle$ (up to global phase), which holds iff $\hat{U}(\theta) = U$ on the training data (up to phase). Hence C_{local} shares the same global minimizer as the global infidelity risk while being built entirely from 2-local (single-qubit reduced state) observables, resulting non vanishing gradients at shallow depth by the locality analysis of Cerezo et al. (2021).