

Spectral, information-theoretic, and perturbative methods for quantum learning and error mitigation

by

Evan Peters

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Physics (Quantum Information)

Waterloo, Ontario, Canada, 2024

© Evan Peters 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Jens Eisert
Professor, Dahlem Center for Complex Quantum Systems,
Freie Universität Berlin

Supervisor: Achim Kempf
Professor, Dept. Applied Mathematics, University of Waterloo

Internal Member: Roger Melko
Professor, Dept. of Physics & Astronomy
University of Waterloo

Internal Member: Norbert Lütkenhaus
Professor, Dept. of Physics & Astronomy
University of Waterloo

Internal-External Member: Graeme Smith
Associate Professor, Dept. of Applied Mathematics,
University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis includes material from five articles or preprints that I authored or co-authored (Refs. [1–5]). Therefore, the work presented in this thesis comes from several collaborations that took place during the course of my PhD studies. My specific contributions to each project are detailed below:

Chapter 2: Quantum kernel methods and spectra

- This chapter introduces quantum kernel learning and spectra of integral kernel operators, and paraphrases some material from Ref. [1] coauthored with collaborators Abdulkadir Canatar, Cengiz Pehlevan, Stefan M. Wild, and Ruslan Shaydulin. This chapter mostly contains well-known background material or my specific contributions to the project, with the exception of Secs. 2.4.1.2 and 2.4.2.1 which were done mainly by Ruslan and Sec. 2.4.2.2 done mainly by Abdulkadir and Ruslan.

Chapter 3 (and Appendix A): Generalization despite overfitting in quantum machine learning models

- This chapter introduces the notion of “benign overfitting” for quantum machine learning models, and is essentially Ref. [2] that I coauthored with Maria Schuld. Maria conceived of the high-level idea for this project, provided close guidance throughout the project, and drafted the abstract and introduction for the work. I wrote the body of the manuscript, and performed the calculations and numerics.

Chapter 4: Improving information transmission using correlated auxiliary noise

- This chapter introduces and analyzes a model for “correlated auxiliary noise” in classical and quantum channels, and contains unpublished work in quantum information theory. Several of the high-level ideas in this chapter developed out of discussions with Achim Kempf and Koji Yamaguchi. I wrote this chapter and derived its mathematical contents independently.

Chapter 5: Bounds and Guarantees for learning and entanglement

- This chapter reproduces the single-author preprint Ref. [3]; I performed this work independently.

Chapter 6 (and Appendix B): Perturbative readout error mitigation for near-term quantum computers

- This chapter reproduces Ref. [4] studying a technique for recovering the probability of measuring the all-zeros bitstring in the presence of readout error. I originally analyzed this technique, and then Andy C. Y. Li proposed and formalized the generalization of this technique appearing in Sec. 6.5 and App. B.3, which we wrote up together. Andy also performed the numerics and IBM experiments for this generalized technique appearing in Sec. 6.6 and App. B.5.1. Gabe Perdue and Andy C. Y. Li both contributed to drafting the body of the manuscript.

Chapter 7 (and Appendix C): Qubit assignment using time reversal

- This chapter reproduces Ref. [5] developing a technique for assigning logical qubits to physical qubits. Prasanth Shyamsundar originally proposed the idea of using simulated annealing for this task, which I later formalized for the manuscript. Andy C. Y. Li developed a generalized analysis for Markovian noise models that I have not included here, and I refer the interested reader to Ref. [5]. I was responsible for developing the theory, running and analyzing all experiments, and writing the bulk of the manuscript that appears here. Prasanth Shyamsundar, Andy CY Li, Gabriel Perdue contributed to drafting the body of the manuscript.

I also contributed to the following publications [6–8] during the course of my PhD, but have excluded them from this dissertation:

Evan Peters et al. “Machine learning of high dimensional data on a noisy quantum processor”. In: *npj Quantum Information* 7.1 (2021), pp. 1–5

Erik Gustafson et al. “Large scale multi-node simulations of $\mathbb{Z}_2$ gauge theory quantum circuits using Google Cloud Platform”. In: *2021 IEEE/ACM Second International Workshop on Quantum Computing Software (QCS)*. 2021, pp. 72–79

Sergei V. Isakov et al. *Simulations of Quantum Circuits with Approximate Noise using qsim and Cirq*. arXiv:2111.02396. 2021. arXiv: [2111.02396](https://arxiv.org/abs/2111.02396) [quant-ph]

Abstract

We present spectral and information-theoretic characterizations of learning tasks involving quantum systems, and develop new perturbative error mitigation techniques for near-term devices. In the first part of this thesis, we explore connections between quantum information and learning theory. We demonstrate theoretically that kernel bandwidth enables quantum kernel methods associated with a high dimensional quantum feature space to generalize. We then characterize quantum machine learning models that generalize despite overfitting their training data, contradicting standard expectations from learning theory. In such learning tasks, the learner may fail due to noise in the input data. So we next consider a setting where the learner has access to correlated auxiliary noise, a resource that contains information about an otherwise unknown noise source corrupting input data. We use classical Shannon theory to relate the strength of these correlations to the classical capacity of a bit flip channel with correlated auxiliary noise, and we extend this analysis to derive the quantum capacity of a quantum bit flip channel given access to an auxiliary system entangled with the environmental source of the noise. Finally, we derive an information-theoretic guarantee for the learnability of data by an optimal learner and, extending this technique to a quantum setting, we introduce and characterize an entanglement manipulation task that generalizes the notion of classical learning.

The second part of this thesis introduces techniques for error mitigation on near-term quantum processors and provides guarantees in the perturbative limit. We introduce a technique for mitigating measurement errors using truncated matrix operations. We then propose and characterize a technique that uses the time-reversibility of a quantum circuit to measure the quality of a subset of qubits, and we apply this technique to assign logical circuits to qubits on a physical device in a nearly optimal manner using a simulated annealing optimization algorithm.

Acknowledgements

Foremost I want to thank Achim Kempf for being so supportive while also letting me pursue whatever topics interested me. Without this degree of freedom I would have never had the chance to get involved in so many interesting projects. I also want to thank Gabe Perdue and Maria Schuld, who supervised me for internships during the course of my PhD and lent valuable, alternative perspectives to my research. Finally, I thank Heidi Schellman for being there to provide professional support when things got a bit rough in my master's.

I thank Nathan Killoran, Angus Lowe, Joseph Bowles, Arsalan Motamedi, Kohdai Kuroiwa, Ernest Tan for their feedback on manuscripts that appear here as chapters, T. Rick Perche and Jose Polo Gomez for helping me with functional analysis, and Alan Ho, Ryan LaRose, Xiao Mi, and Matthew Harrigan for assisting us in using Google's cloud quantum computing services.

I want to acknowledge support and funding from the various institutions and collaborations that I have been a part of in these past four years. My research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development and by the Province of Ontario through the Ministry of Colleges and Universities. I thank Mitacs for their support through the Mitacs Accelerate program, Google Quantum AI for access to their quantum computing hardware and the support they provided for experiments, and Google for supporting me through Achim Kempf's Google Faculty Award. My work at Fermilab was partially supported by several DOE grants: the DOE/HEP QuantISED program grants "HEP Machine Learning and Optimization Go Quantum," (identification number 0000240323), "QCCFP-QMLQCF Consortium," (identification number DE-SC0019219), and Fermilab LDRD project FNAL-LDRD-2018-025 "Towards a Quantum Computing Science Center at Fermilab."

I appreciate the all the friends from the extended lab group that helped me get through the pandemic - Adam, Boris, Bruno, Jose, Kelly, Mattheus, Rick, Zac. And I am quite grateful for all the other friends I've made along the way - Anand, Bossi, Chris, Emilia, Jeremy, Kimia, Nachi, Nina, Noura, Sasha and many others I've surely missed here - and for my family: Amalie, David, Mom, Dad.

Table of Contents

Examining Committee	ii
Author's Declaration	iii
Statement of Contributions	iv
Abstract	vi
Acknowledgements	vii
List of Figures	xiii
List of Tables	xv
1 Introduction	1
2 Quantum theory and kernel methods	5
2.1 Chapter overview	5
2.2 Finite-dimensional quantum theory	5
2.3 Learning with kernels	7
2.3.1 Kernel spectra	10
2.3.2 Quantum feature space is not always quantum state space	12
2.3.3 Kernel spectra and generalization	15

2.4	Quantum kernel bandwidth and generalization	17
2.4.1	Limitations of fixed quantum embeddings	18
2.4.2	Bandwidth enables learning	21
3	Generalization despite overfitting in quantum machine learning models	27
3.1	Chapter overview	27
3.2	Introduction	28
3.3	Interpolating models in the Fourier framework	29
3.3.1	Setting up the learning problem	29
3.3.2	Two extreme cases to understand generalization	33
3.3.3	Generalization trade-offs and benign overfitting	35
3.4	Benign overfitting in single-layer quantum models	39
3.4.1	Simplified quantum model	41
3.4.2	Trainable unitaries reweight the general quantum model	42
3.4.3	Weight behavior in deep quantum models	47
3.5	Discussion	48
4	Improving information transmission using correlated auxiliary noise	50
4.1	Chapter overview	50
4.2	Introduction	51
4.2.1	Prior work	52
4.3	Classical Shannon theory	52
4.4	Correlated auxiliary noise for a bit flip channel	54
4.4.1	The capacity of the correlated-noise bit flip channel	55
4.4.2	Guarantees for hypothesis testing using correlated auxiliary noise	58
4.4.3	Accuracy of the optimal decoder	60
4.4.4	Security of one-time pads from correlated auxiliary noise	62
4.5	Hypothesis testing bounds on QEC decoders	62

4.5.1	Example: 3-qubit code	62
4.6	Correlated auxiliary noise for improving quantum channel capacity	65
4.6.1	The capacity of quantum bit flip channel with correlated auxiliary noise	68
4.6.2	Towards information-theoretic bounds on the helpfulness of UCAN	71
5	Bounds and Guarantees for learning and entanglement	75
5.1	Chapter overview	75
5.2	Introduction	76
5.3	Bounds and guarantees for classical learning	77
5.3.1	Classical learning lower bound via Fano method	78
5.3.2	Classical learning upper bound	81
5.4	Maximal singlet fraction	84
5.4.1	Reduction from singlet fraction to classical hypothesis testing	87
5.5	A quantum generalization of learning	88
5.5.1	An example: Beyond quantum exact learning	89
5.5.2	The entanglement fraction task	91
5.5.3	The relationship to classical learning	93
5.6	Bounds for the entanglement fraction task	95
5.7	Discussion	106
6	Perturbative readout error mitigation for near-term quantum computers	108
6.1	Chapter overview	108
6.2	Introduction	109
6.3	Inverting readout error	110
6.4	Recovering the probability of all-zeros bitstring	111
6.4.1	Exact bounds for a relaxation-only model	113
6.5	Perturbative mitigation for recovering full distributions	115
6.6	Numerical experiments	120
6.7	Discussion	121

7	Qubit assignment using time reversal	124
7.1	Chapter overview	124
7.2	Introduction	125
7.3	Methods	127
7.3.1	Qubit assignment on hardware	127
7.3.2	Loschmidt Echoes for evaluating qubit assignments	128
7.3.3	Simulated annealing for qubit selection	132
7.3.4	Logical and physical connectivity	133
7.4	Experiments	134
7.4.1	Simulations	134
7.4.2	Hardware Experiments	136
7.5	Discussion	139
8	Conclusion	145
	References	148
	APPENDICES	175
A	Supplementary material to Chapter 3	176
A.1	Solution for classical overparameterized model	176
A.1.1	Linearization of overparameterized Fourier model	176
A.1.2	Solutions for special cases	186
A.1.3	Conditions for Benign overfitting	188
A.2	Solution for quantum overparameterized model	192
A.2.1	Computing feature weights of typical quantum models	195
A.2.2	The effect of state preparation on feature weights in the general quantum model	200
A.3	Determining the degeneracy of quantum models	204

A.3.1	Hamming encoding strategy	205
A.3.2	Binary encoding strategy	206
A.3.3	Ternary encoding strategy	206
A.3.4	Nonseparable models	207
A.4	Demonstration of benign overfitting in the general quantum model	208
A.5	Derivation of weight behavior in deep quantum kernels	211
B	Supplementary material to Chapter 6	218
B.1	Proof of Proposition 19	218
B.2	Proof of Theorem 20 and Corollary 21	220
B.3	Reduction of Algorithm 1 to single bitstrings	222
B.4	Modified perturbative technique	224
B.5	Additional numerics	224
B.5.1	Response matrices measured on IBM QPUs	224
B.5.2	Sampling error and comparison to M3	228
C	Supplementary material to Chapter 7	230
C.1	Certifying optimal qubit assignments	230
C.1.1	Local depolarizing noise	230
C.1.2	Alternative noise models	238
C.2	Readout error mitigation	241
C.3	Mapping logical circuits to physical circuits	243
C.3.1	Assignment of circuits with unconstrained logical connectivity	245
C.4	Device specifications and additional experiments	247
C.4.1	Quantum Fourier Transform	250
C.4.2	GHZ State Direct Fidelity Estimation	252

List of Figures

2.1	Intuition for bandwidth in quantum kernel methods	22
2.2	Analytical evidence for generalization improvement from kernel bandwidth	24
3.1	Intuition behind the phenomenon of benign overfitting with spiking models	30
3.2	Toy model of fitting behavior for under- vs. overparameterized quantum models	34
3.3	Bias-variance tradeoff based on engineered feature weights	38
3.4	Degeneracies and fitting behaviors of various quantum models	44
3.5	Benign overfitting for the general quantum model	46
4.1	Circuit diagram for classical bit flip channel with correlated auxiliary noise	55
4.2	Improvement in channel capacity and estimation tasks with and without correlated auxiliary	60
4.3	3-qubit code syndrome circuit	63
4.4	Information-theoretic upper- and lower- bounds for decoding 3-qubit code	64
4.5	Information-theoretic upper- and lower- bounds for decoding 5-qubit code	64
4.6	Quantum bit flip channel with correlated auxiliary noise	68
5.1	Schematic comparison of classical multiple hypothesis testing versus maximizing singlet fraction	85
6.1	Simulated performance of readout error mitigation for all-zeros bitstring	117
6.2	Simulated performance of perturbative readout error mitigation	122

7.1	Loschmidt Echo diagnostic circuit	130
7.2	Simulated comparison of F_{LE} to F_0 and fidelity F	141
7.3	Simulated results of simulated annealing for qubit assignment	142
7.4	Experimental results for F_{LE} versus GHZ state fidelity	143
7.5	Experimental results of simulated annealing for qubit assignment to prepare GHZ state	144
A.1	Concentration of feature weights for different encoding circuits	199
B.1	Failure of perturbative readout error correction due to divergence of Neu- mann series	225
B.2	Modified readout error mitigation using exact inverse	226
B.3	Experimental performance of all-zeros readout error mitigation on IBM device	227
B.4	Comparison to M3 readout error mitigation	229
C.1	Local noise model circuit diagram	232
C.2	Simulated annealing versus random qubit assignment	246
C.3	Error heatmap for Google quantum computer	248
C.4	Circuit diagrams for qubit assignment experiments	249
C.5	Experimental results of simulated annealing for qubit assignment to apply Quantum Fourier Transform	251
C.6	Experimental results for F_{LE} versus random Clifford circuit state fidelity .	253
C.7	Experimental results for F_{LE} versus SWAP network fidelity	254

List of Tables

3.1	Spectra computations for different encoding strategies	43
7.1	Experimental results for the correlation between F_{LE} and fidelity	138
A.1	Variance computations for different encoding strategies	198

Chapter 1

Introduction

Beginning with modest implementations on rudimentary hardware, classical computing has seen tremendous advances in algorithmic efficiency and computational power. The exponential growth in these areas has made it possible to implement Machine Learning (ML) algorithms that essentially automate the solving of a computational problem by consuming large amounts of data. ML algorithms have become ubiquitous and frequently surpass the performance of more traditional (i.e. not data-driven) computational methods implemented by humans in a variety of subject areas. Meanwhile, quantum computing has (under standard assumptions) qualitatively extended the class of algorithms that we might reasonably implement on a computing device. Indeed, a significant motivation for building quantum computers comes from the belief that they may perform certain computations more efficiently than the best known classical algorithms.

Each of these computational paradigms is promising on its own, with a wealth of empirical evidence for the utility of machine learning and complexity-theoretic arguments for the advantage of certain quantum algorithms. Despite this, Quantum Machine Learning (QML) – the use of quantum computers to train and evaluate algorithms by learning from some dataset of classical data or quantum data – sits in an awkward position. The field has inherited a central challenge of quantum algorithms – to demonstrate some kind of advantage over classical computation – without a key advantage of classical machine learning: the ability to demonstrate results empirically, without any need for a theoretical or rigorous characterization of the solution. As such, research in QML frequently involves either theoretically characterizing techniques (that may not actually provide advantage over classical techniques) or proposing entirely new frameworks for applying quantum computers to learning problems. This thesis pursues both of these outcomes.

To provide perspective for QML, we first review kernel methods from classical machine learning. Kernel methods can be understood in the following way [9]: beginning with some data x that has been labelled as belonging to one of two classes, we embed the data into a feature vector $\phi(x)$ in a Hilbert space, with the goal of making embedded data from distinct classes linearly separable in that feature space. This can be done for data that are labelled by a complex, nonlinear function, in part because the feature space may be high-dimensional, or even infinite dimensional. Kernel methods are then a class of learning algorithms that depend only on the inner product $k(x, x') := \langle \phi(x), \phi(x') \rangle$ between any two feature vectors. However, these methods turn out to be broadly applicable because *any* positive semidefinite function k can be used in this way, without the need to describe $\phi(x)$ explicitly [10]. Thus, to use a kernel method for learning, it is sufficient to specify a kernel k and evaluate $k(x, x')$ for all pairs of relevant data points.

Quantum kernel methods [11, 12] therefore seem like a very natural fit between quantum computers and machine learning: If we define an embedding $x \rightarrow \rho(x)$ into a space of quantum states, then we can have access to a high-dimensional feature space using only a few qubits, without the need for explicit access to the components of the density operator. To date, a few flavors of advantage have been proven for learning classical data with quantum kernel methods, usually by constructing a contrived dataset and then repurposing an existing quantum algorithm as a kernel method (e.g. Ref. [13] for the discrete logarithm [14] and Ref. [15] for “correlation” [16]). However, this thesis will avoid the topic of quantum advantage entirely. Instead, we will analyze a scenario that challenges a major premise of quantum kernels: Inner products between normalized vectors in high-dimensional spaces tend to vanish, causing the kernel method to fail. Chapter 2 analyzes *kernel bandwidth*, which mitigates this issue by restricting the effective dimension of the subspace containing embeddings $\rho(x)$, and can be implemented by simply rescaling the angles used by rotation gates to prepare $\rho(x)$.

Proving the efficacy of the bandwidth technique requires a discussion of the spectra of machine learning models [17]. The combination of kernel, a distribution of data, and a target function to be learned is naturally associated with a positive, trace-class operator T_k that can be diagonalized. In this picture, successfully learning from data (loosely) corresponds to *aligning* the model – i.e. preparing a function that is a linear combination of eigenfunctions of T_k with large eigenvalues, and ignoring the small-eigenvalue eigenfunctions that correspond to noise in the data. The latter detail means that a learning algorithm should not *overfit* the training data, i.e. perfectly fit every data point it is provided, since this will entail learning parts of the T_k eigendecomposition that describe noise. One advantage of this perspective is that you can immediately determine whether a learning problem is a candidate for quantum advantage by checking whether the princi-

pal eigenfunction is classically hard but quantumly easy to implement [18]. However, this alignment task can also be used to characterize unusual generalization behavior. In particular, a learning algorithm with support on all eigenfunctions of T_k , with coefficients tuned correctly, can perfectly overfit training data while optimally generalizing in an asymptotic limit. This phenomenon of *benign overfitting* was recently discovered for classical learning models, and Chapter 3 extends this analysis to identify unconventional generalization behavior in certain QML models.

The above discussion raises the question, is there nothing that can be done to mitigate the effect of noise in a learning algorithm? Perhaps out of convenience, most theoretical analyses in ML literature assume independent, identically distributed noise which cannot be removed from data. However, since noise arises from the interactions between physical systems, then given some noisy data is plausible that a discarded system contains correlations with the noise in data that otherwise manifests as i.i.d. Access to such *correlated auxiliary noise* (CAN) has been shown to can enhance the performance of learning algorithms for certain image recognition tasks [19]. Yet, there has not been any rigorous study of the power of this resource. Chapter 4 provides an information-theoretic characterization of CAN in the discrete setting, and improves our understanding of a model of CAN for quantum systems that had briefly been explored in other work [20].

Despite a large body of QML literature, it is unclear whether QML for classical data is a useful direction to pursue. However, there remains unexplored potential for applying quantum computers towards processing fundamentally quantum data types. For instance, we can consider whether providing classical data in superposition as a quantum state offers any interesting advantages for quantum learning algorithms. One such setting involves stochastically sampling a function f_i , providing the learner with a superposition of the form $\sum_x \sqrt{p(x)}|x, f_i(x)\rangle$ [21], and then studying if the learner has an advantage in determining i compared to classical query access to $(x, f_i(x))$ pairs [22–24]. However, at the basic level, this task still involves extracting classical information about a quantum system – what happens if this scenario is further modified to involve only quantum information? Can we characterize such a task, and is it meaningful? In Chapter 5 we will consider an entirely non-classical framework of “learning”, wherein both the input data to the algorithm and the set of possible labelling functions are provided in superposition. We will discuss how this task can be equally well understood as a protocol for maximizing the transmission of a particular entangled state between two parties, thereby bridging a gap between learning theory and entanglement theory.

While unproven to date, the possibility that quantum computers without error correction may provide an advantage over classical computers for useful tasks has received significant attention in recent years [25]. However, in the absence of fault-tolerance, it

will be impossible to extract meaningful results from such devices without mitigating the effects of errors that accumulate during computation due to environmental noise. The second topic of this thesis is devoted to this task. Readout error in such devices is essentially described by a classical, asymmetric bit flip channel involving up to 2^n parameters for n qubits [26]. Chapter 6 introduces a fast, approximate technique for readout error mitigation involving matrix truncations. Furthermore, without error correction, one confronts a scenario in which some qubits are better to use for a computation than others. To improve hardware outcomes in this scenario requires a method to efficiently estimate the performance of subsets of qubits, and then choose the best subset from among all available. To this end, Chapter 7 introduces a technique for optimizing qubit assignment by using the Loschmidt Echo [27] in combination with a simulated annealing algorithm [28].

Chapter 2

Quantum theory and kernel methods

2.1 Chapter overview

This chapter will introduce the linear algebra of quantum theory in finite dimensions, followed by kernel methods, a relatively simple class of techniques used for classical machine learning. We will discuss the relationship between the spectrum of the integral kernel operator and the ability of a kernel method to learn. This material will be useful later for understanding the concept of “benign overfitting” of quantum machine learning models explored in Chapter 3. We will then introduce quantum kernel methods, and show how their ability to learn can be harmed by the high dimensionality of quantum state space. Finally, we demonstrate how *kernel bandwidth* can mitigate this effect to enable generalization of quantum kernel methods.

Parts of this chapter are paraphrased or duplicated from Ref. [1]. The results in Secs. 2.4.1.2 and 2.4.2.1 were developed in close collaboration with Ruslan Shayludin. Sec. 2.4.2.2 contains contributions mainly by Abdulkadir Canatar and Ruslan Shayludin, which I have only included in a summarized form. The remaining material is either well-established in the literature or was developed mainly by myself.

2.2 Finite-dimensional quantum theory

We review quantum information theory in finite-dimensional Hilbert spaces. $\mathcal{H}$ will denote a finite-dimensional Hilbert space, containing pure states (unit-norm vectors) with the

standard inner product $\langle\psi|\phi\rangle$ denoted using bra-ket notation for $|\psi\rangle, |\phi\rangle \in \mathcal{H}$. In the finite-dimensional case, we will write $\mathcal{H} = \mathbb{C}^d$ for some d . $\mathcal{L}(\mathcal{H}, \mathcal{H}')$ then denotes the set of linear operators from $\mathcal{H}$ to $\mathcal{H}'$, which may be represented as matrices in $\mathbb{C}^{d \times d'}$ for $\dim(\mathcal{H}') = d'$, and is abbreviated $\mathcal{L}(\mathcal{H}, \mathcal{H}) := \mathcal{L}(\mathcal{H})$. An *observable* $A \in \mathcal{L}(\mathcal{H})$ is any self-adjoint operator, i.e. $A = A^\dagger$ where † denotes conjugate-transpose of the matrix representation of A . Given a Hilbert space $\mathcal{H}$, $\mathcal{D}(\mathcal{H})$ denotes the *density operators* or *states* defined on $\mathcal{H}$, i.e. the subset of $\mathcal{L}(\mathcal{H})$ acting on $\mathcal{H}$ that are unit trace and positive:

$$\mathcal{D}(\mathcal{H}) = \{\rho : \rho \geq 0, \text{Tr}(\rho) = 1\} \subset \mathcal{L}(\mathbb{C}^d), \quad (2.1)$$

and $A \geq 0$ is shorthand to denote that $A \in \mathcal{L}(\mathcal{H})$ is *positive (semi-definite)*, such that $\langle v|A|v\rangle \geq 0$ for all $|v\rangle \in \mathcal{H}$. For two operators $A, B \in \mathcal{L}(\mathcal{H})$, we can further define a Hilbert–Schmidt inner product

$$\langle A, B \rangle = \text{Tr}(A^\dagger B), \quad (2.2)$$

which is the natural inner product for (self-adjoint) operators in $\mathcal{L}(\mathcal{H})$.

If we associate a physical system labelled X with a Hilbert space $\mathcal{H}_X$, then we may describe a combined system XY in terms of a tensor product of spaces, $\mathcal{H}_{XY} = \mathcal{H}_X \otimes \mathcal{H}_Y$, wherein $\dim(\mathcal{H}_{XY}) = \dim(\mathcal{H}_X) \dim(\mathcal{H}_Y)$. A linear map $\mathcal{N} : \mathcal{L}(\mathcal{H}) \rightarrow \mathcal{L}(\mathcal{H}')$ is positive if it maps positive operators to positive operators, and is completely positive if the map

$$(\mathbb{I}_n \otimes \mathcal{N}) : (\mathcal{L}(\mathbb{C}^n) \otimes \mathcal{L}(\mathcal{H})) \rightarrow (\mathcal{L}(\mathbb{C}^n) \otimes \mathcal{L}(\mathcal{H}')) \quad (2.3)$$

is positive for any integer n . Such a map is trace-preserving if $\text{Tr}(\mathcal{N}(A)) = \text{Tr}(A)$ for all $A \in \mathcal{L}(\mathcal{H})$. A special subset of linear maps – quantum *channels*, or *completely positive trace-preserving (CPTP)* maps – is denoted $\mathcal{C}(\mathcal{H}, \mathcal{H}')$ and contains all maps from $\mathcal{L}(\mathcal{H}) \rightarrow \mathcal{L}(\mathcal{H}')$ that are completely positive and trace preserving. The set of *isometries* from $\mathcal{H} \rightarrow \mathcal{H}'$ is denoted $\mathcal{U}(\mathcal{H}, \mathcal{H}')$, and contains linear operators V such that $V^\dagger V = \mathbb{I}_{\mathcal{H}}$. *Unitary* operators $\mathcal{U}(\mathcal{H})$ consist of isometries $U : \mathcal{H} \rightarrow \mathcal{H}$ that further satisfy $UU^\dagger = \mathbb{I}_{\mathcal{H}}$.

The evolution of quantum states in closed systems is described by applying unitary operators to ρ . An arbitrary pure state may be prepared by applying a unitary operator $U \in \mathcal{U}(\mathcal{H})$ to the state $|0\rangle$, where $|j\rangle$ represents the j th standard basis vector $\hat{e}_j$ in $\mathbb{C}^d$. A *qubit* is the name for a physical system described by a Hilbert space $\mathcal{H} = \mathbb{C}^2$, and so the state of an N -qubit system lives in $\mathbb{C}^{2^N}$. When $N = 1$, quantum states may be represented by the *Bloch ball*. The constraints $\text{Tr}(\rho) = 1$ and $\rho = \rho^\dagger$ (implied by $\rho \geq 0$) mean that ρ is completely described by the *Bloch vector* of three real parameters $\mathbf{n} = (n_x, n_y, n_z)$. Then, ρ can be written in terms of Pauli matrices X, Y, Z which form a basis for $\mathcal{L}(\mathcal{H})$ that is

orthonormal with respect to $\langle \cdot, \cdot \rangle_{HS}$. Denoting

$$\vec{\sigma} = \left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right) := (X, Y, Z), \quad (2.4)$$

then a single qubit state may always be expressed as $\rho = \frac{1}{2} (\mathbb{I} + \mathbf{n} \cdot \vec{\sigma})$, i.e. $n_x = \langle X, \rho \rangle / 2$. Since $\|\mathbf{n}\| \leq 1$ is required for $\rho \geq 0$ subject to $\text{Tr}(\rho) = 1$, we can identify any single-qubit state ρ with a vector $\mathbf{n}$ in the unit ball in $\mathbb{R}^3$. Similarly, any channel acting on a single qubit may be represented as a map between points on the Bloch sphere [29], making this representation convenient for visualizing feature maps associated with quantum states.

2.3 Learning with kernels

We now provide an overview of the theory of learning with kernels. A typical supervised learning problem begins with pairs $(x, y) \in \mathcal{X} \times \mathbb{R}$ distributed jointly according to a density $p_{XY} : \mathcal{X} \times \mathbb{R} \rightarrow \mathbb{R}^+$ (we will take $\mathcal{X} \subset \mathbb{R}^D$ without loss of generality). Thus, each x is associated with a *label* $y \in \mathbb{R}$, and $x \in \mathcal{X}$ is distributed according to a marginal density function $p_X : \mathcal{X} \rightarrow \mathbb{R}$.

The goal of a learning algorithm is, given *training data* consisting of n labelled data points $\mathcal{T} := \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathcal{X}^n \times \mathbb{R}^n$, to output a function f that *generalizes*, such that $f(x) \approx y$ for any x sampled from the underlying distribution p_X . A naive approach to this task would be to choose any function that satisfies $f(x_i) = y_i$ for all $(x_i, y_i) \in \mathcal{T}$ (*overfits* the data), thereby perfectly fitting the observed data. However, this will tend to fail when there is noise present in the labels. For instance, if the labels are generated by some f^* such that $y = f^*(x) + \epsilon$ for some randomly distributed ϵ , a function that overfits the training data may fail to generalize (in Chapter 3 we will demonstrate a counterexample to this general principle).

A learning algorithm must therefore balance the goal of correctly modelling the training data with a tendency to overfit. One way to manage this tradeoff is by fixing a set of functions $\mathcal{C}$ and then selecting a function according to the optimization

$$\arg \min_{f \in \mathcal{C}} \left(\frac{1}{2} \sum_{i=1}^n (f(x_i) - y_i)^2 + \frac{\lambda}{2} \|f\|_*^2 \right) \quad (2.5)$$

where we have introduced the (unspecified) norm $\|\cdot\|_*$ that will be used to penalize functions that are complex in some sense, which includes functions with the capacity to perfectly

fit the provided training data. Since a function that overfits the training data results in a small value for the first term of Eq. 2.5, λ can be used to increase the value of the second term in such scenarios, thereby mitigating overfitting.

We now introduce kernel methods as a technique for constructing solutions to the optimization of Eq. 2.5. A real *kernel* function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is any function that is positive semidefinite in the sense that

$$\langle c, Kc \rangle = \sum_{i,j=1}^n k(x_i, x_j) c_i c_j \geq 0 \quad (2.6)$$

for all $(c_1, \dots, c_n) \in \mathbb{R}^n$ and $(x_1, \dots, x_n) \in \mathcal{X}^n$. Above we have also defined the positive semidefinite *Gram matrix* elementwise by $K_{ij} = k(x_i, x_j)$. A consequence of the Moore–Aronszajn theorem [10] is that every such k may be expressed as

$$k(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{H}} \quad (2.7)$$

for some choice of *feature map* $\psi : \mathcal{X} \rightarrow \mathcal{H}$ and Hilbert space $\mathcal{H}$. Thus, intuitively $k(x, x')$ may be understood to be computing a similarity between two points x, x' , with the specific notion of similarity being determined implicitly by the choice of feature map ψ or kernel k . In particular, the Gram matrix completely describes this measure of similarity between every pair of points in the training set $\mathcal{T}$.

A kernel (or feature map ψ) and fixed training set $\mathcal{T}$ give rise to a *reproducing kernel Hilbert space* (RKHS), which is the completion of the vector space spanned by the set of bounded functions $\{k_{x_i} : i = 1, \dots, n\}$ where $k_{x_i} := k(x_i, \cdot) = \langle \psi(x_i), \cdot \rangle$. The reproducing property of the RKHS $\mathcal{F}$ is

$$k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{F}}. \quad (2.8)$$

This allows us to construct an integral operator that characterizes kernel methods for any specific choice of kernel k and data distribution p_X . Suppose $\mu(dx) := p_X(x)dx$ denotes the probability measure for data distributed on $\mathcal{X}$, and let $L_2^\mu(\mathcal{X})$ be the space of square integrable functions $f : \mathcal{X} \rightarrow \mathbb{R}$ with norm $\|f\|_\mu^2 = \langle f, f \rangle_\mu = \int_{\mathcal{X}} |f(x)|^2 \mu(dx)$. Then the integral operator $T_k : L_2^\mu(\mathcal{X}) \rightarrow L_2^\mu(\mathcal{X})$ is defined according to its action [30]

$$(T_k f)(x) = \int_{\mathcal{X}} k(x, x') f(x') \mu(dx'). \quad (2.9)$$

(It is from this equation for an integral kernel operator that the kernel $k(x, x')$ derives its name). By Mercer’s theorem [31], the eigenfunctions of T_k satisfying $T_k \phi_k = \eta_k \phi_k$

are orthonormal (i.e., $\langle \phi_k, \phi_l \rangle_\mu = \delta_{kl}$) and span $L_2^\mu(\mathcal{X})$. Indeed, T_k is a positive (semi-definite), trace class operator and admits a spectral decomposition. This gives rise to an eigendecomposition of k given by $k(x, x') = \sum_k \eta_k \phi_k(x) \phi_k^*(x')$, where $\{\eta_k\}$ are real, nonnegative eigenvalues of T_k . Then, the inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ in the RKHS of k may be computed with respect to T_k in Eq. 2.9 as $\langle f, g \rangle_{\mathcal{F}} = \langle f, T_k^{-1} g \rangle_\mu$, which is well-defined for all $f, g \in \mathcal{F}$. From the kernel eigendecomposition, assuming that the labels $y = g(x) + \epsilon$ are generated by a *target function* $g \in L_2^\mu(\mathcal{X})$ in the RKHS, then g may also be decomposed as $g(x) = \sum_k \hat{g}_k \phi_k(x)$, where $\hat{g}_k = \langle g, \phi_k \rangle_\mu$ are the target weights or *Fourier coefficients*.

Returning to the optimization problem of Eq. 2.5, Kernel ridge regression (KRR) is a convex optimization problem over functions in an RKHS $\mathcal{F}$ taking the form:

$$f^* = \arg \min_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^n (f(x_i) - y_i)^2 + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2, \quad (2.10)$$

where $\|\cdot\|_{\mathcal{F}}$ denotes the norm with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ on the RKHS and $\lambda \geq 0$ is the ridge parameter introduced for *regularizing* (reducing the complexity of) the solution. The optimum of Eq. 2.10 may always be expressed as a linear combination of elements in the RKHS $\mathcal{F}$ of k [32, 33]. Thus, using these definitions one can show that the solution to Eq. 2.10 takes the form $f^*(x) = \mathbf{k}(x)^\top (K + \lambda \mathbf{I})^{-1} \mathbf{y}$, where $\mathbf{k}(x)$ is a vector with elements $(\mathbf{k}(x))_i = k(x, x_i)$ and $(\mathbf{y})_i = y_i$. This solution, an example of the so-called “kernel trick” [9], allows one to perform ridge regression in an RKHS without explicitly computing elements of either $\mathcal{H}$ or $\mathcal{F}$, instead requiring one only to evaluate the kernel function k on training data in $\mathcal{T}$. This abstraction away from the need to express elements of the Hilbert space $\mathcal{H}$ is what will allow us to compute k using a physical, quantum system without requiring query access to amplitudes of a state.

We now discuss how a quantum Hilbert space may be used to embed classical data, giving rise to a *quantum kernel* [11, 12]. While the literature contains some discussion of the potential for quantum kernel methods to have an advantage over classical techniques in either computational or sample complexity [13, 15, 34], we will not address questions of advantage in this thesis. Fixing $\mathcal{H}_0$, start by defining a data-dependent unitary operator $U(x) \in \mathbf{U}(\mathcal{H}_0)$ that prepares a density matrix $\rho(x) = U(x)|0\rangle\langle 0|U^\dagger(x)$. A natural choice for the feature map $\psi : \mathcal{X} \rightarrow \mathcal{H}$ is then

$$\psi(x) := \rho(x) = |\psi(x)\rangle\langle\psi(x)| \in \mathbf{D}(\mathcal{H}_0)$$

with the corresponding inner product $\langle \rho, \rho' \rangle_{\mathcal{H}} = \text{Tr}(\rho \rho')$. This identifies $\mathcal{H}$ as the subset of Hermitian operators in $\mathcal{L}(\mathcal{H}_0)$ equipped with Hilbert-Schmidt inner product, though

physically realizable feature maps will only ever use the positive, unit trace subset $D(\mathcal{H}_0)$ of $\mathcal{H}$. Under this feature map, the quantum kernel is defined as

$$k(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{H}} = \text{Tr}(\rho(x)\rho(x')), \quad (2.11)$$

which inherits (conjugate) symmetry in its arguments from $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A quantum kernel method is then any learning algorithm that involves computing the Gram matrix K on a quantum computer, for instance by evaluating the value of the observable $|0\rangle\langle 0|$ with respect to the state $U(x)U(x')|0\rangle\langle 0|U^\dagger(x')U^\dagger(x)$. With this association between the feature map $\psi(x)$ and the quantum state $\rho(x)$, we can apply a rich existing theory for kernel methods in classical machine learning to study the spectra and generalization behavior of learning algorithms that use quantum kernels.

2.3.1 Kernel spectra

We now further integral kernel operators for finite-dimensional complex Euclidian Hilbert space, and introduce a covariance operator that shares many properties with T_k . The goal is to relate the spectra of these operators and to better understand how feature maps and distributions of input data combine to affect the learnability of a task.

A quantum kernel may alternatively be expressed as $k(x, x') = \text{Vec}(\rho(x))^\dagger \text{Vec}(\rho(x'))$, where we have introduced the vectorization map $\text{Vec} : \mathcal{L}(\mathbb{C}^d) \rightarrow \mathbb{C}^{d^2}$, which stacks rows of a $d \times d$ matrix into a $d^2 \times 1$ column vector (e.g. [35]). Throughout this section we will continue to consider d -dimensional state space (where $d = 2^N$ in the case of N qubits). We will frequently use the identity

$$\text{Vec}(A)^\dagger \text{Vec}(B) = \langle A, B \rangle = \text{Tr}(A^\dagger B), \quad \forall A, B \in \mathcal{L}(\mathbb{C}^d).$$

Observing that the RKHS of a kernel k is defined (see, e.g., [36]) by functions of the form

$$f(x) = \langle \rho(x), H \rangle,$$

for self-adjoint H , we can rewrite Eq. 2.9 as

$$\begin{aligned} (T_k f)(x) &= \text{Vec}(\rho(x))^\dagger \int_{\mathcal{X}} \text{Vec}(\rho(x')) \text{Vec}(\rho(x'))^\dagger \text{Vec}(H) \mu(dx') \\ &= \langle \text{Vec}(\rho(x)), \Sigma \text{Vec}(H) \rangle, \end{aligned} \quad (2.12)$$

where we have defined the *covariance operator* $\Sigma \in \mathcal{L}(\mathcal{H} \otimes \mathcal{H})$,

$$\Sigma = \int_{\mathcal{X}} \text{Vec}(\rho(x')) \text{Vec}(\rho(x'))^\dagger \mu(dx') = \int_{\mathcal{X}} \rho(x') \otimes \rho(x')^T \mu(dx'). \quad (2.13)$$

We can intuitively understand this operator by analogy to linear feature maps: If we consider n -dimensional real-valued input vectors $x \in \mathcal{X} \subset \mathbb{R}^D$ and assume x are centered such that $\mathbb{E}_{\mathcal{X}}[x] = 0$, then the covariance operator is defined as $\Sigma = \mathbb{E}_{\mathcal{X}}[xx^T]$ such that $(\Sigma)_{ij} = \mathbb{E}[x_i x_j]$. In classical machine learning literature, this operator (and its spectrum) plays a significant role characterizing the generalization behavior of *linear regression* (i.e. Eq. 2.10 with $\lambda = 0$ and $k(x, x') = \langle x, x' \rangle$) [37, 38]. In turn, Equation 2.13 describes a kind of quantum covariance operator, although this is an imperfect analogy since a centered quantum feature map ($\mathbb{E}_{\mathcal{X}}[\rho(x)] = 0$) would violate the requirement $\text{Tr}(\rho(x)) = 1$. We are interested in the eigenvalue equations

$$\begin{aligned}\Sigma \text{Vec}(H) &= \eta \text{Vec}(H) \\ (T_k \phi)(x) &= \eta \phi(x).\end{aligned}$$

Given the operators defined in Eq. 2.9 and Eq. 2.13, one may verify (see, e.g., [39]) that the following statements hold under the identification $\psi(x) \rightarrow \text{Vec}(\rho(x))$ described above.

- (S1) For every eigenfunction ϕ satisfying $(T_k \phi)(x) = \eta \phi(x)$, there is a corresponding eigenvector $\text{Vec}(H)$ of Σ given by

$$\text{Vec}(H) = \int_{\mathcal{X}} \phi(x) \text{Vec}(\rho(x)) \mu(dx)$$

such that $\Sigma \text{Vec}(H) = \eta \text{Vec}(H)$. Furthermore, from the bijectivity of the Vec operation and the fact that Hermiticity is preserved under convex combination, it follows that H is Hermitian.

- (S2) For every eigenvector $\text{Vec}(H)$ of Σ satisfying $\Sigma \text{Vec}(H) = \eta \text{Vec}(H)$, there is a corresponding eigenfunction of T_k given by

$$\phi(x) = \text{Tr}(\rho(x)H) = \text{Vec}(\rho(x))^{\dagger} \text{Vec}(H)$$

such that $(T_k \phi)(x) = \eta \phi(x)$.

- (S3) Assuming that eigenfunctions ϕ_k of T_k indexed by eigenvalue η_k are orthonormal,

$$\langle \phi_k, \phi_l \rangle_{\mu} = \int_{\mathcal{X}} \phi_k(x)^* \phi_l(x) \mu(dx) = \delta_{kl},$$

the eigenvectors $\text{Vec}(H_k)$ of Σ indexed according to eigenvalue η_k satisfy

$$\text{Vec}(H_k)^{\dagger} \text{Vec}(H_l) = \text{Tr}(H_k^{\dagger} H_l) = \eta_k^{-1} \delta_{kl}$$

whenever $\eta_k > 0$ (when $\eta_k = 0$, we ignore $\text{Vec}(H_k)$ in the null space of Σ).

Letting $\text{spec}(A)$ denote the sequence of eigenvalues of an operator A sorted in non-increasing order, it then follows from (S1) and (S2) that

$$\text{spec}(T_k) = \text{spec}(\Sigma).$$

2.3.2 Quantum feature space is not always quantum state space

In the interest of exposition, here we provide an example of a quantum kernel for which the feature space used to compute the kernel does not coincide with the quantum state space $\mathcal{D}(\mathcal{H})$ (for any choice of $\mathcal{H}$), and in fact the feature space is infinite dimensional for a quantum circuit defined on any finite dimensional system (though the resulting kernel may always be approximated by some choice of quantum feature space [40]). This further supports the argument that machine learning using quantum computers is a much richer field than simply computing quantum kernels on physical devices [41]. This analysis will also foreshadow further analyses in Chapter 3, based on translation-invariant kernels.

Fixing a Hilbert space $\mathcal{H}$ and data $x \in \mathcal{X} \subset \mathbb{R}^D$, and consider a single-layer translation-invariant quantum feature map given by

$$\phi(x) = \rho(x) = |\psi(x)\rangle\langle\psi(x)| \quad (2.14)$$

$$|\psi(x)\rangle = S(x)U|0\rangle := S(x)|\Gamma\rangle, \quad (2.15)$$

where $U, S(x) = \exp(-ix \cdot H) \in \mathcal{U}(\mathcal{H})$ and $H = (H_1, \dots, H_D)$ is a sequence of Hermitian operators $H_k \in \mathcal{L}(\mathcal{H})$. This induces a kernel

$$k(x, x') = \langle\phi(x), \phi(x')\rangle_{HS} = \text{Tr}(\rho(x)\rho(x')) = |\langle\Gamma|S(x - x')|\Gamma\rangle|^2, \quad (2.16)$$

where we have used linearity in the argument $x \cdot H$ generating S . This is a *translation-invariant* kernel, since we can take $\delta = x - x'$ to define

$$k(x, x') = k(\delta) = |\langle\Gamma|S(\delta)|\Gamma\rangle|^2. \quad (2.17)$$

Using the techniques of [42], we rewrite this as a Fourier series

$$k(\delta) = |\langle 0|U^\dagger \exp(-i\delta \cdot \mathbf{H}) U|0\rangle|^2 \quad (2.18)$$

$$= \sum_{\mathbf{j}, \mathbf{k}=0}^{2^N-1} e^{-i(\lambda_{\mathbf{j}} - \lambda_{\mathbf{k}}) \cdot \delta} |U_{j0}|^2 |U_{k0}|^2 \quad (2.19)$$

$$= \sum_{\omega \in \Omega} e^{-i\omega \cdot \delta} \sum_{\substack{\mathbf{j}, \mathbf{k}: \\ \lambda_{\mathbf{j}} - \lambda_{\mathbf{k}} = \omega}} |\gamma_{\mathbf{j}}|^2 |\gamma_{\mathbf{k}}|^2 \quad (2.20)$$

$$= \sum_{\omega \in \Omega} k_{\omega} e^{-i\omega \cdot \delta}. \quad (2.21)$$

Eq. 2.21 provides a Fourier representation for the translation-invariant kernel. Then, Bochner's theorem [43] states that a translation-invariant kernel is positive semidefinite if and only if its (Fourier) spectrum is a positive measure. By inspection, we see that $k_{\omega} \geq 0$ and $\sum_{\omega \in \Omega} k_{\omega} = k(0) = 1$. Thus, k is a positive semi-definite kernel.

The above setup provides a generic template for constructing kernels that are computed on a quantum computer, but for which the feature space does not contain quantum states. One strategy is to introduce a preprocessing function $g : \mathbb{R}^D \rightarrow \mathbb{R}$ that acts on $\delta = x - x'$ and employ the Fourier series of Eq. 2.21 for a scalar input. For instance, letting $g(\delta) = \|\delta\|$, choose a diagonal Hamiltonian H and define

$$k(\delta) = |\langle 0|U^\dagger \exp(-i\|\delta\|H) U|0\rangle|^2 = \sum_{\omega \in \Omega} k_{\omega} e^{-i\omega \|\delta\|} \quad (2.22)$$

We now need to demonstrate that k is positive semidefinite before we may identify it as a kernel. We can no longer rely on the fact that $k(\delta) = \langle \rho(x), \rho(x') \rangle$ since encoding $\|x - x'\|$ eliminates the choice of $|\psi(x)\rangle = S(x)|\Gamma\rangle$ as the explicit feature map giving rise to k . Following Ref. [44] loosely, we use the series expansion

$$e^{-i\omega \|x - x'\|} = e^{-i\omega(\|x\|^2 + \|x'\|^2 - 2x \cdot x')} \quad (2.23)$$

$$= e^{-i\omega \|x\|^2} e^{-i\omega \|x'\|^2} \sum_{k=0}^{\infty} \frac{(i2\omega)^k}{k!} (x \cdot x')^k. \quad (2.24)$$

Then, we expand the dot product term using multinomial theorem,

$$(x \cdot x')^k = \sum_{\|\mathbf{r}\|_1=k, \mathbf{r} \geq 0} c_{\mathbf{r}} p_{\mathbf{r}}(x) p_{\mathbf{r}}(x') \quad (2.25)$$

$$c_{\mathbf{r}} = \binom{k!}{r_1! \dots r_D!} \quad (2.26)$$

$$p_{\mathbf{r}}(x) = \prod_{j=1}^D x_j^{r_j}, \quad (2.27)$$

where we index over a countable number of vectors $\mathbf{r} \in \mathbb{Z}^D$. Substituting, we get

$$e^{-i\omega\|x-x'\|} = e^{-i\omega\|x\|^2} e^{-i\omega\|x'\|^2} \sum_{k=0}^{\infty} \frac{(i2\omega)^k}{k!} \sum_{\|\mathbf{r}\|_1=k, \mathbf{r} \geq 0} c_{\mathbf{r}} p_{\mathbf{r}}(x) p_{\mathbf{r}}(x') \quad (2.28)$$

$$= e^{-i\omega\|x\|^2} e^{-i\omega\|x'\|^2} \sum_{\mathbf{r} \geq 0} b_{\mathbf{r},\omega} c_{\mathbf{r}} p_{\mathbf{r}}(x) p_{\mathbf{r}}(x') \quad (2.29)$$

$$= \langle \phi_{\omega}(x), \phi_{\omega}(x') \rangle, \quad (2.30)$$

where $b_{\mathbf{r},\omega} = (i2\omega)^{\|\mathbf{r}\|_1} / (\|\mathbf{r}\|_1!)$ and we recognize a feature vector in (countably) infinite dimensional space indexed by $\mathbf{r}$

$$[\phi_{\omega}(x)]_{\mathbf{r}} = e^{-i\omega\|x\|^2} (b_{\mathbf{r},\omega} c_{\mathbf{r}})^{1/2} p_{\mathbf{r}}(x). \quad (2.31)$$

Finally, we substitute into Eq. 2.23 to recover

$$k(x - x') = \sum_{\omega \in \Omega} k_{\omega} \langle \phi_{\omega}(x), \phi_{\omega}(x') \rangle. \quad (2.32)$$

Since $k_{\omega} \geq 0$ and a convex sum of positive functions is also positive, k is also positive. Indeed, the explicit feature map we have derived is

$$[\psi(x)]_{\mathbf{r},\omega} = e^{-i\omega\|x\|^2} (k_{\omega} b_{\mathbf{r},\omega} c_{\mathbf{r}})^{1/2} p_{\mathbf{r}}(x) \quad (2.33)$$

for which $k(x, x') = \langle \psi(x), \psi(x') \rangle$ where the inner product is defined in the countably infinite-dimensional space indexed by $(\mathbf{r}, \omega)$. Similar types of quantum kernels (i.e. ones that do not arise from inner products in a quantum state space) may be defined for scalar arguments like $\|\delta\|^2$ and $\langle x, x' \rangle$ in which case $k(x, x') = k_r(\langle x, x' \rangle)$ is rotation invariant rather than translation-invariant. Presumably, by the arguments above *any* positive semi-definite function on $\mathcal{X} \times \mathcal{X}$ can be substituted where we used $\|\cdot\|$, though this is not obviously practical.

2.3.3 Kernel spectra and generalization

We will now discuss how the generalization capability of a kernel method is associated with the spectrum of the integral kernel operator. To begin, we clarify what is meant by “generalization” here. We suppose there is a target function $g : \mathcal{X} \rightarrow \mathbb{R}$ to learn, and a learning model outputs some function $f : \mathcal{X} \rightarrow \mathbb{R}$. We may indicate that the output function is trained on a dataset $X = (x_1, \dots, x_n)$ and sequence of labels $\mathbf{y} = (y_1, \dots, y_n)$ by writing $f(x; X, \mathbf{y})$. Moreover, the labels $y_i = g(x_i) + \epsilon$ contain random noise. Then, for some f trained on a fixed dataset X , we are interested in the mean squared error of f for predicting on an unseen data point $x \in \mathcal{X}$ sampled randomly with respect to the distribution of data $p_X(x)$, averaged with respect to the noise in the labels, i.e.

$$E(n) := \mathbb{E}_{x, \mathbf{y}} (f(x; X, \mathbf{y}) - g(x))^2 := \mathbb{E}_{\mathbf{y}} \int_{\mathcal{X}} \mu(dx) (f(x; X, \mathbf{y}) - g(x))^2 \quad (2.34)$$

Assuming that there is a kernel k associated with the learning algorithm that outputs f , we can use the existence of an orthonormal basis for the integral kernel operator T_k to simplify the analysis (for now we only summarize the steps, see App. A.1.1.2 for more details). T_k has eigenvalues η_i associated with eigenfunctions ϕ_i that are orthonormal with respect to the measure $\mu(dx) = p(x)dx$, so that $\langle \phi_i, \phi_j \rangle_{\mu} = \delta_{ij}$. We can then write f and g in this basis of eigenfunctions,

$$f(x; X, \mathbf{y}) = \sum_i \hat{f}_i(X, \mathbf{y}) \phi_i(x), \quad g(x) = \sum_i \hat{g}_i \phi_i(x). \quad (2.35)$$

where the coefficients of f in the eigenbasis of T_k depend on the training data received by the learning algorithm. Defining $\phi(x) = (\phi_1(x), \dots, \phi_i(x), \dots)$ and $\hat{f}_i := \hat{f}_i(X, \mathbf{y})$ to write $\hat{f} = (\hat{f}_1, \dots, \hat{f}_i, \dots)$, each function is simply

$$f(x; X, \mathbf{y}) = \langle \hat{f}, \phi(x) \rangle \quad (2.36)$$

$$g(x) = \langle \hat{g}, \phi(x) \rangle. \quad (2.37)$$

We will consider the expected value $\mathbb{E}_{\mathbf{y}} f(x; X, \mathbf{y})$ of the function with respect to the noise in the labels $\mathbf{y}$. This describes the behavior of the model averaged over all possible noisy

labels corresponding to X . Using this, we decompose Eq. 2.34 as the sum of two terms:

$$E(n) = \mathbb{E}_{x,\mathbf{y}}(f(x; X, \mathbf{y}) - \mathbb{E}_{\mathbf{y}}f(x; X, \mathbf{y}) + \mathbb{E}_{\mathbf{y}}f(x; X, \mathbf{y}) - g(x))^2 \quad (2.38)$$

$$= \mathbb{E}_{x,\mathbf{y}}(f(x; X, \mathbf{y}) - \mathbb{E}_{\mathbf{y}}f(x; X, \mathbf{y}))^2 + \mathbb{E}_{x,\mathbf{y}}(\mathbb{E}_{\mathbf{y}}f(x; X, \mathbf{y}) - g(x))^2 \quad (2.39)$$

$$= \mathbb{E}_{x,\mathbf{y}}|\langle \phi(x), \hat{f} - \mathbb{E}_{\mathbf{y}}\hat{f} \rangle|^2 + \mathbb{E}_{x,\mathbf{y}}|\langle \phi(x), \mathbb{E}_{\mathbf{y}}\hat{f} - \hat{g} \rangle|^2 \quad (2.40)$$

$$= \mathbb{E}_{x,\mathbf{y}}(\hat{f} - \mathbb{E}_{\mathbf{y}}\hat{f})^\dagger \phi(x) \phi(x)^\dagger (\hat{f} - \mathbb{E}_{\mathbf{y}}\hat{f}) + \mathbb{E}_{x,\mathbf{y}}(\mathbb{E}_{\mathbf{y}}\hat{f} - \hat{g})^\dagger \phi(x) \phi(x)^\dagger (\mathbb{E}_{\mathbf{y}}\hat{f} - \hat{g}) \quad (2.41)$$

$$= \text{Tr} \left(\Sigma \cdot \mathbb{E}_{\mathbf{y}}(\hat{f} - \mathbb{E}_{\mathbf{y}}\hat{f})(\hat{f} - \mathbb{E}_{\mathbf{y}}\hat{f})^\dagger \right) + (\mathbb{E}_{\mathbf{y}}\hat{f} - \hat{g})^\dagger \Sigma (\mathbb{E}_{\mathbf{y}}\hat{f} - \hat{g}) \quad (2.42)$$

$$:= \text{Tr} \left(\Sigma \cdot \text{Cov}(\hat{f}, \mathbb{E}_{\mathbf{y}}\hat{f}) \right) + \|\mathbb{E}_{\mathbf{y}}\hat{f} - \hat{g}\|_\Sigma^2, \quad (2.43)$$

where in line 2.39 we have eliminated cross terms due to the orthonormality of $\{\phi_i\}$, and Σ is the covariance operator for the orthonormal eigenfunctions as defined in Eq. 2.13, with components $\Sigma_{ij} = \eta_i \delta_{ij}$. Here, $\|\cdot\|_\Sigma$ denotes a norm with respect to a (diagonal) metric Σ . In the final line, we have arrived at terms that are typically called the *variance* and the (squared) *bias* respectively.

We can use these terms to interpret how the generalization error depends on the spectrum of T_k contained on the diagonal of Σ . The second term of Eq. 2.43 gets smaller as the Fourier coefficients of $f(x; X, \mathbf{y})$ become more similar to those of $g(x)$, with the metric acting to punish larger discrepancies between f and g in subspaces of the RKHS where T_k has larger eigenvalues. The first term of Eq. 2.43 gets smaller as the statistical fluctuations of $\hat{f}$ (with respect to the noise in $\mathbf{y}$) shrink in the directions within the RKHS where T_k has larger eigenvalues. Thus, minimizing the error of a model depends strongly on the spectrum of the corresponding kernel. This analysis will be used in Chapter 3 to prove that certain quantum machine learning models can simultaneously generalize optimally while perfectly overfitting on training data.

A powerful tool that we will use later is Bochner's theorem [43], which states that a translation-invariant kernel is positive semidefinite if and only if its Fourier spectrum is a positive measure. This provides a reliable formula for deriving the spectra of certain kernels for certain data inputs: For a carefully chosen distribution of data, if you can express both the target function and the learning model as a Fourier series, then you can rephrase the learning problem in terms of a translation-invariant kernel $k(x, x')$ for which the eigendecomposition of operator T_k is known. Combined with the observation that functions computed by parameterized quantum circuits are expressible as a finite Fourier series [42], this provides a recipe for tractably studying the generalization behaviors of quantum machine learning models – an advantage that we will exploit in Chapter 3.

While the above analysis provides some intuition for the relationship between kernel spectra and generalization, it cannot yield further insight without a detailed analysis of

the learning algorithm for f and the distribution of learned Fourier coefficients. Analyzing such scenarios is an active area of research. In particular, Ref. [45] provides an analytical expression for the generalization error of the model f , averaged over training samples $(X, \mathbf{y})$. Specifically, for an operator T_k with $\text{spec}(T_k) = (\eta_1, \dots, \eta_i, \dots)$, and for a dataset of fixed size n , one may show that the solution f^* to Kernel Ridge Regression (Eq. 2.10) results in an error [1, 45]

$$E_g(n) := \mathbb{E}_{(X, \mathbf{y})} \mathbb{E}_x (f(x) - g(x))^2 \quad (2.44)$$

$$= \frac{\kappa^2}{1 - \gamma} \sum_i \frac{|\hat{g}_i|^2}{(n\eta_i + \kappa)^2} + \sigma^2 \frac{\gamma}{1 - \gamma}, \quad (2.45)$$

where

$$\kappa = \lambda + \kappa \sum_i \frac{\eta_i}{n\eta_i + \kappa}, \quad \gamma = \sum_i \frac{n\eta_i^2}{(n\eta_i + \kappa)^2}. \quad (2.46)$$

Here, $\sigma^2 := \mathbb{E}_y (y - \mathbb{E}_y)^2$ is the variance of the noise in $\mathbf{y}$. This equation will be used later in this chapter to show that the generalization error of a particular quantum kernel improves when we apply the *kernel bandwidth* technique.

Despite the mathematical elegance of spectra for integral operators, for any given $k(x, x')$ (translation-invariant or otherwise) and fixed data distribution, it can be difficult to diagonalize the operator T_k . The literature includes just a few closed-form solutions for the spectra of T_k in some scenarios, for instance:

- Ref. [46], for $x \in \mathbb{R}$, $k(x, x') = e^{-b|x-y|^2}$, and $p_X(x) = e^{-2ax^2}$
- Ref. [47], for $x \in \mathbb{R}$, $k(x, x') = e^{|x-x'|}$, and $p_X(x) = \frac{1}{2}\mathbb{I}\{x \in [-1, 1]\}$.
- Ref. [48], for $x \in \mathbb{R}^d$, all $k(x, x') = f(\langle x, x' \rangle)$, and $p_X(x) = \text{Uniform}(S^d)$

Such closed-form solutions tend to be rare, and typically rely on unrealistic data distributions. This speaks to the practical challenge of applying spectral techniques towards the analysis of kernel methods.

2.4 Quantum kernel bandwidth and generalization

While there exist techniques for empirically evaluating the potential performance of quantum kernels on classical datasets [34] and examples of successful implementation on currently available quantum computers [6, 49–51], it is still unclear how to construct a quantum kernel that will be suitable for learning on a real-world classical dataset. In particular,

there are regimes in which generalization is apparently impossible with a subexponential amount of training data. These regimes arise from the same feature of quantum computing that originally motivated quantum kernel methods: the availability of exponentially large feature spaces.

This section introduces a technique to mitigate the harmful effects of using a high-dimensional quantum state space. We will begin by motivating this technique using two examples of high-dimensional quantum feature maps that result in flat spectra, thereby linking complete utilization of high-dimensional quantum feature space to poor generalization. We will then introduce the notion of kernel bandwidth, and show intuitively and analytically how this can potentially address poor generalization behavior associated with exponentially large quantum feature spaces.

2.4.1 Limitations of fixed quantum embeddings

In this section, we discuss simple examples where the high dimensionality of the feature space precludes learning with fixed quantum embeddings, in order to introduce the notion of kernel bandwidth and provide intuition as to how it may enable generalization. To this end, we present two extreme (and somewhat contrived) examples where a quantum feature map associated with random state vectors leads to a flat kernel spectrum.

2.4.1.1 Example 1: Uniformly random features

A toy classical example of how high dimensionality of the feature space precludes learning is motivated by uniformly random feature maps¹. Given a feature map ψ consisting entirely of independent and identically distributed Gaussian features (and therefore independent of x), the operator $\Sigma = \mathbb{E}_{\mathcal{X}}[\psi\psi^\dagger]$ is proportional to the identity operator. Since Σ shares eigenvalues with T_k of Eq. 2.9 (Section 2.3.1), the eigenvalues of K concentrate around unity at a rate $\mathcal{O}(n^{-1/2})$ [17].

Analogously, suppose there exist a data distribution and feature map $x \rightarrow \rho(x)$ for which unitaries $U(x)$ are distributed uniformly over the space of N -qubit unitaries $\mathcal{U}(2^N)$. Such a distribution may be achieved by sampling uniformly with respect to the Haar measure $\mu(dU)$ (e.g., [53]). As described in [17, Proposition 9], the spectrum of K concentrates

¹Not to be confused with random Fourier features [52], a powerful tool for speeding up the evaluation of classical kernel methods.

around the spectrum of Σ . We therefore explicitly compute the spectrum of Σ corresponding to random quantum features. Using standard results for integration with respect to the Haar measure [54, 55], we compute the average elements of Σ as

$$\begin{aligned}\langle ij|\Sigma|k\ell\rangle &= \int_{U(2^N)} \langle i|\rho(x)|k\rangle \langle j|\rho(x)|\ell\rangle \mu(dU) \\ &= \int_{U(2^N)} U_{i0}U_{k0}^*U_{j0}U_{\ell0}^* \mu(dU) \\ &= \frac{2^N - 1}{2^N(2^{2N} - 1)} (\delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}).\end{aligned}\tag{2.47}$$

The terms associated with $\delta_{ik}\delta_{j\ell} = 1$ correspond to the identity operator $\mathbb{I}_{4^N}$, while the terms $\delta_{i\ell}\delta_{jk}$ contribute to either 2^N -many diagonal components $\Sigma_{ii,ii}$ for $i = 1, \dots, 2^N$ or $2^N(2^N - 1)$ -many off-diagonal components $\Sigma_{ij,ji}$ whenever $j \neq i$:

$$\Sigma \rightarrow \mathbb{I}_{4^N} + \sum_{i=1}^{2^N} |ii\rangle\langle ii| + \sum_{i=1}^{2^N} \sum_{j \neq i} |ij\rangle\langle ji|.\tag{2.48}$$

Equation 2.48 is therefore a direct sum of subspaces proportional to $2\mathbb{I}_2$ and subspaces proportional to $\mathbb{I} + X$, the latter of which may be easily diagonalized. Combining with Eq. 2.47 yields the spectrum

$$\text{spec}(\Sigma) = \begin{cases} \frac{2}{2^N(2^N+1)} & \text{with multiplicity } 2^N + \frac{2^N(2^N-1)}{2} \\ 0 & \text{with multiplicity } \frac{2^N(2^N-1)}{2}. \end{cases}$$

Coincidentally, the *mean embedding* $\rho_\mu = \int_{\mathcal{X}} \rho(x) \mu(dx)$ associated with this feature map is proportional to the identity; that is, $\rho_\mu = \mathbb{E}_{\mathcal{X}}[\rho(x)] = \mathbb{I}/2^N$. This computation yields $2^{N-1}(2^N + 1)$ nonzero eigenvalues with magnitude $2^{1-N}(2^N + 1)^{-1}$, and thus the nonzero eigenvalues of K become uniform as $n \rightarrow \infty$. Since the largest eigenvalue is exponentially small in N , generalization requires the number of data points to be exponentially large in N . The connection between the magnitude of the largest eigenvalue and the required number of training samples can be seen directly from [18, Theorem 3], although it is a straightforward consequence of many older results, for example, [44, 45, 56–58]. In other words, efficient generalization becomes impossible when the quantum feature map uses the full extent of the quantum state space uniformly. Later we will demonstrate that the converse is true: restricting embeddings to a smaller region of quantum state space recovers the possibility of efficient learning.

2.4.1.2 Example 2: Product embeddings

We now provide a second example of a quantum embedding that fails to generalize on data. This learning problem and the failure of fixed quantum embeddings on it were considered in [34]. For an input $x \in \{0, \pi\}^D$, the goal is to learn a target function $f(x) = \cos(x^{(D)})$, where $x^{(i)}$ is the i -th element of the vector x . Since learning this function is equivalent to learning the value of the last element of x , it is trivial classically, and a simple linear regression succeeds. We now show how KRR with a badly designed quantum kernel fails on this trivial task, and we show how the introduction of bandwidth allows the KRR with a quantum kernel to learn the target. Consider a quantum kernel equipped with feature map

$$U(x) = \bigotimes_{j=1}^D U(x^{(j)}), \quad U(x^{(j)}) = R_x(x^{(j)}) = \begin{pmatrix} \cos(x^{(j)}/2) & i \sin(x^{(j)}/2) \\ i \sin(x^{(j)}/2) & \cos(x^{(j)}/2) \end{pmatrix}, \quad (2.49)$$

where $\bigotimes_{j=1}^D$ is the tensor product of single-qubit operations and $R_x(\theta)$ represents a rotation gate by θ about the x -axis of the single-qubit Bloch sphere (see Fig. 2.1A). For this feature map, the embedding factorizes over qubits as $\rho(x) = U(x)|0\rangle\langle 0|U^\dagger(x) = \bigotimes_{j=1}^D \rho(x^{(j)})$, and

$$\rho(x^{(j)}) = \begin{pmatrix} \cos^2(x^{(j)}/2) & -i \cos(x^{(j)}/2) \sin(x^{(j)}/2) \\ i \cos(x^{(j)}/2) \sin(x^{(j)}/2) & \sin^2(x^{(j)}/2) \end{pmatrix}, \quad (2.50)$$

with the kernel given by

$$k(x, x') = \text{Tr}(\rho(x)\rho(x')) = \prod_{j=1}^D \cos^2((x^{(j)} - x'^{(j)})/2). \quad (2.51)$$

While this kernel may be estimated using a quantum system, the fact that it has a tensor product closed form expression makes it classically easy to simulate. Nevertheless, it has been a useful toy model [18, 34] for the analytical analysis of exponentially large quantum feature spaces, and so we will continue to use it here. Since the input data is $x \in \{0, \pi\}^D$, the kernel becomes a delta function: $k(x, x') = \delta_{x, x'}$. Therefore, any two points in the feature space $\rho(x)$ and $\rho(x')$ for $x \neq x'$ are orthogonal and the kernel cannot capture the correlations in data. KRR with this kernel simply memorizes the training set and cannot generalize to any target function with a subexponential (in D) training set size.

Then under the following assumptions on the input data the kernel matrix associated with $\rho(x)$ becomes diagonal: Ref. [34, Appendix I] considers data that are drawn from $x \in \{0, \pi\}^D$ while Ref. [18] considers data drawn from $\text{Uniform}(0, \pi)^{\otimes D}$. In the former case,

since $\rho(x) = xx^T$ one can immediately compute $k(x_i, x_j) = |x_i^T x_j|^2 = \delta_i^j$, demonstrating that the kernel matrix is diagonal and full rank. In the latter case, it can be shown that the largest eigenvalue of the integral kernel operator vanishes exponentially in D ; normalization conditions then impose an exponential number of non-negligible eigenvalues.

In both of the above examples, the kernel exhibits a flat spectrum which precludes generalization on any training set whose size is subexponential in the number of qubits. More intuitively, these failures arise because the quantum kernel completely utilizes the high dimension of the feature space, and consequently lose the ability to distinguish the images of input data. This behavior is not surprising in light of classical notions of a curse of dimensionality [59] for high-dimensional feature spaces. In classical machine learning, one technique for successfully learning in high-dimensional spaces is bandwidth, which allows for optimal performance in nearest neighbor settings and universal classification with radial basis function (RBF) kernels. Analogously, the remainder of this work will be devoted to showing how use of bandwidth in quantum kernels can lead to significant performance gains despite the use of an exponentially growing quantum state space, and recovers the possibility of using large-scale QML algorithms to successfully classify data.

2.4.2 Bandwidth enables learning

The preceding examples highlighted how quantum kernel methods utilizing high dimensional spaces can fail to generalize. Our central technique for mitigating this limitation will be to introduce bandwidth to quantum kernels [6, 60].

2.4.2.1 Recovering a usable spectrum

We will begin with an intuitive explanation of bandwidth (and its effects) for a single qubit. As shown in [18, Lemma 1], the inequality $\eta_{\max}(\Sigma) \leq (\text{Tr}(\rho_\mu^2))$ provides a bound for the largest eigenvalue of Σ . Let $\mathbf{n}$ be the Bloch vector for ρ_μ defined on $n = 1$ qubits, which can be parameterized as

$$\mathbf{n} = (\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta) \quad (2.52)$$

Then $\text{Tr}(\rho_\mu^2) = (1 + \|\mathbf{n}\|^2)/2$. This provides a geometric argument for the use of bandwidth in a single qubit system (or product thereof). For instance, consider each shaded region in Fig. 2.1A representing the extreme points of the subspace of $D(\mathbb{C}^2)$ spanned by quantum states associated with the feature map giving rise to some k . Then we have that the maximum eigenvalue of Σ is suppressed like $\eta_{\max}(\Sigma) \leq (1 + \|\mathbf{n}\|^2)^{1/2}/\sqrt{2}$ as the centroid

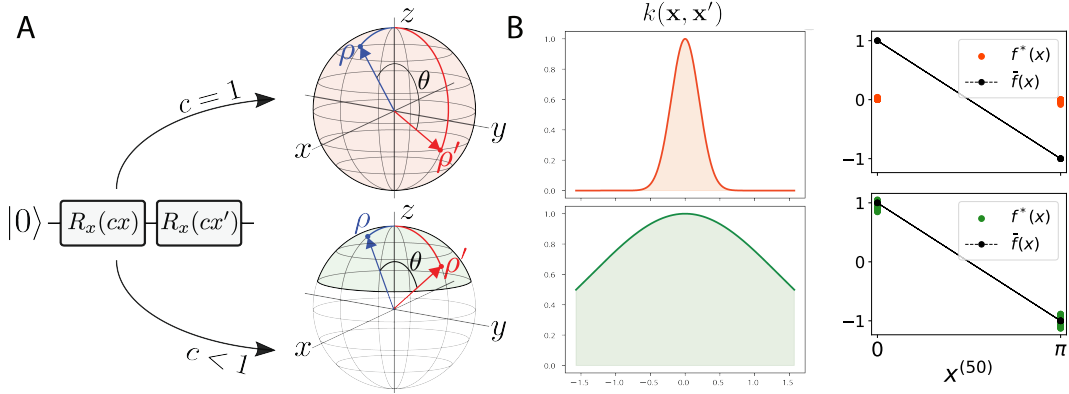


Figure 2.1: **A** A quantum kernel method involves embedding data using a quantum circuit, often involving rotation by some *angle* about an axis. When angles are not rescaled properly, data can be embedded far apart in a space with dimensionality $O(2^N)$ (Sec. 2.4). Similarly, $\lambda_{\max}$ is suppressed as the mean embedding $\mathbb{E}_{\mathcal{X}}[\rho(x)]$ approaches the center of the Bloch sphere [18]. **B** High-dimensional feature space results in $k(x, x')$ with narrow width (Eq. 2.53 with $n = 50$ and (top) $c = 1$ (bottom) $c = 0.25$) Tuning bandwidth escapes the “curse of dimensionality” associated with high-dimensional feature space. For $n = 50$ and $\tilde{f}(x) = \cos(x^{(50)})$, quantum features that are nearly orthogonal result in a *narrow* kernel ($c = 1$, top) and failure to generalize, while tuning bandwidth ($c < 1$, bottom) recovers KRR generalization performance.

$\mathbf{n}$ of each region (representing ρ_{μ}) approaches the center of the Bloch sphere. By limiting bandwidth (and therefore restricting the shaded region of Fig. 2.1B to a polar cap of the Bloch sphere), the upper bound on $\eta_{\max}(\Sigma)$ is lifted, and the possibility of learning on data is restored.

We now reconsider the kernel with the feature map of Eq. 2.49 and introduce a scaling parameter $c \in [0, 1]$ that controls the bandwidth of the kernel. The feature map and the kernel become

$$U(x^{(j)}) = R_x(cx^{(j)}), \quad k(x, x') = \prod_{j=1}^N \cos^2(c(x^{(j)} - x'^{(j)})/2). \quad (2.53)$$

Geometrically, the factor c restricts features $\rho(x)$ to a smaller region of the Bloch sphere (Fig. 2.1A). Consequently, the kernel matrix is no longer diagonal, and we can straightforwardly check that simply tuning c allows KRR with kernel Eq. 2.53 to generalize (see Fig. 2.1B).

2.4.2.2 Explicit formulas for generalization of bandwidth-equipped kernels

We will now discuss how the bandwidth introduced into the feature map given by Eq. 2.53 for the input distribution $x \sim \text{Unif}([- \pi, \pi]^N)$ enables generalization. Ref. [1] analytically derives the spectrum of this kernel for $c \in [0, 1]$ using $N = 1$ qubits, and then uses techniques from Ref. [45] to compute generalization error analytically in the large- n limit. For brevity, we will only present the results of this analysis, and refer the interested reader to Ref. [1]. For $N = 1$ qubits, the kernel becomes $k(x, x') = \cos^2(c(x - x')/2)$ with bandwidth parameter c and the eigenvalues are:

$$\begin{aligned} \lambda_1 &= \frac{3}{8} + \frac{1}{8} \text{sinc}(2\pi c) + \frac{1}{8} \sqrt{(1 - \text{sinc}(2\pi c))^2 + 16 \text{sinc}(\pi c)^2}, & \lambda_2 &= \frac{1}{4} - \frac{1}{4} \text{sinc}(2\pi c), \\ \lambda_3 &= \frac{3}{8} + \frac{1}{8} \text{sinc}(2\pi c) - \frac{1}{8} \sqrt{(1 - \text{sinc}(2\pi c))^2 + 16 \text{sinc}(\pi c)^2}, & \lambda_4 &= 0. \end{aligned} \quad (2.54)$$

For an N -qubit system, the largest eigenvalue $\eta_{\max}$ of the kernel in Eq. 2.53 falls exponentially with N for $c \sim \mathcal{O}(1)$, which makes generalization impossible with $n \sim \text{poly}(N)$ amount of data [18]. To prevent $\eta_{\max}$ from decreasing with N , we can choose a bandwidth that scales with N (i.e., $c = aN^{-\alpha}$ with $a \sim \mathcal{O}(1)$ and $\alpha > 0$). This choice of bandwidth enables sufficient decay in the spectrum of the kernel for learning to occur. Fig. 2.2 shows the results of simulating the kernel of Eq. 2.53 with $N = 40$ qubits for a target function given by $\bar{f}(x) = e^{-\|x\|^2/N^2}$. For $c = 1$, the kernel has a flat spectrum (Fig. 2.2A) with poor alignment with the task (Fig. 2.2B). On the other hand, bandwidth introduces sufficient decay in the eigenspectrum so that learning with a polynomial amount of data (in N) becomes possible. Fig. 2.2C shows kernel regression with this toy kernel, and uses Eq. 2.44 to confirm that generalization improves with bandwidth up to an optimal value, after which it degrades again. This is due to the fact that larger bandwidths cause underfitting of the data [61] since only a very few eigenfunctions become learnable while the target cannot be fully explained by those modes.

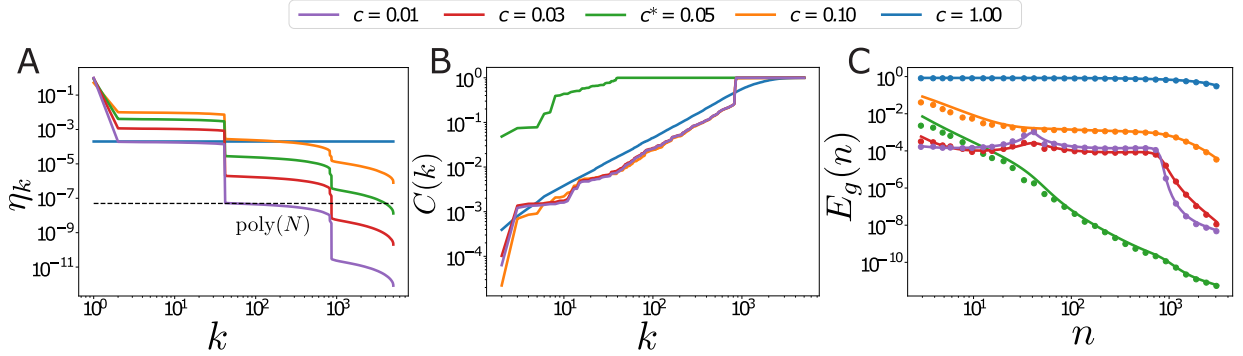


Figure 2.2: **A** Eigenvalues of the kernel Eq. 2.53 with respect to data $x \sim \text{Unif}([-\pi, \pi]^{40})$. Learnability is explained by the preservation of large-eigenvalue eigenspaces; without bandwidth, the model provably cannot learn in $\text{poly}(N)$ complexity due to the flat spectrum. **B** The projections of the target $\bar{f}(x) = e^{-\|x\|^2/N^2}$ on the eigenvectors of the kernel for each bandwidth. Apart from the flatness of the $c = 1$ kernel, its eigenfunctions align poorly with the target. **C** Generalization error as a function of the number of training samples computed by using theory (solid lines) and performing kernel ridge regression empirically (dots). Bandwidth $c = 1$ yields a constant learning curve. While all $c < 1$ kernels provide improvement, an optimal bandwidth parameter $c^* \approx 2/N$ gives the best task-model alignment.

2.4.2.3 Lifting a no-go result for quantum kernel learning

We further show how the concept of bandwidth circumvents prior no-go results for quantum kernel methods. Specifically, Ref. [18, Theorem 1] states that, with probability at least $1 - \epsilon - \eta_{\max} P^4$, the error for the solution f_m^λ of the KRR optimization of Eq. 2.10 with penalty λ and n training data is lower bounded by

$$R_{\text{emp}}(f_m^\lambda) \geq \left(1 - \sqrt{\frac{2\eta_{\max} n^2}{\epsilon}}\right) \|f\|_2. \quad (2.55)$$

Here, $\eta_{\max}$ is the largest eigenvalue of T_k and, consequently, the error of this model grows as $\eta_{\max} \rightarrow 0$. While it is certainly untrue that bandwidth can guarantee low error for a KRR optimization, we can show how the bound of Eq. 2.55 becomes vacuous by introducing a bandwidth to k . Specifically, we suppose there is a choice of bandwidth such that the magnitude of $k(x, x')$ is lower bounded by some bandwidth-dependent constant. We therefore demonstrate how an upper bound on the ridge regression classifier accuracy based

on the largest eigenvalue $\eta_{\max}$ no longer guarantees failure of any particular quantum kernel method. We assume that we can use bandwidth to require that states are “not too far away” in Hilbert space. More specifically, we suppose that there is some function Δ_c such that

$$k(x, x') \geq \Delta_c. \quad (2.56)$$

Note that this does not interfere with the requirement that k is L_2^μ integrable since we assume that k is defined on a restricted support $x \in [-\pi, \pi]^D$ (without this assumption, k is not integrable in a more general treatment). We now show that some choice of Δ_c always ensures that the largest eigenvalue of K (and therefore the largest eigenvalue of T_k with high probability) is bounded. Denote the eigenvalues of K as η_k , and in particular let $\eta_{\max}$ be the largest eigenvalue of K and $Ku = \eta_{\max}u$ for some eigenvector $u \in \mathbb{R}^n$. By definition, $\eta_{\max} = \max_{\|v\|=1} \langle v, Kv \rangle$ for all $v \in \mathbb{R}^n$, and in particular

$$\eta_{\max} = \frac{\langle u, Ku \rangle}{\langle u, u \rangle} \geq \frac{\langle \mathbf{1}, K\mathbf{1} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} = \frac{1}{n} \sum_{i,j=1}^n K_{i,j} \geq (n-1)\Delta_c,$$

where $\mathbf{1}$ is the vector of all ones, and we have applied the inequality of Eq. 2.56. Ref. [17, Proposition 10] then states that

$$\Pr \left(\sup_k \left| \frac{\eta_k}{n} - \gamma_k \right| \leq 2\sqrt{\frac{\log(4/\delta^2)}{n}} \right) \geq 1 - \delta,$$

where we contrast the empirical eigenvalues η_k with the eigenvalues γ_k of the integral operator T_k . Note that the empirical eigenvalues $\eta_k \rightarrow \gamma_k$ as $n \rightarrow \infty$. Hence, with probability at least $1 - \delta$ it holds that

$$\left| \eta_{\max} - \left(\frac{n-1}{n} \right) \Delta_c \right| \leq 2\sqrt{\frac{\log(4/\delta^2)}{n}}. \quad (2.57)$$

From Eq. 2.57, however, we see that $\eta_{\max}$ approaches a constant Δ_c at a rate of $O(1/\sqrt{n})$; for sufficiently large n , Eq. 2.55 holds with vanishing probability, and for all other choices of n the bound becomes vacuous with high probability under the condition that

$$\sqrt{2\eta_{\max}n^2/\epsilon} \geq 1, \quad (2.58)$$

which may be achieved within $O(1/\sqrt{n})$ precision by choosing

$$\Delta_c \geq \sqrt{\frac{2\epsilon}{n^2}}.$$

Importantly, this outcome does *not* guarantee that KRR using k satisfying the inequality in Eq. 2.56 will achieve good generalization error. In particular, the choice $\Delta_c = 1$ will result in T_k having a single nonzero eigenvalue associated with the constant function. Rather, this demonstration reemphasizes that there are two conditions to successful classification using quantum kernels: (i) the kernel should be chosen such that the eigenfunctions of T_k align with the target function, and (ii) the corresponding eigenvalues should be large. By lower bounding k in this way, we can guarantee condition (ii); however, successful generalization will still depend on a choice of k satisfying condition (i). Indeed, follow-up work has suggested that kernels equipped with a bandwidth may start to lose any presupposed advantage over classical learning algorithms, perhaps for the very same reason that they restrict the size and/or complexity of the quantum state space that is exploited to compute the quantum kernel [62]. Future work may shed light on the interplay between the advantage offered by a kernel and the spectral problems induced by using a high-dimensional feature space.

Chapter 3

Generalization despite overfitting in quantum machine learning models

3.1 Chapter overview

Chapter 2 introduced a relationship between the spectrum of a kernel operator to the success of a learning algorithm, as well as the concept of overfitting and how it tends to harm generalization. This chapter will analyze a scenario known as *benign overfitting* in which certain learning models trained using a quantum computer may, counterintuitively, successfully generalize on unseen data after having completely overfit the given training data. First, we describe a classical Fourier framework for the study of interpolating models and develop explicit formulae for the error in these models to produce a basic example of benign overfitting (Sec. 3.3). We will then construct a quantum model with analogous components to the classical model, and demonstrate how each of these components is related to the structure of the corresponding quantum circuit and measurement (Sec. 3.4). We then analyze specific cases that give rise to “spikiness” that is related to benign overfitting in these quantum models (Sec. 3.4.2). Appendix A contains derivations and proofs for these results.

This chapter (and Appendix A) is reproduced directly from Ref. [2], except for Sec. 3.4.3 which contains unpublished work. Maria Schuld contributed heavily to the introduction (Sec. 3.2), and helped organize the presentation of the work. I was responsible for developing the theory and numerics necessary to execute this idea.

3.2 Introduction

A long-standing paradigm in machine learning is the trade-off between the complexity of a model family and the model’s ability to generalize: more expressive model classes contain better candidates to fit complex trends in data, but are also prone to overfitting noise [63, 64]. *Interpolation*, defined for our purposes as choosing a model with zero training error, was hence long considered bad practice [65]. The success of deep learning – machine learning in a specific regime of extremely complex model families with vast amounts of tunable parameters – seems to contradict this notion; here, consistent evidence shows that among some interpolating models, more complexity tends *not to harm* the generalisation performance¹, a phenomenon described as “benign overfitting” [66].

In recent years, a surge of theoretical studies have reproduced benign overfitting in simplified settings with the hope of isolating the essential ingredients of the phenomenon [66, 67]. For example, Ref. [68] showed how interpolating linear models in a high complexity regime (more dimensions than datapoints) could generalize just as well as their lower-complexity counterparts on new data, and analyzed the properties of the data that lead to the “absorption” of noise by the interpolating model without harming the model’s predictions. Ref. [69] showed that there are model classes of simple functions that change quickly in the vicinity of the noisy training data, but recover a smooth trend elsewhere in data space (see Figure 3.1). Such functions have also been used to train nearest neighbor models that perfectly overfit training data while generalizing well, thereby directly linking “spiking models” to benign overfitting [70]. Recent works try to recover the basic mechanism of such spiking models using the language of Fourier analysis [71–73].

In parallel to these exciting developments in the theory of deep learning, quantum computing researchers have proposed families of parametrised quantum algorithms as model classes for machine learning (e.g. Ref. [74]). These quantum models can be optimised similarly to neural networks [75, 76] and have interesting parallels to kernel methods [11, 12] and generative models [77, 78]. Although researchers have taken some first steps to study the expressivity [51, 79–81], trainability [82, 83] and generalisation [34, 84–86] of quantum models, we still know relatively little about their behaviour. In particular, the interplay of overparametrisation, interpolation, and generalisation that seems so important for deep learning is yet largely unexplored.

In this paper we develop a simplified framework in which questions of overfitting in quantum machine learning can be investigated. Essentially, we exploit the observation

¹The fact that interpolation does not always harm generalization performance is in fact well-known for models like boosting and nearest neighbours, which are both interpolating models.

that quantum models can often be described in terms of Fourier series where well-defined components of the quantum circuit influence the selection of Fourier modes and their respective Fourier coefficients [42, 87, 88]. We link this description to the analysis of spiking models and benign overfitting by building on prior works analyzing these phenomena using Fourier methods. In this approach, the complexity of a model is related to the number of Fourier modes that its Fourier series representation consists of, and overparametrised model classes have more modes than needed to interpolate the training data (i.e., to have zero training error). After deriving the generalization error for such model classes these “superfluous” modes lead to spiking models, which have large oscillations around the training data while keeping a smooth trend everywhere else. However, large numbers of modes can also harm the recovery of an underlying signal, and we therefore balance this trade-off to produce an explicit example of benign overfitting in a quantum machine learning model.

The mathematical link described above allows us to probe the impact of important design choices for a simplified class of quantum models on this trade-off. For example, we find why a measure of redundancy in the spectrum of the Hamiltonian that defines standard data encoding strategies strongly influences this balance; in fact to an extent that is difficult to counterbalance by other design choices of the circuit.

3.3 Interpolating models in the Fourier framework

In this section we will provide the essential tools to probe the phenomenon of overparametrized models that exhibit the spiking behaviour from Figure 3.1 using the language of Fourier series. We will review and formalize the problem setting and several examples from Refs. [71–73], before extending their framework by incorporating standard results from linear regression to derive closed-form error behavior and examples of benign overfitting.

3.3.1 Setting up the learning problem

We are interested in functions g defined on a finite interval that may be written in terms of a linear combination of Fourier basis functions or *modes* $e^{i2\pi kx}$ (each describing a complex sinusoid with integer-valued frequency k) weighted by their corresponding Fourier coefficients $\hat{g}_k$:

$$g(x) = \sum_{k=-\infty}^{\infty} \hat{g}_k e^{i2\pi kx}. \quad (3.1)$$

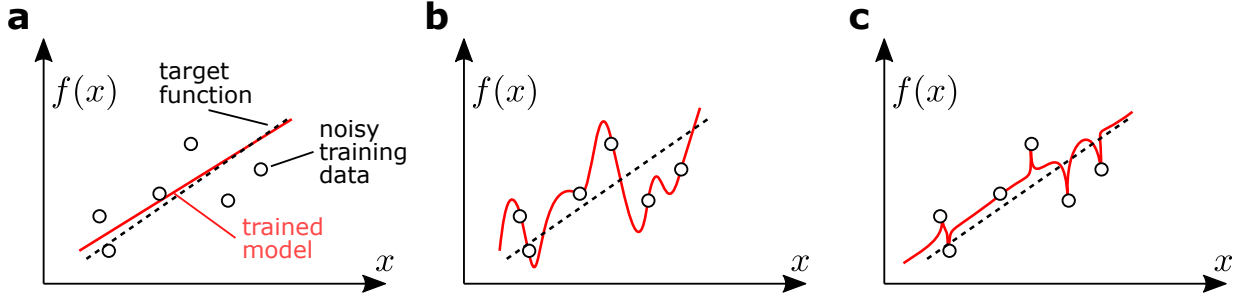


Figure 3.1: Intuition behind the phenomenon of benign overfitting with spiking models. **a** Training a model typically involves a trade-off between fitting noisy training data well and recovering an underlying target function **b** Traditional learning theory associates interpolating models (reaching zero training error by fitting every data point) with low generalization capability (or high test error) by failing to recover a simple target function. **c** “spiking models” that change quickly in the vicinity of training data but otherwise exhibit simple behavior can explain how both kinds of errors may be kept low.

We restrict our attention to well-behaved functions that are sufficiently smooth and continuous to be expressed in this form. We will now set up a simple learning problem whose basic components – the model and target function to be learned – can be expressed as Fourier series with only few non-zero coefficients, and define concepts such as over-parametrization and interpolation.

Consider a machine learning problem in which data is generated by a target function of the form

$$g(x) = \sum_{k \in \Omega_{n_0}} \hat{g}_k e^{i2\pi kx} \quad (3.2)$$

which only contains frequencies in the discrete, integer-valued spectrum

$$\Omega_{n_0} = \left\{ -\frac{n_0 - 1}{2}, -\frac{n_0 - 1}{2} + 1, \dots, 0, \dots, \frac{n_0 - 1}{2} \right\} \quad (3.3)$$

for some odd integer n_0 . We call functions of this form *band-limited* with bandwidth $n_0/2$ (that is, $|k| < n_0/2$ for all frequencies $k \in \Omega_{n_0}$). The bandwidth limits the complexity of a function, and will therefore be important for exploring scenarios where a complex model is used to overfit a less complex target function. We are provided with n noisy samples $y_j = g(x_j) + \epsilon$ of the target function g evaluated at points $x_j = j/n$ spaced uniformly on interval $[0, 1]$ (we will assume input data has been rescaled to $[0, 1]$ without loss of generality), where $n > n_0$ and we require n to be odd for simplicity.

The model class we consider likewise contains band-limited Fourier series, and since we are interested in interpolating models, we always assume that they have enough modes to interpolate the noisy training data, namely:

$$f(x) = \sum_{k \in \Omega_d} \alpha_k \sqrt{\nu_k} e^{i2\pi kx}, \quad (3.4)$$

such that $f(x_j) = y_j$ for all $j \in [n]$.

Similarly to Eq. (3.3) we define the spectrum

$$\Omega_d = \left\{ -\frac{d-1}{2}, -\frac{d-1}{2} + 1, \dots, 0, \dots, \frac{d-1}{2} \right\}. \quad (3.5)$$

Following Ref. [71], this model class has two components: The set of weighted Fourier basis functions $\sqrt{\nu_k} e^{i2\pi kx}$ describe a feature map applied to x for some set of feature weights $\nu_k \in \mathbb{R}^+$, while the trainable weights $\alpha_k \in \mathbb{C}$ are optimized to ensure that f interpolates the data. The theory of trigonometric polynomial interpolation [89] ensures that f can always interpolate the training data for some choice of trainable weights α_k under these conditions. In the following, we will therefore usually consider the α_k as being determined by the data and interpolation condition, while the ν_k serve as our “turning knobs” to create different settings in which to study spiking properties and benign overfitting. We call the model class described in Eq. (3.4) *overparameterized* when the degree of the Fourier series is much larger than the degree of the target function, $d \gg n_0$, in which case the model has many more frequencies available to perform fitting than there are present in the underlying signal to fit.

To complete the problem setup, we have to impose one more constraint. Consider that $\exp(i2\pi kx_j)$ with frequency $k < n/2$ and uniformly spaced points x_j is equal to $\exp(i2\pi k'x_j)$ for any choice of *alias frequency* $k' = k \bmod n$. The presence of these aliases means that the model class described in Eq. (3.7) contains many interpolating solutions in the overparameterized regime. Motivated by prior work exploring benign overfitting for linear features [68], Fourier features [71, 90], and other nonlinear features [44, 91], we will study the minimum- ℓ_2 norm interpolating solution,

$$\boldsymbol{\alpha}^{opt} = \arg \min_{\boldsymbol{\alpha}: f(x_j)=y_j \forall j} \|\boldsymbol{\alpha}\|_2. \quad (3.6)$$

Minimizing the ℓ_2 norm is a typical choice for imposing a penalty on complex functions (regularization) in the underparameterized regime, though we will see that this intuition does not carry over to the overparameterized regime. The remainder of this section will

explore how this learning problem results in a trade-off in interpolating Fourier models: Overparameterization introduces alias frequencies that increase the error in fitting simple target functions but can also reduce error by absorbing noise into high-frequency modes with spiking behavior.

Note that one can rewrite the Fourier series of Eq. (3.4) in a linear form

$$f(x) = \langle \boldsymbol{\alpha}, \phi(x) \rangle \quad (3.7)$$

where

$$\boldsymbol{\alpha} = (\alpha_{-(d-1)/2}, \dots, \alpha_k, \dots, \alpha_{(d-1)/2}) \in \mathbb{C}^d, \quad (3.8)$$

$$\phi(x) = (\sqrt{\nu_{-(d-1)/2}} e^{i\pi(d-1)x}, \dots, \sqrt{\nu_k} e^{i2\pi kx}, \dots, \sqrt{\nu_{(d-1)/2}} e^{i\pi(d-1)x}) \in \mathbb{C}^d. \quad (3.9)$$

From this perspective, optimizing f amounts to learning trainable weights $\boldsymbol{\alpha}$ by performing regression on observations $\phi(x)$ sampled from a *random Fourier features* model [52] for which ν_k and $\phi(x)_k$ are precisely the eigenvalues and eigenfunctions of a translation-invariant kernel [92]. This provides a direct link between this work and the kernel spectra introduced in Chapter 2. Specifically, for the above choice of data distribution $x_j = j/n$ occuring with probability $1/n$, we choose a kernel of the form

$$\kappa(x, x') = \sum_k \nu_k e^{-i2\pi k(x-x')} = \sum_k \nu_k \psi_k(x) \psi_k^*(x'), \quad (3.10)$$

for $\psi_k(x) := e^{-i2\pi kx}$. This results in an integral kernel operator which acts on a function h according to

$$(T_\kappa h)(x) = \sum_k \nu_k e^{i2\pi kx} \frac{1}{n} \sum_{j=0}^{n-1} h\left(\frac{j}{n}\right) e^{-i2\pi kj} := \sum_k \nu_k \hat{h}_k \psi_k(x), \quad (3.11)$$

where $\hat{h}_k$ are the Fourier coefficients of h . This means that both the model f and the target function g are already provided in the eigenbasis $\{\exp(i2\pi kx)\}_k$ of T_κ , and the remainder of this work may be understood as choosing a set of coefficients α_k for $f(x)$ that correctly *aligns* f with the target function g . In a more standard setting (not benign overfitting), an optimal model would learn α such that the Fourier coefficients of f match the the Fourier coefficients $\hat{g}_k$. However, we will see that the additional constraints above of perfect interpolation and minimum- ℓ_2 solution will result in a rather different form of alignment.

3.3.2 Two extreme cases to understand generalization

To better understand the trade-off that overparametrization – or in our case, a much larger number of Fourier modes than needed to interpolate the data – introduces between fitting noise and generalization error, we revisit two extreme cases explored in Ref. [71], involving a pure-noise signal and a noiseless signal.

Case 1: Noise only

The first case demonstrates how alias modes can help to fit noise without disturbing the (here trivial) signal. We set $g = 0$ and consider n observations $y_j = \epsilon_j$ of zero-mean noise with known variance $\mathbb{E}[\epsilon^2] = \sigma^2$. After making n uniformly spaced observations, we compute the discrete Fourier transform of the observations as the sequence of values $\hat{\epsilon}_j$ satisfying

$$\hat{\epsilon}_j = \frac{1}{n} \sum_{k \in \Omega_n} \epsilon_k e^{-i2\pi jk/n}, \quad (3.12)$$

which characterizes the frequency content of the noisy signal that can be captured and learned from using only n evenly spaced samples. Suppose that the degree of the model (controlling the model complexity) is given by $d = n(m + 1)$ for some even integer m and that $\nu_k = 1$ for every mode, so that there are exactly m equally-weighted aliases for each frequency in the spectrum of the Fourier series for g . Then the optimal (i.e., the minimum ℓ_2 -norm, interpolating) trainable weight vector $\boldsymbol{\alpha}^{opt}$ has entries

$$\alpha_{k+n\ell}^{opt} = \frac{\hat{\epsilon}_k}{n(m + 1)} \quad (3.13)$$

for $\ell = -m/2, \dots, m/2$, with all other entries being zero (see Appendix A.1.2). From Eq. (3.13), the minimum- $\|\boldsymbol{\alpha}\|_2$ solution distributes noise Fourier coefficients $\hat{\epsilon}_k$ evenly into many alias frequencies $k + n\ell$, while enforcing that the sum of trainable weights $\alpha_{k+n\ell}$ for all of these aliases is $\hat{\epsilon}_k$ to guarantee interpolation. As shown in Figure 3.2, the higher-frequency aliases suppress the optimal model $f^{opt}(x) = \langle \boldsymbol{\alpha}^{opt}, \phi(x) \rangle$ to near-zero at every point away from the interpolated points, resulting in a test error of $\mathcal{O}(\sigma^2/m)$ that decreases monotonically with the complexity of the model. As m increases, the optimal model f^{opt} remains close to the true signal $y = 0$ while becoming “spiky” near the noisy samples. By conforming to the true signal everywhere except in the vicinity of noise, this behavior embodies the mechanism of how overparameterized models can absorb noise into high frequency modes. In this case the generalization error, measuring how close the model is to the target function on average, *decreases* with increasing complexity of the model class.

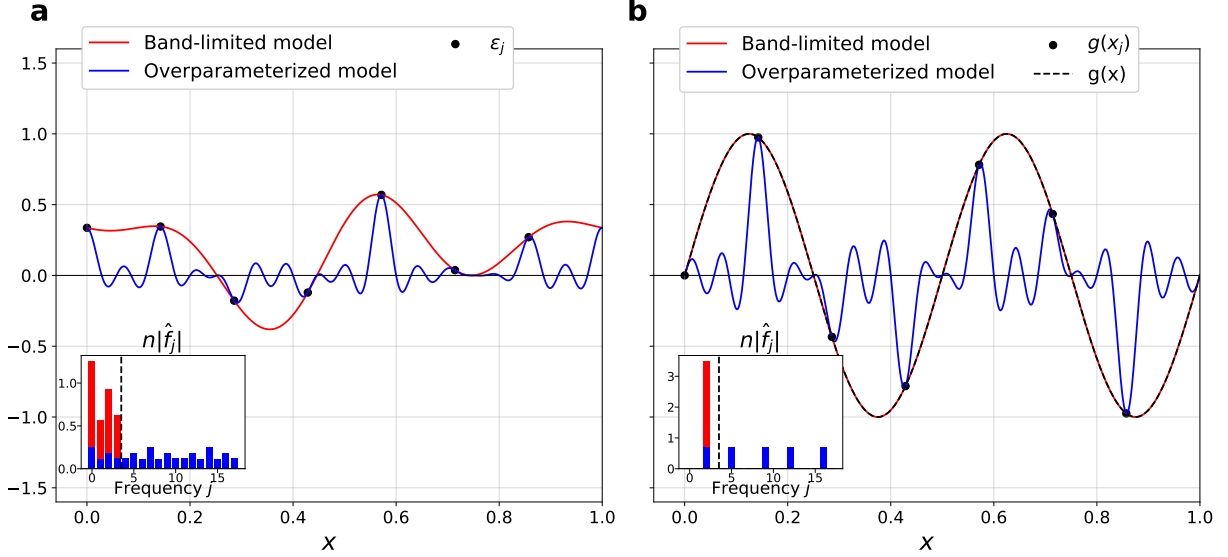


Figure 3.2: Comparison of overparametrized and simple models that interpolate data from different target functions. **a** *Noise only*: The overparameterized model ($n = 7$, $d = 35$, blue) is more effective at recovering the target function $g(x) = 0$ when provided with noisy data, while a band-limited model (red) cannot do so effectively. Inset: The distribution of trainable weights for the minimum $\|\alpha\|_2$ model are distributed among many aliases suppressing the error of the optimized model (proportional to $\sum_j |\hat{f}_j|^2$) in this case. **b** *Signal only*: The opposite occurs in the case of noiseless input, for which a band-limited model ($n_0 = 5$) perfectly recovers $g(x) = \sin(2\pi(2x))$ while the overparameterized model fails to capture the behavior of the input signal. Inset: Minimizing the ℓ_2 norm of α distributes Fourier coefficients $\hat{g}_2$ among aliases $\hat{f}_{2+\ell n}$ (and $\hat{g}_{-2}$ among aliases $\hat{f}_{-2+\ell n}$); this bleeding of the signal g into higher frequencies results in higher error in the overparameterized model [71, 73].

Case 2: Signal only

While the above case shows how overparametrization can help to absorb noise to reduce error without harming the signal, the second case will illustrate how alias frequencies in the overparameterized model can harm the model’s ability to learn the target function. To demonstrate this, we now consider a noiseless, single-mode signal $g(x) := \hat{g}_p e^{i2\pi p x}$ of frequency $p \leq n_0/2$. The data is hence of the form

$$y_j = g(x_j) := \hat{g}_p e^{i2\pi p x_j}. \quad (3.14)$$

Once again we choose $d = n(m+1)$ and for simplicity we assume an unweighted model, $\nu_k = 1$ for $k \in \Omega_d$. By orthonormality of Fourier basis functions, the interpolation condition requires that only the modes of the model f with integer multiples of the frequency p are retained. The interpolation constraint can then be rewritten as

$$\sum_{\ell=-m/2}^{m/2} \alpha_{p+n\ell} = \hat{g}_p. \quad (3.15)$$

The choice of trainable weights α_k that satisfy Eq. (3.15) while minimizing ℓ_2 -norm is

$$\alpha_{p+n\ell}^{opt} = \frac{\hat{g}_p}{m+1} \quad (3.16)$$

for $k = p + n\ell$ and $\alpha_k = 0$ otherwise (see Appendix A.1.2). Eq. (3.15) distributes the Fourier coefficient $\hat{g}_p$ among the trainable weights $\alpha_{p+n\ell}$ corresponding to frequencies $p+n\ell$. Therefore, minimizing $\|\alpha\|_2$ in this case “bleeds” the target function into higher frequency aliases and results in the opposite effect compared to fitting a noisy signal (see Fig. 3.2b): The generalization error of the overparameterized model now *increases* with the number of aliases m and the complexity of the model harms its ability to fit a noiseless target function.

In order to recover a trade-off in generalization error for more general cases, we will need to consider more interesting distributions of feature weights ν_k (instead of $\nu_k = 1$) that provide finer control over fitting the target function with low-frequency modes while spiking in the vicinity of noise with high-frequency aliases.

3.3.3 Generalization trade-offs and benign overfitting

The opposing effects of higher-frequency modes in overparameterized models in the cases discussed above hint at a trade-off in model performance that depends on the underlying

signal and the feature weights of the Fourier feature map. Returning to the more general case of input samples $y_j = g(x_j) + \epsilon_j$, in Appendix A.1 we show that the task of fitting uniformly spaced samples using weighted Fourier features may be transformed into a linear regression problem, thereby generalizing the results of [71] to derive the following general solution to the minimum- ℓ_2 interpolating problem of Eq. (3.6):

$$\alpha_{k+\ell n}^{opt} = \hat{y}_k \frac{\sqrt{\nu_{k+\ell n}}}{\sum_{j \in S(k)} \nu_j}, \quad (3.17)$$

where $\hat{y}_k$ is the discrete Fourier transform of y_k and $k \in \Omega_{n_0}$, where

$$S(k) = \{\ell : k + n\ell \in \Omega_d\} \quad (3.18)$$

denotes the set of alias frequencies of k appearing in the overparameterized model with spectrum Ω_d . The optimal model is then expressed as

$$f^{opt}(x) = \sum_{k \in \Omega_{n_0}} \tilde{y}_k \sum_{\ell \in S(k)} e^{i2\pi \ell x} \frac{\nu_\ell}{\sum_{j \in S(k)} \nu_j}. \quad (3.19)$$

Recalling that our model f is trained on n noisy samples $(y_0, \dots, y_{n-1})$ of the target function g , we are interested in the squared error of the model f averaged over (noisy) samples over the input domain,

$$L(f) := \mathbb{E}_{x, \mathbf{y}} (f(x) - g(x))^2. \quad (3.20)$$

and we call L the *generalization error* of f , as it captures the behavior of f with respect to g over the entire input domain $x \in [0, 1]$ instead of just the uniformly spaced training points x_j . In Appendix A.1 we derive

$$L(f^{opt}) = \underbrace{\frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{j \in S(k)} \nu_j\right)^2}}_{\text{VAR}} + \underbrace{\sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \left[\left(\frac{\sum_{\ell \in S(k) \setminus k} \nu_\ell}{\sum_{j \in S(k)} \nu_j} \right)^2 + \frac{\sum_{\ell \in S(k) \setminus k} \nu_\ell^2}{\left(\sum_{j \in S(k)} \nu_j\right)^2} \right]}_{\text{BIAS}^2} \quad (3.21)$$

We use this generalization error now to explore two interesting behaviors of the interpolating model in our setting: The tradeoff between noise absorption and signal recovery exemplified by the cases in Sec. 3.3.2, and the ability of an overparameterized Fourier features model to *benignly* overfit the training data.

The first behavior involves a trade-off in the *generalization error* $L(f^{opt})$ between absorbing noise (reducing VAR) and capturing the target function signal (reducing BIAS²) that recovers and generalizes the behavior of the two cases in Sec. 3.3.2. This trade-off is controlled by three components: The noise variance σ^2 , the input signal Fourier coefficients $\hat{g}_k$, and the distribution of feature weights $\sqrt{\nu_k}$. As described in the two cases above, when $\sigma^2 \rightarrow 0$ (signal only) the variance term VAR vanishes and the model is biased for any choice of nonzero ν_k where $k > n$. Conversely, when $\hat{g} \rightarrow 0$ (noise only) the bias term BIAS² vanishes, and the variance term is minimized by choosing uniform ν_k for all $k \in \Omega_d$.

The second interesting behavior occurs when the generalization error of the overparameterized model decreases at a nearly optimal rate as the number of samples n increases, known as *benign overfitting*. Prior work on benign overfitting in linear regression studied scenarios where the distribution of input data varied with the dimensionality of data and size of the training set in such a way that the excess generalization error of the overparameterized model (compared to a simple model) vanished [68]. However, since the dimensionality of the input data for our model is fixed, we instead consider sequences of feature weights that vary with respect to the problem parameters (n_0, n, d) in a way that results in BIAS² and VAR vanishing as $n \rightarrow \infty$. In this case, by fitting an increasing number of samples n using such a sequence of feature weights, the overparameterized model both perfectly fits the training data and generalizes well for unseen test points, and therefore exhibits benign overfitting.

These behaviors are exemplified by a straightforward choice of feature weights that incorporate some prior knowledge of the spectrum Ω_{n_0} available to the target function g . For all $k \in \Omega_{n_0}$, let $\nu_k = c/n_0$ for some positive c and normalize the feature weights so that $\sum_{k \in \Omega_k} \nu_k = 1$. We show in Appendix A.1.3.1 that the error terms of $L(f^{opt})$ scale as

$$\text{VAR} = \mathcal{O}\left(\frac{1}{n} + \frac{n}{d}\right), \quad (3.22)$$

$$\text{BIAS}^2 = \mathcal{O}\left(\frac{1}{n^2}\right). \quad (3.23)$$

Thus, as long as the dimension of the overparameterized Fourier features model grows strictly faster than n (i.e., $d = \omega(n)$), the model exhibits benign overfitting. In Appendix A.1.3.2 we demonstrate how this simple example actually captures the mechanisms of benign overfitting for much more general choices of feature weights. Fig. 3.3 summarizes this behavior and provides an example of the bias-variance tradeoff that occurs for overparameterized models. In particular, Fig. 3.3a exemplifies the setting in which benign overfitting occurs, wherein the feature weights of the Fourier features model are strongly

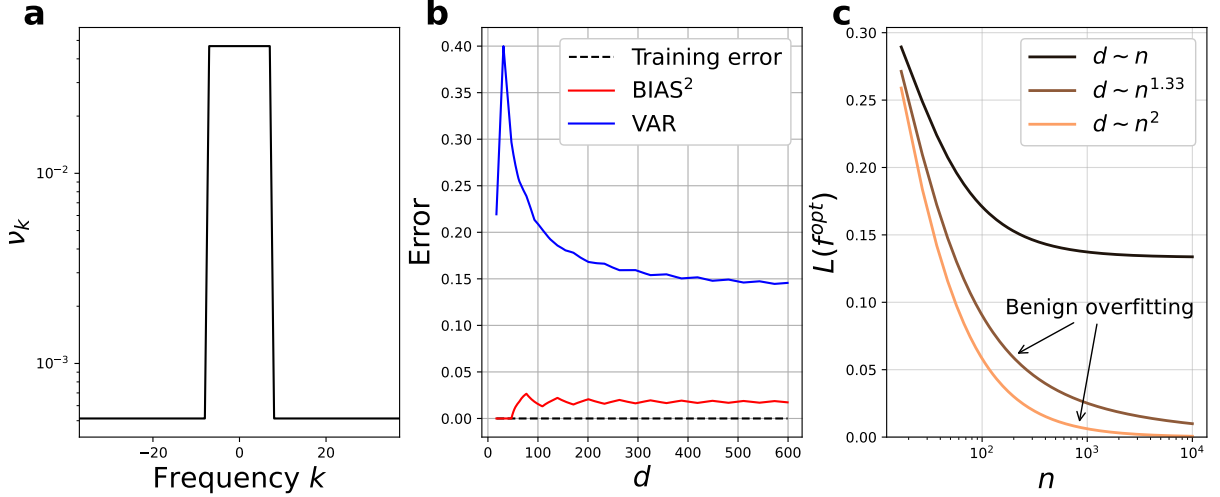


Figure 3.3: The feature weights ν_k may be engineered to exhibit a bias-variance tradeoff and benign overfitting. **a** A demonstration of feature weights ν_k which account for prior knowledge of the target function g , with large ν_k for $k \in \Omega_{n_0}$ and small ν_k for $k \in \Omega_d \setminus \Omega_{n_0}$. **b** After sampling a bandlimited target function ($n_0 = 15$) at a fixed number of uniformly spaced points ($n = 31$), the generalization error of the minimum ℓ_2 -norm model f^{opt} consists of a tradeoff between decreasing VAR and increasing BIAS^2 for larger d . Note the peaking behavior for VAR at $d/n = 1$. **c** This choice of feature weights ν_k benignly overfits the input signal such that $\lim_{n,d \rightarrow \infty} L(f^{opt}) = 0$ for any scaling of $d = n^\alpha$ with $\alpha > 1$, and fails to generalize on the interval $[0, 1]$ when $\alpha \leq 1$.

concentrated over frequencies in Ω_{n_0} but extend over a large range of alias frequencies for each $k \in \Omega_{n_0}$.

The generalization behavior described here is fundamentally different from many generalization guarantees typically found in statistical learning theory. While prior work has derived guarantees for the generalization of quantum models by constructing bounds on the complexity of the model class [84], Eqs. 3.22-3.23 demonstrate that generalization may occur as the complexity (i.e. dimension) of a model grows arbitrarily large.

So far, we have reviewed the Fourier perspective on fitting periodic functions in a classical setting and extended the analysis to characterize benign overfitting. However, if we can link the basic components of quantum models to the terms appearing in the error of Eq. (3.21), then we will be able to study a similar trade-off in the error of overparameterized

quantum models and the conditions necessary for benign overfitting. The remainder of this work is devoted to showing that analogous mechanisms exist in certain quantum machine learning models, and to studying the choices of feature weights for which quantum models can exhibit tradeoffs in generalization error and benign overfitting.

3.4 Benign overfitting in single-layer quantum models

In the previous section we have seen that the feature weights $\sqrt{\nu_k}$ balance the trade-off between absorbing the noise and hurting the signal of overparametrized models. To understand how different design choices of quantum models impact this balance, we need to link their mathematical structure to the model class defined in Eq. (3.4), and in particular to the feature weights, which is what we do now.

The type of quantum models we consider here are simplified versions of parametrized quantum classifiers (also known as *quantum neural networks*) that have been heavily investigated in recent years [75, 93, 94]. They are represented by quantum circuits that consist of two steps: first, we encode a datapoint $x \in [0, 1]$ into the state of the quantum system by applying a d -dimensional unitary operator $V(x)$, and then we measure the expectation value of some d -dimensional (Hermitian) observable M . This gives rise to a general class of quantum models of the form

$$f(x) = \langle 0 | V^\dagger(x) M V(x) | 0 \rangle. \quad (3.24)$$

To simplify the analysis, we will consider a quantum circuit $V(x)$ that consists of a data-independent unitary U and a diagonal data-encoding unitary generated by a d -dimensional Hermitian operator H ,

$$S(x) = \exp(i2\pi \cdot \text{diag}(\lambda_0, \dots, \lambda_{d-1})x) = \exp(i2\pi Hx), \quad (3.25)$$

which includes a large class of quantum models commonly studied but excludes schemes involving data re-uploading [41, 95]. Defining $U|0\rangle = |\Gamma\rangle = \sum_{j=1}^d \gamma_j |j\rangle$, the output of this quantum model becomes

$$f(x) = \langle \Gamma | S^\dagger(x) M S(x) | \Gamma \rangle, \quad (3.26)$$

where $|\Gamma\rangle$ can be treated as an arbitrary input quantum state. We call the corresponding quantum circuit for this model single-layer in the sense that it contains a single diagonal data-encoding layer in which all data-dependent circuit operations could theoretically be

executed simultaneously (though in general the operation U and measurement M may require significant depth to implement).

Applying insights from Refs. [42, 87], quantum models of this form can be expressed in terms of a Fourier series

$$f(x) = \sum_{k \in \Omega} e^{i2\pi kx} \sum_{\ell, m \in R(k)} \gamma_\ell \gamma_m^* M_{m\ell}, \quad (3.27)$$

where the *spectrum* Ω as well as the *partitions* $R(k)$ depend on the eigenspectrum $\lambda(H)$ of the data-encoding Hamiltonian H :

$$\Omega = \{(\lambda_\ell - \lambda_m) : \lambda_\ell, \lambda_m \in \lambda(H)\} \quad (3.28)$$

$$R(k) = \{(\ell, m) : \lambda_\ell - \lambda_m = k; \lambda_\ell, \lambda_m \in \lambda(H)\}. \quad (3.29)$$

Comparing Eq. (3.27) to Eq. (3.4) we see that the quantum model may be expressed as a linear combination of weighted Fourier modes, but it is not yet clear how the input state γ_j and the trainable observable M of the quantum model correspond to feature weights ν_k for each Fourier mode. To reveal this correspondence, we will need to first find the minimum-norm interpolating observable that solves the optimization problem

$$M^{opt} = \arg \min_{M=M^\dagger: f(x_j)=y_j \forall j} \|M\|_F, \quad (3.30)$$

where $\|M\|_F = \sqrt{\text{Tr } M^\dagger M}$ denotes the Frobenius norm of M . Solving Eq. (3.30) is analogous to the minimization the ℓ_2 norm of α in the classical optimization problem of Eq. (3.6), and serves a role similar to regularization commonly applied to quantum models by introducing a penalty term proportional to $\|M\|_F^2$ [34, 36, 96]. In Appendix A.2 we prove that subject to the condition that $\gamma_i > 0$, the minimum- $\|\cdot\|_F$ interpolating observable that solves Eq. (3.30) is given as

$$(M^{opt})_{m\ell} = \hat{y}_k \frac{\gamma_\ell^* \gamma_m}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2}, \quad (3.31)$$

for all $\ell, m \in R(S_k)$, and the corresponding optimal quantum model is

$$f^{opt}(x) = \sum_{k \in \Omega_{n_0}} \hat{y}_k \sum_{\ell \in S(k)} e^{i2\pi \ell x} \frac{\nu_\ell^{opt}}{\sum_{j \in S(k)} \nu_j^{opt}}, \quad (3.32)$$

where $S(k)$ denotes the set of aliases of k appearing in Ω from Eq. (3.18). By comparison to the optimal classical model of Eq. (3.19) we have identified the feature weights of the optimized quantum model as

$$\nu_k^{opt} := \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2. \quad (3.33)$$

Interestingly, while there was initially no clear way to separate the building blocks of the quantum model in Eq. (3.27) into trainable weights α_k and feature weights ν_k , this separation has now appeared after solving for the optimal observable M^{opt} . Furthermore, the optimal quantum model depends on $|\gamma_i|$ and is independent of phases associated with amplitudes γ_i (an effect that stems from using only a single data-encoding layer $S(x)$ in the quantum model).

From Eq. (3.33) it is clear that the partitions $R(k)$ of Eq. (3.29) arising from the choice of data-encoding unitary $S(x)$ have a strong relationship with the feature weights ν_k of the quantum model. We will now consider a simplified quantum model to highlight this relationship, thereby identifying a tradeoff between noise absorption and target signal recovery and the possibility of observing benign overfitting in quantum models.

3.4.1 Simplified quantum model

To explicitly highlight the role of $R(k)$ in controlling the feature weights ν_k^{opt} of the optimized quantum model, we will now simplify the model by using an equal superposition input state $|\Gamma\rangle = \frac{1}{\sqrt{d}} \sum_{j=0}^{d-1} |j\rangle$ and by restricting the set of observables considered during optimization. If we fix every entry of the observable with respect to elements in a partition $R(k)$ to be proportional to some complex constant $M(k)$:

$$M_{m\ell} = \frac{M(k)}{\sqrt{|R(k)|}} \quad \text{for all } \ell, m \in R(k), \quad (3.34)$$

then we can simplify the quantum model of Eq. (3.27) to

$$f(x) = \sum_{k \in \Omega} M(k) \frac{\sqrt{|R(k)|}}{d} e^{i2\pi kx}. \quad (3.35)$$

Comparing Eq. (3.35) to Eq. (3.4) we identify a direct correspondence between the trainable weights α_k in the classical model with $M(k)$, as well as a correspondence between

the feature weights $\sqrt{\nu_k}$ and the the *degeneracy* $|R(k)|$ of the quantum model. Making the substitutions $\alpha_k \rightarrow M(k)$, $\sqrt{\nu_k} \rightarrow \sqrt{|R(k)|}/d$, one can verify that $\sum_{k \in \Omega} |M(k)|^2 = \|M\|_F^2$ for this restricted choice of M and so the solution to the optimization of Eq. (3.30) is essentially the same as that of the classical problem in Eq. (3.17).

The crucial property of the simplified model is that the degeneracy $|R(k)|$ – and hence the combinatorial structure introduced by the data encoding Hamiltonian’s eigenvalues – completely controls the trade-off in the generalization error (Eq. (3.21)). We can hence study different types of partitions $R(k)$ to show a direct effect of the data-encoding unitary $S(x)$ on the fitting and generalization error behaviors for this simplified, overparameterized quantum model.

To study these behaviors we will now consider specific families of H which we call *encoding strategies* since the choice of H completely determines how the data is encoded into the quantum model. While $R(k)$ and Ω may be computed for an arbitrary $S(x)$ using brute-force combinatorics, some encoding strategies lead to particularly simple solutions. We have derived a few such examples of simple degeneracies and spectra for different encoding strategies in Appendix A.3 and present the results in Table 3.1. These choices highlight the extreme variation in Ω resulting from minor changes to $S(x)$, for example $|\Omega| \propto n_q$ for the “Hamming” encoding strategy compared to $|\Omega| \propto 3^{n_q}$ for the “Ternary” encoding strategy. These examples also highlight the limitations in constructing Hamiltonians with specific properties such as uniform $|R(k)|$ or evenly-spaced frequencies in Ω .

Since the feature weights ν_k of the Fourier modes are fixed by the choice of the data-encoding unitary, we can understand a choice of $S(x)$ as providing a structural bias of a quantum model towards different overfitting behaviors, and conversely the choices of feature weights available to quantum models are limited and are particular to the structure of the associated quantum circuit. Figure 3.4 shows distributions for feature weights arising from the example encoding strategies presented in Table A.1, and demonstrates a broad connection between the degeneracies $|R(k)|$ of the model (giving rise to feature weights ν_k^{opt}) and the generalization error $L(f^{opt})$.

3.4.2 Trainable unitaries reweight the general quantum model

We now return to the general quantum model of Eq. (3.27) to understand how different choices of the state preparation unitary U (giving rise to input state $|\Gamma\rangle$) affect the feature weights ν_k^{opt} of the general quantum model, thereby influencing benign overfitting of the target function. While Eq. (3.33) differs from the correspondence $\nu_k = |R(k)|/d^2$ that we observed for the simplified quantum model, we see that the feature weights of the optimal

Encoding strategy	Hamiltonian example	Degeneracy $ R(k) $	Spectrum Ω	$ \Omega $
Hamming	$\frac{1}{2}(Z_0 + Z_1 + Z_2 + \dots)$	$\binom{2n_q}{n_q - k }$	$\{-n_q, 1 - n_q, \dots, n_q\}$	$2n_q + 1$
Binary	$\frac{1}{2}(Z_0 + 2Z_1 + 4Z_2 + \dots)$	$2^{n_q - k }$	$\{-2^{n_q} + 1, \dots, 2^{n_q} - 1\}$	$2^{n_q+1} - 1$
Ternary	$\frac{1}{2}(Z_0 + 3Z_1 + 9Z_2 + \dots)$	$2^{n_q - \ T(k)\ _1}$	$\left\{-\left\lfloor \frac{3^{n_q}}{2} \right\rfloor, \dots, \left\lfloor \frac{3^{n_q}}{2} \right\rfloor\right\}$	3^{n_q}
Golomb	diag $(0, 1, 4, 6)$	$\begin{cases} d & k = 0 \\ 1 & k \neq 0 \end{cases}$	varies	$2 \binom{d}{2} + 1$

Table 3.1: Spectra Ω and degeneracies $R(k)$ computed for various data-encoding Hamiltonians H defined for either d dimensions or n_q qubits. The Hamming, Binary, and Ternary data encoding strategies are realized on n_q qubits using a separable Hamiltonian consisting of Pauli- Z operators with different prefactor schemes. The Ternary encoding strategy results in the largest $|\Omega|$ possible for a separable data encoding Hamiltonian (see also Ref. [97]), while the Golomb encoding strategy (named in reference to Golomb rulers, e.g. [98]) results in the largest $|\Omega|$ possible for any choice of d -dimensional data-encoding Hamiltonian. Note that the spectrum is preserved under permutations and additive shifts of the diagonal of the Hamiltonian and so we use “ $\sim$ ” to denote equivalence up to these operations. The function T converts an integer to a signed ternary string, as defined in Eq. (A.197) (see Appendix A.3).

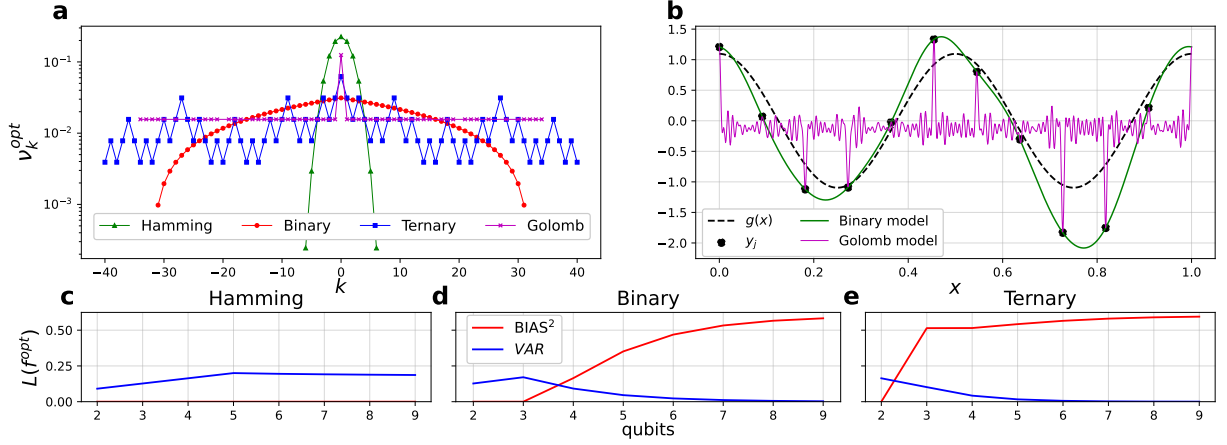


Figure 3.4: The degeneracies $|R(k)|$ for select quantum models directly influence the tendency of the models to overfit noise or bleed the underlying signal. **a** The degeneracies $|R(k)|$ for each frequency k for the four models introduced in Table 3.1 induce significantly different feature weights in the simplified quantum model. **b** Two models exemplify the extreme cases of fitting behaviors from Sec. 3.3.2: The Golomb encoding strategy (magenta) has significant weight on feature weights ν_k for large frequencies k , resulting in a “spiky” overfitting behavior that interpolates each noisy sample y_j but sacrifices recovery of the underlying signal $g(x)$; the Hamming encoding strategy (green) fits noisy samples smoothly and the error of the model is sensitive to the noise in the sampled data. **c-e** The BIAS^2 and VAR of the optimized Hamming encoding strategy (**c**), Binary encoding strategy (**d**), and Ternary encoding strategy (**e**) demonstrate the general effect of $|R(k)|$ and ν_k on the generalization error. The concentration of feature weights ν_k on $k \in \Omega_{n_0}$ in the Hamming encoding strategy reduces the BIAS (**c**), while ν_k having broad support over $k \in \Omega$ (e.g. Ternary, **e**) reduces VAR as the resulting f is close to its average (with respect to noise) except for becoming “spiky” in the vicinity of samples y_j , but increases the BIAS by introducing alias frequencies into the optimized model f^{opt} .

quantum model still depend heavily on the degeneracies $|R(k)|$. For instance, the *average* ν_k^{opt} with respect to Haar-random d -dimensional input states $|\Gamma\rangle$ is proportional to $|R(k)|$ whenever $k \neq 0$:

$$\mathbb{E}_{U \sim \mathcal{U}(d)} \nu_k^{opt} = \frac{|R(k)| + d\delta_0^k}{d(d+1)}. \quad (3.36)$$

where δ_i^j denotes the Kronecker delta. Furthermore, in Appendix A.2 we observe that the variance of ν_k^{opt} around its average tends to be small for encoding strategies considered in this work, for instance scaling like $\mathcal{O}(d^{-3})$ for the Binary encoding strategy and $\mathcal{O}(d^{-4})$ for the Golomb encoding strategy. This demonstrates that the feature weights of generic quantum models (i.e., ones for which U is randomly sampled) will be dominated by the degeneracy $|R(k)|$ introduced by the data-encoding operation $S(x)$.

Despite the behavior of ν_k being dominated by $|R(k)|$ in an average sense, there are specific choices of U for which the feature weights deviate significantly from this average. We will now use one such choice of U to provide a concrete example of an interpolating quantum model that exhibits benign overfitting. Suppose we choose U such that the elements of $|\Gamma\rangle$ are given by

$$\gamma_j = \begin{cases} \sqrt{a}, & j \in [c_1, c_2), \\ \sqrt{b}, & \text{otherwise} \end{cases} \quad (3.37)$$

for $j \in [d]$, and some integers $0 < c_1 < c_2 < d$ and amplitudes a, b subject to normalization. We show that given a band-limited target function g with access to spectrum Ω_{n_0} , there is a specific choice of c_1, c_2 dependent on d and n_0 for which the interpolating quantum model f^{opt} of Eq. (3.32) also benefits from vanishing generalization error in the limit of many samples, namely we show in Appendix A.4 that

$$L(f^{opt}) = \mathcal{O}\left(\frac{1}{n} + \frac{n}{d}\right). \quad (3.38)$$

Thus, by perfectly fitting the training data and exhibiting good generalization in the $n \rightarrow \infty$ limit, the quantum model exhibits benign overfitting. This behavior is outlined in Figure 3.5, which highlights the role that $|\Gamma\rangle$ plays in concentrating the feature weights ν_k within the spectrum Ω_{n_0} of g while preserving a long tail that provides the model with low-variance “spiky” behavior in the vicinity of noisy samples. In contrast, the feature weights for the Binary encoding strategy with a uniform input state has little support on Ω_{n_0} and resulting in a large BIAS error.

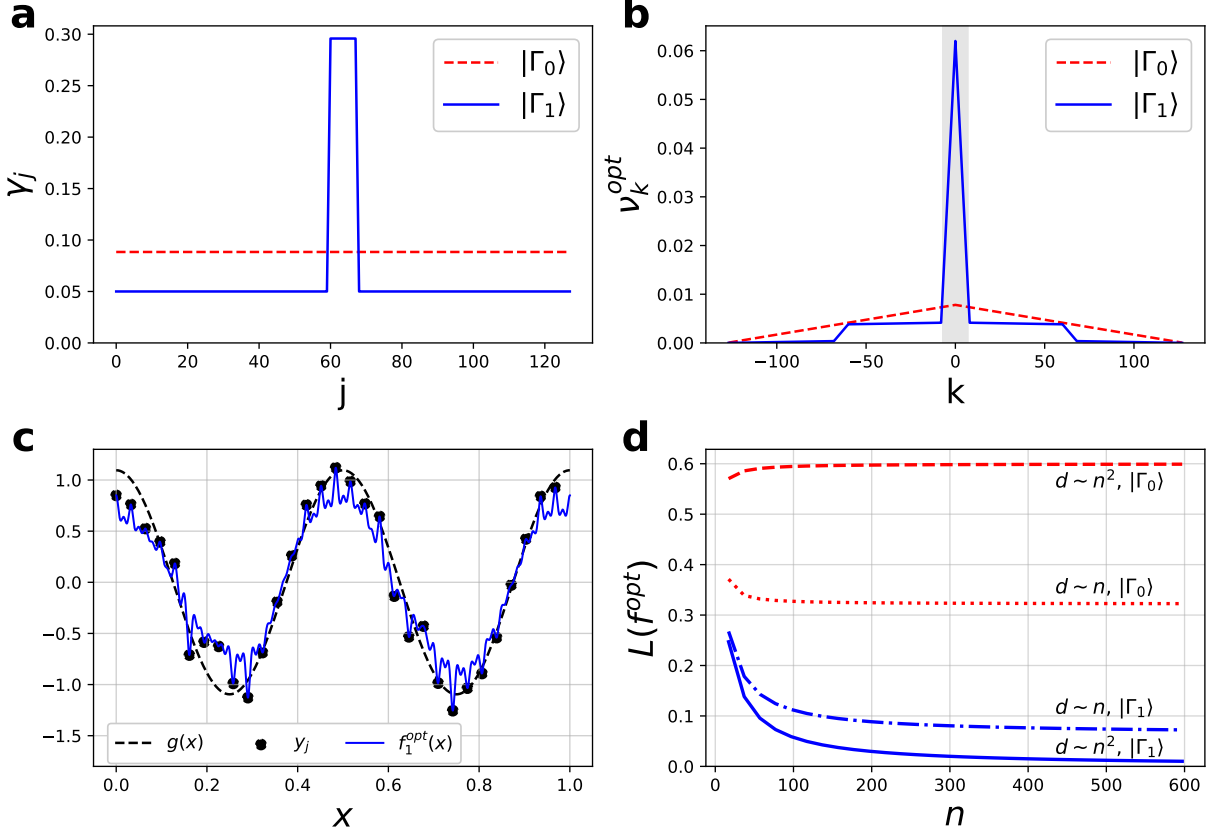


Figure 3.5: Benign overfitting for the general quantum model. **a** For the Hamiltonian $H \sim \text{diag}((0, 1, \dots, d-1))$ we construct a simple input state $|\Gamma_1\rangle$ (Eq. (3.37)) that is dependent on n_0 and d (see Appendix A.4), compared to the uniform state $|\Gamma_0\rangle = \frac{1}{\sqrt{d}} \sum_k |k\rangle$. **b** $|\Gamma_1\rangle$ induces feature weights ν_k^{opt} in the optimized model that are heavily concentrated in the target function spectrum Ω_{n_0} (gray band) and decay like $1/d$ and $1/d^2$ elsewhere for $k \in \Omega_d \setminus \Omega_{n_0}$, compared to the feature weights induced by $|\Gamma_0\rangle$ which have similar magnitude over all of Ω_d . **c** Plotting the $\|M\|_F$ -minimizing quantum model f_1^{opt} with input state $|\Gamma_1\rangle$ for a sample signal ($d = 128, n = 31$) highlights how $|\Gamma_1\rangle$ and H work together to interpolate y_j via high-frequency “spiky” behavior near the sampled points, but otherwise closely matches the underlying signal $g(x)$. **d** The resulting quantum model benignly overfits the band-limited target function: As long as $d = \omega(n)$, the generalization error $L(f_1^{opt})$ of the minimum- $\|M\|_F$ interpolating quantum model vanishes for large enough sample size n . Circuit simulations were performed in PennyLane [99].

The above discussion shows how the input state amplitudes γ_i provide additional degrees of freedom with which the feature weights ν_k^{opt} can be tuned in order to modify the generalization behavior of the interpolating quantum model, and to exhibit benign overfitting in particular. It is therefore worthwhile to consider what other kinds of feature weights ν_k might be prepared by some choice of input state $|\Gamma\rangle$. We may use simple counting arguments to demonstrate the restrictions in designing particular distributions of feature weights. Suppose we define $\Omega_+ = \{k : k \in \Omega, k > 0\}$ containing the positive frequencies of a quantum model. Then the introduction of an arbitrary input state $|\Gamma\rangle$ provides us with $2^{n_q} - 1$ free parameters with which to tune $|\Omega_+|$ -many terms in the distribution of ν_k^{opt} (subject to $\nu_k = \nu_{-k}$ and $\sum_k \nu_k = 1$). Clearly, there are distributions of feature weights ν_k^{opt} that can not be achieved for models where $|\Omega_+| \geq 2^{n_q}$.

Conversely, the condition $|\Omega_+| < 2^{n_q}$ does not necessarily mean that we can thoroughly explore the space of possible feature weights by modifying the input state $|\Gamma\rangle$. For example, consider the Hamming encoding strategy for which the number of free parameters controlling the distribution of feature weights ν_k^{opt} is $|\Omega_+| = n_q$, which is exponentially smaller the number of parameters in $|\Gamma\rangle$. While this might suggest significant freedom in adjusting ν_k^{opt} , the opposite is true: For any choice of input state $|\Gamma\rangle$, there is another state of the form

$$|\Gamma'\rangle = \sum_{i=0}^{n_q} \phi_i |\Phi_i\rangle, \quad (3.39)$$

that achieves exactly the same distribution of feature weights ν_k . In Eq. (3.39), $|\Phi_i\rangle$ describes a uniform superposition over all computational basis state bitstrings with weight i , and so the distribution of ν_k^{opt} actually only depends on $n_q + 1$ real parameters ϕ_i , $i = 0, 1, \dots, n_q$, and the feature weights are invariant under any operations in U that preserve $|\Phi_i\rangle$ (see Appendix A.2). An example of such operations are the *particle-conserving unitaries* well-known in quantum chemistry, which act to preserve the number of set bits (i.e., the Hamming weight) when each bit represents a mode occupied by a particle in second quantization [100, 101]. This example demonstrates how symmetry in the data-encoding Hamiltonian (e.g. Refs. [102, 103]) can have a profound influence on the ability to prepare specific distributions of feature weights ν_k^{opt} , and consequently affect the generalization and overparameterization behavior of the associated quantum models.

3.4.3 Weight behavior in deep quantum models

Using the formalism developed in the previous section, we return to discussing quantum kernels. In particular, we will use the tools of the previous sections to compute the average

value of a quantum kernel $k(x, x')$ with respect to trainable unitaries sampled uniformly at random. We will demonstrate that in this setting, a quantum kernel computed using a very deep quantum circuit will be translation-invariant and therefore easy to simulate classically, by performing combinatorics that depend only on knowledge of the structure of the quantum circuit.

Consider a single-qubit embedding $\rho(x)$ prepared by applying L layers, each consisting of trainable unitaries interleaved with data-encoding unitaries of the form $S_k(x) := \exp(i\alpha_k x Z/2)$ for $k = 1, \dots, L$. That is, define

$$|\psi(x)\rangle = U_{L+1} S_L(x) U_L \cdots S_1(x) U_1 |0\rangle, \quad (3.40)$$

and say $\rho(x) = |\psi(x)\rangle\langle\psi(x)|$ with implicit dependence on $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_L) \in \mathbb{R}^L$. Note that this belongs to the so-called “data-reuploading” setting of Ref. [95]. Let $k(x, x') = \text{Tr } \rho(x)\rho(x')$ be the associated quantum kernel. Define the sets

$$Q(\omega) = \{\mathbf{k} : \boldsymbol{\alpha} \cdot \mathbf{k} = \omega, \mathbf{k} \in \{-1, 0, 1\}^L\}, \quad (3.41)$$

$$\Omega = \{\omega : \boldsymbol{\alpha} \cdot \mathbf{k} = \omega, \mathbf{k} \in \{-1, 0, 1\}^L\}. \quad (3.42)$$

Then, for trainable unitaries $\{U_1, \dots, U_{L+1}\}$ sampled uniformly with respect to the single-qubit Haar measure, the corresponding kernel is translation-invariant $k(x, x') = k(x - x') = \sum_{\omega} k_{\omega} e^{-i(x-x')\omega}$ with Fourier coefficients

$$\mathbb{E}_{U \sim \text{Haar}} k_{\omega} = \frac{3^{L-1} \delta_0^{\omega} + |Q(\omega)|}{2 \cdot 3^L} \quad (3.43)$$

for all $\omega \in \Omega$. This statement is proven in App. A.5, and is somewhat similar to results derived in Ref. [104]. This is significant in that it demonstrates such deep models are, on average with respect to trainable unitaries, easy to simulate. The easiness argument follows from Bochner’s theorem and the random Fourier features technique: because k is translation-invariant, its Fourier coefficients are distributed with respect to a (positive) measure and so the corresponding feature map may be sampled (with good convergence guarantees) according to Eq. 3.43. In other words, the quantum embedding corresponding to this learning model is well-approximated by Fourier features. This reproduces observations that have been made in a series of subsequent works addressing classical techniques for simulating quantum machine learning models [105–107].

3.5 Discussion

The work in this chapter represents a first step towards characterizing the phenomenon of benign overfitting in a quantum machine learning setting. We derived the error for an

overparameterized Fourier features model that interpolates the (noisy) input signal with minimum ℓ_2 -norm trainable weights and connected the feature weights associated with each Fourier mode to a trade-off in the generalization error of the model. We then demonstrated an analogous simplified quantum model for which the feature weights are induced by the choice of data-encoding unitary $S(x)$. Finally, we discussed how introducing an arbitrary state-preparation unitary U gives rise to effective feature weights in the optimized general quantum model, presenting the possibility of connecting U and $S(x)$ to benign overfitting in more general quantum models.

Our discussion of interpolating quantum models presents an interpretation of overparameterization (i.e., the size of the model spectrum Ω) that departs from other measures of quantum circuit complexity discussed in the literature [79, 108, 109], as even the simplified quantum models studied here are able to interpolate training data using a fixed circuit U and optimized measurements. We also reemphasize that – unlike much of the quantum machine learning literature – we do not consider a setting where the model is optimized with respect to a trainable circuit, as the model of Eq. (3.32) is constructed to exhibit zero training error (and can therefore not be improved via optimization). Finding the input state $|\Gamma\rangle$ that will result in a specific distribution of feature weights ν_k^{opt} generally requires solving a $|\Omega_+|$ -dimensional system of equations that are second order in 2^{n_q} many real parameters $|\gamma_i|^2$ (i.e., inverting the map of the form $\mathbb{R}^{2^{n_q}} \rightarrow \mathbb{R}^{|\Omega_+|}$ in Eq. (3.33)) or otherwise performing a variational optimization that will likely fail due to the familiar phenomenon of *barren plateaus* [82, 83, 110, 111].

While we have shown an example of benign overfitting by a quantum model in a relatively restricted context, future work may lead to more general characterizations of this phenomenon. Similar behavior likely exists for quantum kernel methods and may complement existing studies on these methods’ generalization power [1]. An exciting possibility would be to demonstrate benign overfitting in quantum models trained on distributions of quantum states which are hard to learn classically [112, 113], thereby extending the growing body of statistical learning theory for quantum learning algorithms [85, 86, 114].

Chapter 4

Improving information transmission using correlated auxiliary noise

4.1 Chapter overview

Communication over noisy channels involves an information source being corrupted by noise during transmission from a sender to a receiver. This chapter introduces and analyzes *correlated auxiliary noise* for communication with noisy channels. In this setting, based on a similar scenario explored in the context of machine learning [19], the receiver has access to an additional resource that is correlated with the noise that affects the transmitted signal. We study this setting through the lens of (quantum) Shannon theory for discrete random variables and finite dimensional quantum systems. In Sec. 4.4 we derive the capacity and optimal estimation probability for communication through a classical bit flip channel with correlated auxiliary noise. Sec. 4.4 demonstrates how this idea may be applied to derive information-theoretic error rates for decoding quantum error correcting codes. Then, in Sec. 4.6 we extend these ideas to derive the quantum channel capacity of a quantum bit flip channel with correlated auxiliary noise.

This chapter consists entirely of unpublished work. I developed and derived its mathematical contents independently based on high-level discussions with Achim Kempf and Koji Yamaguchi.

4.2 Introduction

Our ability to transmit and process information is generally limited by noise. Characterizing the transmission of information through noisy channel is a fundamental task in information theory [115], and has implications for topics as varied as statistics, error correction, machine learning, and data compression. In all of these tasks, some external source of noise interferes with our attempts to process or transmit information. Yet, from a physical perspective, we know that noise arises from interactions between the system containing desirable information, and some external environment. In the course of these interactions, the state of a system and the state of its surrounding environment become correlated. When the environment is “discarded” or otherwise made inaccessible, this leads to information loss from the original system, which manifests as noise that increases the difficulty of processing information.

Conversely, detailed knowledge about the noise in a system can greatly improve our ability to learn from noisy data. When a signal is transmitted, information about that signal need not be completely lost to the environment due to the effects of noise. For instance, the environment itself could be correlated with an auxiliary system that the receiver could then retrieve and use to learn about the original signal. In this case, we may be capable of utilizing this *correlated auxiliary noise* to extract information about the errors that occurred in our system, thereby improving our ability to use the noisy channel to transmit information. Prior work showed that access to an auxiliary source of randomness that is correlated with noise in images could improve a learner’s performance at image recognition [19]. In that scenario, machine learning was used to automate the process of incorporating correlated auxiliary noise into an image classifier. Ref. [116] showed that preexisting correlations in a quantum field could be used to improve quantum channel capacities. These works provide evidence that correlated auxiliary noise is beneficial in machine learning and quantum communication, but leave open questions about the details of utilizing such a resource.

While there is empirical evidence for the benefits of correlated auxiliary noise in information processing tasks, in this chapter we will zoom in on a simple system in which we can analytically characterize the benefits of correlated auxiliary noise. We will consider several models for information transmission of the same basic form: A sender transmits a symbol through a channel, and the recipient receives a noisy version of that symbol along with some additional information (in the form of correlated auxiliary noise) about the noise that was sampled for this specific use of the channel.

4.2.1 Prior work

Several existing proposals have demonstrated improved efficiency for encoding a source that is correlated with another variable (in the environment or otherwise). Our setting is related to, but distinct from, providing *side information* at the receiver [117, 118]. Slepian and Wolf considered achievable rates for noiseless transmission of correlated sources in classical noisy channels [119], which was further expanded in Refs. [120, 121] to consider transmission of binary sources with side information. Similarly, a function of two binary variables may be encoded using a simple scheme [122]. Our work involves key differences: The correlated auxiliary noise is initially uncorrelated with the source (only becoming correlated after transmission), and we do not assume that the sender has any ability to encode these variables jointly. Our work relates more closely to Ref. [123] considering side information at the transmitter, and the specific analysis of modulo additive noise of Ref. [124].

Prior works have also considered how access to the environment of a quantum system could improve one's ability to transmit information through that system. Ref. [125] derived conditions for which measurement of a system's environment could be used to perfectly restore the original state of the system that had been subjected to noise. Furthermore, giving the receiver of a noisy system access to part of a system that is entangled with the environment has been shown to increase the classical [126] and quantum [20] capacity of a given channel in some cases.

4.3 Classical Shannon theory

We now introduce some basic quantities from classical information theory that we will apply towards understanding the transmission of data over noisy channels. We will refer to a (capital) *random variable* to describe an object that takes (lowercase) values in a (calligraphic) finite set with some probability, e.g. the random variable X will take values $x \in \mathcal{X}$ (where $|\mathcal{X}| < \infty$) with probability p_X , where $p_X : \mathcal{X} \rightarrow \mathbb{R}$ is a *probability density* defined on $\mathcal{X}$, i.e. $p_X(x) \geq 0 \forall x \in \mathcal{X}$ and $\sum_{x \in \mathcal{X}} p_X(x) = 1$. Let $\mathcal{P}(\mathcal{X})$ denote the simplex of probability distributions defined on $\mathcal{X}$. We may denote functions of random variables in two ways. A (random) function $f : \mathcal{A} \rightarrow \mathcal{B}$ may be used to denote a random variable B that takes values depending on A and some source of randomness, e.g. $B = f(A)$ ¹. Alternatively, we may denote a function, or *stochastic map* $g : \mathcal{P}(\mathcal{A}) \rightarrow \mathcal{P}(\mathcal{B})$ such that

¹In other words, we use “function” to refer to f even if each element $a \in \mathcal{A}$ is not uniquely mapped to a unique $b \in \mathcal{B}$

$p_B = g(p_A)$. Since fixing one of g or f determines the other, we may occasionally mix these notations up, e.g. $p_B = f(p_A)$ and $B = g(A)$. The (Shannon) *entropy* of A is defined

$$H(A) := - \sum_{a \in \mathcal{A}} p_A(a) \log p_A(a), \quad (4.1)$$

and for the special case $\mathcal{A} = \{0, 1\}$ and $p_A(0) := p$ we will define the binary entropy function as

$$h_2(p) = -p \log p - (1 - p) \log(1 - p). \quad (4.2)$$

The conditional entropy of a random variable A given B is defined

$$H(A|B) = - \sum_{a,b} p_{AB}(a,b) \log p_{A|B}(a|b) = H(AB) - H(B), \quad (4.3)$$

where $p_{\cdot|B}(\cdot|b) \in \mathcal{P}(\mathcal{B})$ denotes a *conditional probability density*. From the conditional entropy, we define the mutual information $I(A:B) := H(A) - H(A|B)$, which satisfies the *data processing inequality*

$$I(A:B) \geq I(A:C) \quad (4.4)$$

whenever $p_C = f(p_B)$. Similarly, we may make a straightforward statement that a deterministic function cannot increase the entropy of a random variable:

Proposition 1. Let f be a deterministic function mapping each $x \in \mathcal{X}$ to some $y \in \mathcal{Y}$. For random variables X, Y satisfying $Y = f(X)$, $H(X) \geq H(Y)$, with equality iff f^{-1} exists.

We now review classical hypothesis testing and its relation to information theory before discussing information-theoretic bounds and guarantees for learning problems. In classical hypothesis testing (or multiple hypothesis testing) a *target* random variable A taking values $a \in \mathcal{A}$ is sampled according to a discrete probability distribution $p_A : \mathcal{A} \rightarrow \mathbb{R}^+$. Then, *observations* are provided in the form of a random variable $B = \mathcal{N}(A)$ taking values in a finite set $\mathcal{B}$ according to probability $p_{B|A}$ conditioned on the value of A . The observations may not be a single data point - for instance, we might have that $\mathcal{B} = \mathcal{X}^n \times \mathcal{Y}^n$ represents a sequence of n many (x, y) pairs of a training set. Given the observations $B = b$ conditioned on sampling $A = a$, the objective is to output an *estimate* $\hat{A}(B)$ such that the success probability $p(A|\hat{A}) := \Pr_{p_A}(\hat{A}(B) = A)$ is large. See Fig. 5.1 for a schematic representation of classical hypothesis testing.

We are interested in upper and lower bounds on the error probability for this estimate. Intuitively, the estimate $\hat{A}$ is likely to be wrong if the uncertainty about variable A (given B) is large. Fano's inequality formalizes this intuition by lower bounding the failure probability

for any estimator $\hat{A} = \hat{A}(B)$ in terms of the conditional entropy between the inputs of that estimator and the target random variable [127]:

$$H(A|B) \leq h_2(p(A|\hat{A})) + (1 - p(A|\hat{A})) \log(|\mathcal{A}| - 1). \quad (4.5)$$

Furthermore, the probability of failure for an optimal estimate also is upper bounded according to

$$\max_{\hat{A}} p(A|\hat{A}) \geq 2^{-H(A|B)}, \quad (4.6)$$

where the maximization is over all possible estimators $\hat{A} : \mathcal{B} \rightarrow \mathcal{A}$. Bounds similar to (and tighter than) Eq. 4.6 have been derived in various works [128–130], and we will see later that this inequality follows immediately from the definition of *conditional min-entropy*.

In contrast to the above setting, the channel capacity of $\mathcal{N}$ is defined allowing for the possibility of encoding of the source variable. In our setting, the channel capacity for the channel $\mathcal{N} : \mathcal{P}(\mathcal{A}) \rightarrow \mathcal{P}(\mathcal{B})$ is defined [131]

$$\mathcal{C}(\mathcal{N}) = \sup_{p_A \in \mathcal{P}(\mathcal{A})} I(A:B). \quad (4.7)$$

Shannon famously showed that $\mathcal{C}(\mathcal{N})$ is also equal to the highest possible rate (bits per channel use) of information transmission through $\mathcal{N}$ with arbitrarily low probability of error [115].

4.4 Correlated auxiliary noise for a bit flip channel

We will study correlated auxiliary noise for a simple model that involves transmitting single bits through a bit flip channel. We consider a binary source $X \in \{0, 1\}$ with probability distribution $p_X : \{0, 1\} \rightarrow \mathbb{R}^+$. This source is then subjected to a bit flip conditioned on the value of a noise variable $\Gamma \in \{0, 1\}$, with one of the channel outputs being $Z = X \oplus \Gamma$. An additional variable $\Delta \in \{0, 1\}$ represents correlated auxiliary noise with Γ : For each use of the channel, a pair (Γ, Δ) is drawn according to a joint distribution $p_{\Gamma\Delta}$, such that Δ is correlated with Z through the addition of noise Γ to the source X . A bit flip occurs with probability $p_{\Gamma}(1) = \sum_{\delta} p_{\Gamma\Delta}(1, \delta)$. In this setting, the sender and receiver have prior knowledge of $p_{\Gamma\Delta}$ but the sender does not have direct access to Δ nor Γ , while the receiver has access to (Z, Δ) and is tasked with recovering the bit X . We call this scenario a bit flip channel with correlated auxiliary noise $\mathcal{M} : \{0, 1\} \rightarrow \{0, 1\}^2$ producing outputs $\mathcal{M}(X) = (Z, \Delta)$ (see Fig. 4.1).

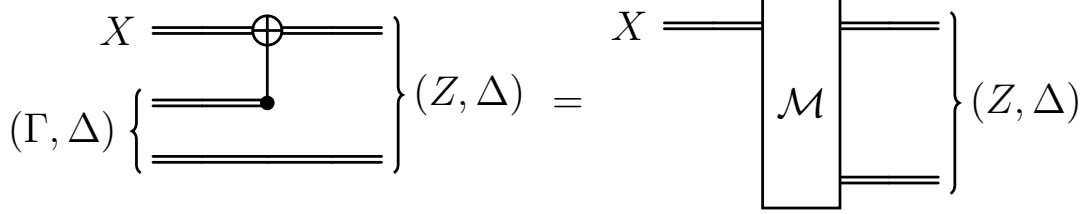


Figure 4.1: Circuit model for classical bit flip channel with onetime pad³. Double lines represent classical bits of information. The sender initially holds X , while the receiver has access to Z and Δ .

Since there are many possible decoders to recover a single bit from the output of $\mathcal{M}$, we make several simplifying assumptions. While there are $2^4 = 16$ possible decoders of the form $\hat{X}^c : \{0, 1\}^2 \rightarrow \{0, 1\}$ that could be used to recover X , the fact that X is initially uncorrelated with (Γ, Δ) means that, up to symmetries in the distribution p_X , the only relevant strategies involving Δ are $\hat{X}^c = Z \oplus \Delta$ (“controlled-NOT” conditioned on Δ) or $\hat{X}^c = (Z \oplus \Delta) \oplus 1$. We will usually only consider the former decoding strategy that applies a controlled-NOT on Z conditioned on the value of Δ (in regimes where this choice not optimal, the latter decoder is the better choice). In this case, the scenario in Fig. 4.1 may be considered as a channel from $\{0, 1\} \rightarrow \{0, 1\}$ with an output $\hat{X}^c := Z \oplus \Delta$.

4.4.1 The capacity of the correlated-noise bit flip channel

The ability for correlated auxiliary noise to improve the transmission of signals over a noisy channel makes intuitive sense for a particular choice of decoder. We first suppose the receiver implements a fixed decoder $\hat{X}^c(Z, \Delta) = Z \oplus \Delta$. We can treat the combined channel $\mathcal{M}$ plus this decoder as a bit flip channel with effective bit flip probability

$$p_{eff} := \Pr(\hat{X}^c = 1 | X = 0) \quad (4.8)$$

$$= \Pr(Z \oplus \Delta = 1 | X = 0) \quad (4.9)$$

$$= \Pr(\Gamma \neq \Delta) \quad (4.10)$$

$$= p_{\Gamma\Delta}(1, 0) + p_{\Gamma\Delta}(0, 1), \quad (4.11)$$

compared to the bit flip probability of $p_\Gamma(1)$ when Δ is not available. The condition for correlated auxiliary noise to improve the channel is $p_\Gamma(1) \geq p_{eff}$, or equivalently,

$$p_{\Gamma\Delta}(1, 1) \geq p_{\Gamma\Delta}(0, 1). \quad (4.12)$$

Intuitively, the success of decoder $\hat{X}^c$ requires that $\Delta = \Gamma$ more often than not. By the same reasoning above, $\Pr(\hat{X}^c = 0|X = 1) = p_{eff}$ and so we have recovered a binary symmetric channel with capacity $1 - h_2(p_{eff})$ [131]. By monotonicity of $h_2(x)$ for $0 < x < 0.5$, the capacity of the channel with correlated auxiliary noise $\hat{X} \circ \mathcal{M}$ exceeds the capacity of $\mathcal{M}$ whenever Eq. 4.12 holds. Fig. 4.2a shows the improvement in channel capacity when correlated noise is provided to the channel recipient.

We now show that the capacity of the bit flip channel using correlated auxiliary noise increases as Γ and Δ become more strongly correlated. Our main tool will be the following statement that correlations between Δ and Γ are never stronger than correlations between X and the pair (Z, Δ) :

Proposition 2. For binary random variables $X, Z, \Gamma, \Delta \in \{0, 1\}$ with $Z = X \oplus \Gamma$ and X is independent of Γ, Δ , then for any fixed distributions p_X and $p_{\Gamma\Delta}$,

$$H(X|Z\Delta) \leq H(\Gamma|\Delta) \quad (4.13)$$

Proof. To prove Proposition 2, we will need a few inequalities characterizing the behavior of entropy in general, and specifically for a bit flip channel.

Lemma 3. (Round robin) Let $X, Z, \Gamma, \Delta \in \{0, 1\}^n$ be classical binary-valued random variables such that $Z = X \oplus \Gamma$ (bitwise XOR) and (Γ, Δ) are distributed jointly according to probability $p_{\Gamma\Delta} : \{0, 1\}^{2n} \rightarrow \mathbb{R}^+$. Then,

$$H(XZ\Delta) = H(X\Gamma\Delta) = H(\Gamma Z\Delta).$$

This follows by noting that $(X, Z) = f(X, \Gamma)$ and $(\Gamma, Z) = f(X, \Gamma)$ for $f(a, b) := (a, a \oplus b)$ (thus $f^{-1} = f$ exists), and then applying Proposition 1. Lemma 3 implies that the uncertainty of Z given Γ is exactly $H(X)$:

$$H(Z|\Gamma) = H(Z\Gamma) - H(\Gamma) = H(X). \quad (4.14)$$

The next result shows that the correlation between Z and Δ obeys a kind of data processing inequality:

Lemma 4. Let $X, Z, \Gamma, \Delta \in \{0, 1\}^n$ be classical binary-valued random variables described in Proposition 2. Then,

$$H(Z|\Delta) \geq H(Z|\Gamma) \quad (4.15)$$

Proof. Using the chain rule for mutual information to expand $I(Z:\Gamma\Delta)$, we find

$$I(Z:\Delta) + I(Z:\Gamma|\Delta) = I(Z:\Gamma) + I(Z:\Delta|\Gamma) \quad (4.16)$$

$$= I(Z:\Gamma), \quad (4.17)$$

where we have used the conditional independence of Δ and Z given Γ :

$$p_{Z\Delta|\Gamma}(z, \delta|\gamma) = \frac{p_{Z\Delta\Gamma}(z, \delta, \gamma)}{p_{\Gamma}(\gamma)} \quad (4.18)$$

$$= \frac{p_{X\Delta\Gamma}(z \oplus \gamma, \delta, \gamma)}{p_{\Gamma}(\gamma)} \quad (4.19)$$

$$= p_X(z \oplus \gamma) \frac{p_{\Delta\Gamma}(\delta, \gamma)}{p_{\Gamma}(\gamma)} \quad (4.20)$$

$$= \frac{p_{X\Gamma}(z \oplus \gamma, \gamma)}{p_{\Gamma}(\gamma)} p_{\Delta|\Gamma}(\delta|\gamma) \quad (4.21)$$

$$= p_{Z|\Gamma}(z|\gamma) p_{\Delta|\Gamma}(\delta|\gamma). \quad (4.22)$$

Since conditional mutual information is positive and $I(Z:\Delta) = H(Z) - H(Z|\Delta)$ and $I(Z:\Gamma) = H(Z) - H(Z|\Gamma)$, we rewrite Eq. 4.17 according to

$$H(Z|\Delta) = H(Z|\Gamma) + I(Z:\Gamma|\Delta) \geq H(Z|\Gamma). \quad (4.23)$$

□

We may now bound $H(X|Z\Delta)$ as follows:

$$H(X|Z\Delta) = H(XZ\Delta) - H(Z\Delta) \quad (4.24)$$

$$= H(X\Gamma\Delta) - H(Z\Delta) \quad (\text{round robin}) \quad (4.25)$$

$$= H(X) + H(\Gamma\Delta) - H(Z\Delta) \quad (X \perp \Delta, \Gamma) \quad (4.26)$$

$$= H(X) + H(\Gamma|\Delta) - H(Z|\Delta) \quad (4.27)$$

$$\leq H(X) + H(\Gamma|\Delta) - H(Z|\Gamma) \quad (\text{Lemma 4}) \quad (4.28)$$

$$= H(X) + H(\Gamma|\Delta) - H(X) \quad (\text{round robin, Eq. 4.14}) \quad (4.29)$$

$$= H(\Gamma|\Delta). \quad (4.30)$$

Proof of Proposition 2 follows by setting $n = 1$. □

Proposition 2 allows us to prove the following:

Theorem 5. Let $\mathcal{M} : \{0, 1\}^n \rightarrow \{0, 1\}^{2n}$ be defined $\mathcal{M}(X^n) = (X_1 \oplus \Gamma_1, \dots, X_n \oplus \Gamma_n, \Delta^n)$ with $(\Gamma^n, \Delta^n) \in \{0, 1\}^n \times \{0, 1\}^n$ distributed jointly according to $p_{\Gamma^n \Delta^n}$. The classical capacity of $\mathcal{M}$ is

$$\mathcal{C}(\mathcal{M}) = n - H(\Gamma^n | \Delta^n) \quad (4.31)$$

Proof. Defining $Z_i := X_i \oplus \Gamma_i$, we may compute

$$I(X^n : Z^n \Delta^n) = H(X^n) - H(X^n | Z^n \Delta^n) \quad (4.32)$$

$$= H(Z^n \Delta^n) - H(\Gamma^n \Delta^n) \quad (4.33)$$

$$= H(Z^n | \Delta^n) - H(\Gamma^n | \Delta^n) \quad (4.34)$$

$$\leq H(Z^n) - H(\Gamma^n | \Delta^n) \quad (4.35)$$

$$\leq n - H(\Gamma^n | \Delta^n) \quad (4.36)$$

where line 4.33 follows by the same computation as line 4.26, and line 4.35 follows since conditioning reduces entropy. The final inequality is saturated when p_X is chosen to be the uniform distribution, in which case Z^n is uniformly distributed and $H(Z^n) = n$. As this choice saturates the upper bound for $I(X^n : Z^n \Delta^n)$, it corresponds to the channel capacity. $\square$

Setting $n = 1$ immediately leads to the capacity of transmitting a single bit with correlated auxiliary noise:

Corollary 6. Let $\mathcal{M} : \{0, 1\} \rightarrow \{0, 1\}^2$ be defined for input X and outputs $Z = X \oplus \Gamma$ and Δ , with (Γ, Δ) distributed jointly according to $p_{\Gamma \Delta}$. Then, the capacity of $\mathcal{M}$ is

$$\mathcal{C}(\mathcal{M}) = 1 - H(\Gamma | \Delta) \quad (4.37)$$

We can see that Corollary 6 recovers the capacity of a binary symmetric channel whenever $H(\Gamma | \Delta) = H(\Gamma)$, in which case Δ is uncorrelated with Γ and the (symmetric) probability of a bit flip is $p := p_\Gamma(1)$ with $H(\Gamma | \Delta) = h_2(p)$. On the other extreme end, perfect knowledge of Γ given Δ implies $H(\Gamma | \Delta) = 0$, such that the capacity of the channel is always one bit (since the receiver can reverse the effect of the bit flip perfectly).

4.4.2 Guarantees for hypothesis testing using correlated auxiliary noise

Having shown that correlated auxiliary noise improves channel capacity for a bit flip channel, we will now demonstrate a similar result for binary hypothesis testing. We consider

guarantees on the performance of an optimal hypothesis $\hat{X}^c$, i.e. how well the receiver can estimate X given Z, Δ and prior knowledge of $p_X, p_{\Gamma\Delta}$. We will derive a lower bound for the improvement of an optimal estimator when it is given access to correlated auxiliary noise. Our primary tool for this analysis will be information-theoretic guarantees for the performance of an optimal estimator, initially described in Refs. [128, 129] and summarized in the following theorem [130]:

Theorem 7. (Feder, 1994) Define the piecewise function $\phi_N : [0, 1] \rightarrow \mathbb{R}$ and $\Phi_N : [0, 1] \rightarrow \mathbb{R}$ as

$$\phi_N(\lambda) = \begin{cases} a_k(\lambda - \frac{k-1}{k}) + \log(k), & \frac{k-1}{k} \leq \lambda \leq \frac{k}{k+1} \end{cases} \quad (4.38)$$

$$\Phi_N(\lambda) = h_2(\lambda) + \lambda \log(N-1) \quad (4.39)$$

with $a_k = k(k+1) \log((k+1)/k)$ for $k = 1 \dots N$. Let X be a random variable taking values in alphabet $\mathcal{X}$ with size $|\mathcal{X}| = N$, and let Y be a random variable taking values in alphabet $\mathcal{Y}$. Define the error probability of an optimal estimator as $p_e^*(X|Y) := \min_{\hat{X}: \mathcal{Y} \rightarrow \mathcal{X}} \Pr(\hat{X}(Y) \neq X)$. Then,

$$\Phi_N(p_e^*(X|Y)) \geq H(X|Y) \geq \phi_N(p_e^*(X|Y)) \quad (4.40)$$

Corollary 8. For binary variable $X \in \{0, 1\}$ and any variable Y taking values in alphabet $\mathcal{Y}$, the error of the optimal estimator for X given Y satisfies

$$\min_{\hat{X}: \mathcal{Y} \rightarrow \{0,1\}} \Pr(\hat{X} \neq X) \leq \frac{1}{2} H(X|Y) \quad (4.41)$$

These results mean that the (optimal) performance of hypothesis testing using correlated auxiliary noise is related to the uncertainty $H(X|Z\Delta)$. Consequently, Proposition 2 allows us to lower bound the improvement of estimation in the presence of correlated noise. Let $\hat{X}^c := \hat{X}^c(Z, \Delta)$ be an optimal estimator for X (given correlated auxiliary noise) with error

$$p_e^c(X|Z\Delta) = \Pr(\hat{X}^c \neq X) \quad (4.42)$$

and let $\hat{X} = \hat{X}(Z)$ (with no correlated noise) be an optimal estimator with error $p_e(X|Z) = \Pr(\hat{X} \neq X)$. Clearly, any estimator $\hat{X}(Z)$ has at least error $p_\Gamma(1)$. Combining Corollary 8 with Proposition 2, the prediction error for an optimal $\hat{X}(Z, \Delta)$ is upper bounded as

$$p_e^c(X|Z\Delta) \leq \frac{1}{2} H(X|Z\Delta) \leq \frac{1}{2} H(\Gamma|\Delta). \quad (4.43)$$

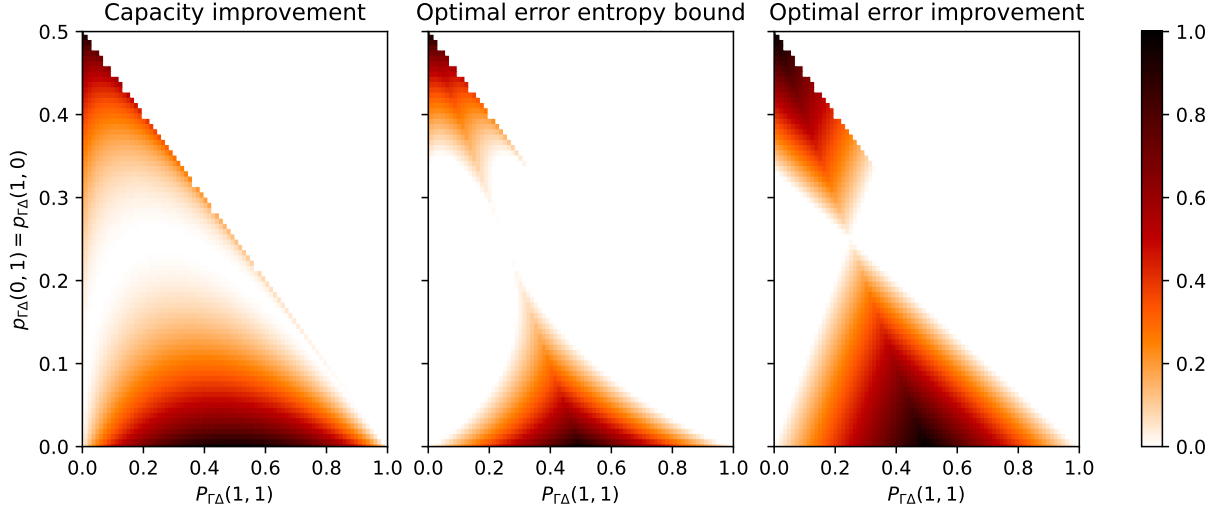


Figure 4.2: Improvement in channel capacity and improvement (bounds) for estimation tasks with and without correlated auxiliary noise when the distribution $p_{\Gamma\Delta}$ is symmetric. **a** The capacity improvement resulting from access to correlated auxiliary noise is greatest for (anti-)correlated variables (Γ, Δ) . **b** The information-theoretic lower bound for improving estimator accuracy (Eq. 4.44) identifies regions where correlated auxiliary noise is beneficial. **c** The true improvement for optimal estimator accuracy using correlated noise shows that correlations in the joint distribution $p_{\Gamma\Delta}$ are not sufficient for an improvement: The correlations must be of a particular form to be useful.

Thus, in an optimal scenario the improvement in estimation performance given correlated noise $p_e(X|Z) - p_e^c(X|Z\Delta)$ is at least

$$\max_{\hat{X}^c} (p_e(X|Z) - p_e^c(X|Z\Delta)) \geq p_\Gamma(1) - \frac{1}{2} H(\Gamma|\Delta) \quad (4.44)$$

Figure 4.2b plots this bound for the case when $p_{\Gamma\Delta}(0, 1) = p_{\Gamma\Delta}(1, 0)$, and provides comparison to the true difference in performance for optimal strategies shown in Fig. 4.2c.

4.4.3 Accuracy of the optimal decoder

For completeness, we derive the optimal decoder for a bit flip channel with correlated auxiliary noise, and prove that its error is never greater than the optimal scheme without

auxiliary noise. We define a decoder that takes as input the pair (Z, Δ) and outputs an estimate $\hat{X}^c(Z, \Delta)$ for X using knowledge of the joint distribution $q(\gamma, \delta) := \Pr(\Gamma = \gamma, \Delta = \delta)$. For our choice of $Z = X \oplus \Gamma$,

$$\Pr(X = x|Z = z, \Delta = \delta) = \Pr(\Gamma = x \oplus z, \Delta = \delta), \quad (4.45)$$

resulting in an optimal decoder

$$\hat{X}^c(z, \delta) = \begin{cases} z, & \Pr(\Gamma = 0|\Delta = \delta) \geq \Pr(\Gamma = 1|\Delta = \delta), \\ z \oplus 1, & \text{otherwise} \end{cases}. \quad (4.46)$$

The error of this decoder is then

$$\Pr_{Z, \Delta, X}(\hat{X}^c(Z, \Delta) \neq X) \quad (4.47)$$

$$= \mathbb{E}_{z, \delta} \min [\Pr(X = 1|Z = z, \Delta = \delta), 1 - \Pr(X = 1|Z = z, \Delta = \delta)] \quad (4.48)$$

$$= \sum_{z, \delta} \min [\Pr(Z = z, \Delta = \delta|X = 1)p_X(1), \Pr(Z = z, \Delta = \delta|X = 0)p_X(0)] \quad (4.49)$$

$$= \min [q(1, 0)p_X(1), q(0, 0)p_X(0)] + \min [q(1, 1)p_X(1), q(0, 1)p_X(0)] \quad (4.50)$$

$$+ \min [q(1, 0)p_X(0), q(0, 0)p_X(1)] + \min [q(1, 1)p_X(0), q(0, 1)p_X(1)]$$

$$\leq \min [(q(1, 0) + q(1, 1))p_X(1), (q(0, 0) + q(0, 1))p_X(0)] \quad (4.51)$$

$$+ \min [(q(1, 0) + q(1, 1))p_X(0), (q(0, 0) + q(0, 1))p_X(1)]$$

$$= \min [p_\Gamma(1)p_X(1), p_\Gamma(0)p_X(0)] + \min [p_\Gamma(1)p_X(0), p_\Gamma(0)p_X(1)] \quad (4.52)$$

$$\leq \Pr(\hat{X}(Z) \neq X) \quad (4.53)$$

for any $\hat{X} : \{0, 1\} \rightarrow \{0, 1\}$, where Line 4.51 applies the following inequality for scalars a, b, c, d, x, y :

$$\min(ax, by) + \min(cx, dy) \leq \min((a + c)x, (b + d)y), \quad (4.54)$$

and Line 4.53 holds since the previous expression is the optimal error rate for any decoder with access only to Z . And so we find, Eq. 4.54. If we choose $q(\gamma, \delta) = p_\Gamma(\gamma)p_\Delta(\delta)$, we may factor out every p_Δ term appearing in Line 4.51 to recover strict equality. This implies that the error of $\hat{X}^c(Z, \Delta)$ is always less than the error of $\hat{X}(Z)$ with equality if Γ and Δ are independent.

Conversely, correlations in $p_{\Gamma\Delta}$ are *not* sufficient for an improvement in error rate. A simple counter example is $q(0, 1) = q(1, 0) = 1/4$, $q(0, 0) = 1/2$, $q(1, 1) = 0$ which may not be expressed as a product distribution over (Γ, Δ) but nevertheless results in the same error as a decoder without correlated auxiliary noise. Intuitively, this is because the bits Γ and Δ are correlated but agree or disagree with the same probability.

4.4.4 Security of one-time pads from correlated auxiliary noise

The one-time pad (OTP) is an encryption scheme in which a sender encrypts their message by performing modular addition with a perfectly random, secret key. The sender may then perfectly recover the original message by using the same key. From this perspective, the above analysis may be understood as characterizing an OTP in which the assumption of a secret key is violated. For instance, if X represents the message and Γ represents a secret key (both assumed to be unbiased coins), an eavesdropper may attempt to recover X from Z and some correlated noise Δ . As the above results show, any amount of correlation between Γ and Δ may allow the eavesdropper to recover X with better-than-random accuracy. Our framework therefore characterizes, analytically, the degradation of a OTP scheme when an eavesdropper has access to correlations with the sender's secret key, e.g. through mishandling of private keys or information leakage in a physical system.

4.5 Hypothesis testing bounds on QEC decoders

We now show that there is a relationship between hypothesis testing and the decoding of a quantum error correcting code (QECC). We will demonstrate this connection by showing that the syndrome of a quantum error-correcting code is related to the identity of the error that occurred through a hypothesis testing characterization.

4.5.1 Example: 3-qubit code

We first consider the 3-qubit repetition code for correcting bit-flip errors to demonstrate a simple relationship between correlated auxiliary noise and error correction. In this code, a one-qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is encoded into a logical state $|\psi_L\rangle := \alpha|000\rangle + \beta|111\rangle$. Each qubit is then subjected to a bit flip channel $\mathcal{N}(p) : \rho \rightarrow (1-p)\rho + X\rho X$ where X is the Pauli-X operator, which can be understood as a bit flip operation being applied with probability p . Taken together, the channel $\mathcal{N}(p)^{\otimes 3}$ stochastically applies one of 8 possible operations to $|\psi_L\rangle$. Let the random variable $E = (E_1, E_2, E_3)$ taking values in $\{0, 1\}^3$ label each of these possible errors, such that the effect of the channel becomes

$$|\psi_L\rangle \rightarrow (X^{E_1} \otimes X^{E_2} \otimes X^{E_3}) |\psi_L\rangle. \quad (4.55)$$

After the error occurs, a set of operations are performed on two ancilla qubits to try to detect which error occurred so that a correction may be applied to $|\psi_L\rangle$. In this code, a pair

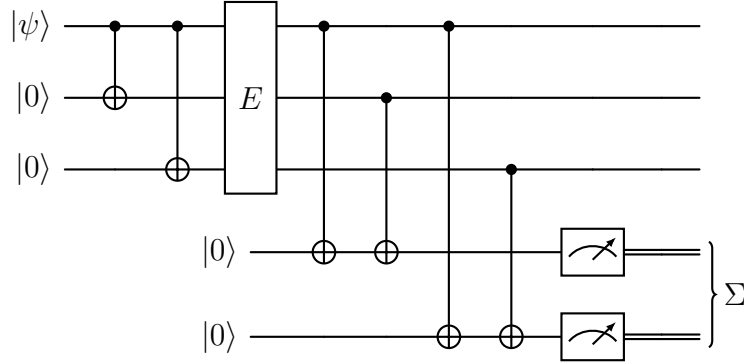


Figure 4.3: A circuit depicting the three-qubit code for correcting bit flip errors. E represents one of eight possible errors involving a bit flip on each qubit with probability p , while a pair of bits Σ describe the parity of encoded qubits and allow for a single bit flip to be corrected.

of syndrome bits are described by a random variable $\Sigma \in \{0, 1\}^2$ generated by performing parity checks on pairs of qubits in $|\psi_L\rangle$ (see Fig. 4.3). In this simplified example, if we can determine the value of $E = (E_1, E_2, E_3)$ from the syndrome Σ then we may reverse the effect of the noisy channel by applying $X_i^{E_i}$ to each qubit $i = 1, 2, 3$. In other words, correctly estimating the value of E is sufficient to perform error correction in this setting (though this is not a *necessary* condition for success, as we will discuss in the next section). Define the optimal probability of successfully identifying the error random variable E given the syndrome random variable Σ as

$$p^*(E|\Sigma) = \max_{\hat{E}: \mathcal{S} \rightarrow \mathcal{E}(\Sigma)} \Pr(\hat{E}(\Sigma) = E), \quad (4.56)$$

where the probability is with respect to the joint distribution p_{ES} of errors and syndromes. Then using Theorem 7, the success probability for the nondegenerate 3-qubit code obeys the bounds

$$\frac{1}{2} H(E|\Sigma) \geq p^*(E|\Sigma) \quad (4.57)$$

$$H(E|\Sigma) \leq h_2(p^*(E|\Sigma)) + p^*(E|\Sigma) \cdot \log 7, \quad (4.58)$$

where the latter inequality may be used to bound $p^*(E|\Sigma)$ implicitly. Figures 4.4-4.5 show the behavior of Eqs. 4.57-4.58 as a function of the probability p of a bit flip occurring on each qubit. We also provide the true success probability for the 3-qubit code, which is equal to $p^*(E|\Sigma)$.

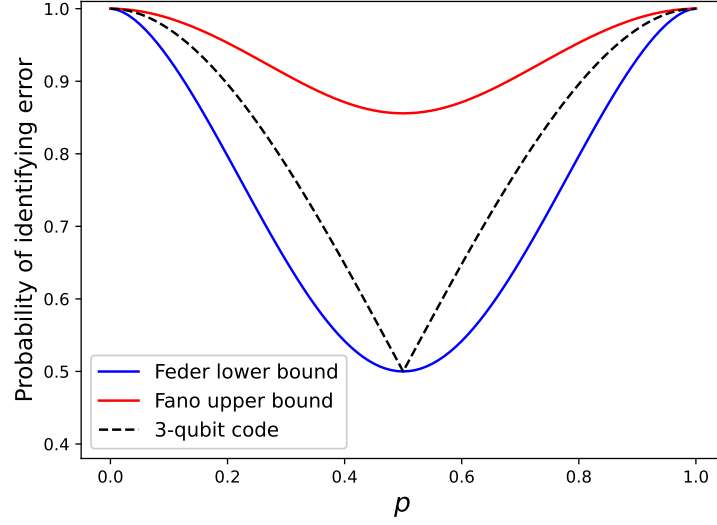


Figure 4.4: Information-theoretic upper (Eq. 4.57) and lower (Eq. 4.58) bounds for succeeding at QEC when treating the syndrome as a correlated random variable with the error event. The performance of the optimal decoder - the standard decoder for the 3-qubit code with success probability $p^*(E|\Sigma)$ - is also shown, providing indication of the looseness of the information-theoretic bounds.

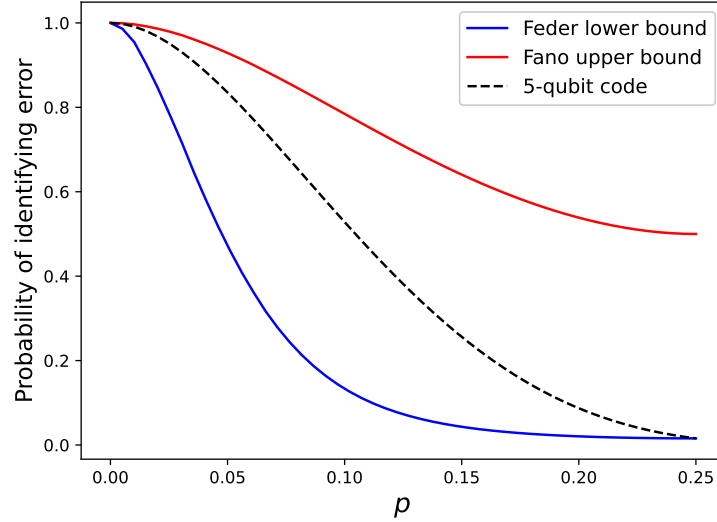


Figure 4.5: Corresponding bounds for the 5-qubit code [133, 134] for single-qubit depolarizing channels depolarization probability $4p/3$.

To expand on this toy model, for a random variable E taking values in $\mathcal{E}$ with probability p_E , and a resulting syndrome Σ with probability $p_{\Sigma|E}(\sigma|e)$, we assume the syndrome depends deterministically on the error such that $\Sigma = f(E)$ for some $f : \mathcal{E} \rightarrow \mathcal{S}$. Then, to successfully correct for an error E , it is sufficient (though not necessary) to determine the value of error E given a syndrome Σ . However, the Knill-Laflamme conditions for quantum error correction mean that this is not always possible [135]. We will have a probability for failure whenever two or more errors map to the same syndrome. Thus, syndrome values σ will be distributed according to a probability

$$p_{\Sigma}(\sigma) = \sum_{\substack{e \in \mathcal{E}: \\ f(e) = \sigma}} p_E(e) \quad (4.59)$$

where the sum is over all errors e giving rise to the same syndrome. Once the probability distribution of syndrome events is calculated, the conditional entropy of the error event given a syndrome is computed as

$$H(E|\Sigma) = H(E, \Sigma) - H(\Sigma) \quad (4.60)$$

$$= H(E) - H(\Sigma) \quad (4.61)$$

where the second equality follows since Σ is a function of E . However, in the general case it is not *necessary* to determine which error E occurred, since many (but not all!) errors that give rise to the same syndrome will also be corrected by the same recovery operation. However, it is straightforward to extend the above ideas to the setting of, e.g. *stabilizer codes* wherein the goal of decoding is to determine the label of an equivalence class of errors given a syndrome.

4.6 Correlated auxiliary noise for improving quantum channel capacity

We will now develop a model for correlated auxiliary noise in quantum channels and demonstrate how this resource can improve quantum channel capacity. Every quantum channel acting on a system may be described by the system interacting with an environment according to a unitary operation, followed by discarding the environment. For any channel $\mathcal{N}$ acting on states $\rho \in \mathcal{D}(\mathcal{H}_B)$, there is an environment $\mathcal{H}_E$ and a unitary⁴ $U \in \mathcal{U}(\mathcal{H}_{AE}, \mathcal{H}_{BE})$

⁴Note when an isometry $U : \mathcal{L}(\mathcal{H}_{AE}) \rightarrow \mathcal{L}(\mathcal{H}_{BE})$ is unitary, we require $\mathcal{H}_A \simeq \mathcal{H}_B$ but distinguish the co-domain and domain of U for clarity.

such that [136],

$$\mathcal{N}(\rho) = \text{Tr}_E \left(U_{AE}(\rho \otimes |\psi\rangle\langle\psi|_E) U_{AE}^\dagger \right). \quad (4.62)$$

A natural model for quantum correlated auxiliary noise involves modifying the output of the channel $\mathcal{N}$ to contain an auxiliary system that is correlated with the subsystem of the environment interacting with ρ . We will therefore consider an environment $\mathcal{H}_E = \mathcal{H}_{\Gamma\Delta}$ that is initially prepared in a state $\psi_{\Gamma\Delta} \in \mathcal{D}(\mathcal{H}_{\Gamma\Delta})$, such that the subsystem $\mathcal{H}_\Gamma$ interacts with $\mathcal{H}_B$ directly while the output of the channel contains $\mathcal{H}_\Delta$. The action of the channel with correlated auxiliary noise $\mathcal{M} : \mathcal{L}(\mathcal{H}_A) \rightarrow \mathcal{L}(\mathcal{H}_{B\Delta})$ is described by the map

$$\mathcal{M}(\rho) = \text{Tr}_\Gamma \left((U_{A\Gamma} \otimes \mathbb{I}_\Delta)(\rho \otimes \psi_{\Gamma\Delta})(U_{A\Gamma}^\dagger \otimes \mathbb{I}_\Delta) \right). \quad (4.63)$$

The choice of U giving rise to the channel $\mathcal{N}$ of Eq. 4.62 is not unique, and so the channel $\mathcal{M}$ must be defined with respect to a specific choice of U that can be justified based on, e.g., knowledge of the environmental interaction Hamiltonian giving rise to $\mathcal{N}$ in an experimental setting.

Several prior works have studied settings involving (partial) preparation or measurement of the environment generating a quantum channel. For instance, when the recipient of $\mathcal{N}(\rho)$ is allowed to projectively measure the entire environment, then they may recover the original transmitted state ρ if and only if the channel is a mixed unitary channel [125]. Previous works have characterized the classical [137] and quantum [138] environment-assisted quantum capacity, wherein the state of the environment is fixed (or optimized) and arbitrary measurements are allowed on the entire environment. However, we consider a single-use setting in which the environment is inaccessible before/during use of the channel. Refs. [20, 126] characterized a family of channels like Eq. 4.63, deriving formulae for the capacity of such channels optimized with respect to $|\psi\rangle_{\Gamma\Delta}$. Here we complement this prior work by deriving the quantum capacity of the channel with respect to *any* choice of correlated state $|\psi\rangle_{\Gamma\Delta}$, allowing us to identify the properties of such an entangled state that lead to improvements in the transmission of quantum information.

We briefly review quantum communication theory. Given an finite set $\mathcal{X}$, we associate a system X with the complex Euclidean space $\mathcal{H}_X = \mathbb{C}^{|\mathcal{X}|}$ containing states of the form $\rho \in \mathcal{D}(\mathcal{H}_X)$. For a given spectral decomposition $\rho = \sum_{x \in \mathcal{X}} \lambda p_x |x\rangle\langle x|$, the *von Neumann entropy* of system X in state ρ is defined

$$\mathcal{H}(X)_\rho := - \sum_x p_x \log p_x. \quad (4.64)$$

An important difference between this entropy and Shannon entropy is that the von Neumann entropy is only defined at one “time slice” of a quantum system, as reflected by the

need to specify the state of the system via subscript. This differs from Shannon entropy, which allows computation of entropies of the form $H(AB)$ even when $p_B = f(p_A)$, and reflects a fundamental difference between quantum and classical information. Eq. 4.64 leads immediately to the fact that the entropies of the two bipartitions of a pure state are equal, i.e. for $\sigma \in D(\mathcal{H}_{AB})$ and $\sigma_B := \text{Tr}_A \sigma$ and $\sigma_A := \text{Tr}_B \sigma$, then if σ is pure (defined by $\text{Tr}(\sigma^2) = 1$) we have

$$H(A)_{\sigma_A} = H(B)_{\sigma_B}. \quad (4.65)$$

The conditional entropy of a state is defined analogously to the classical case,

$$H(A|B)_\sigma := H(AB)_\sigma - H(B)_{\sigma_B}. \quad (4.66)$$

Importantly, $H(A|B)_\sigma$ may be negative. For a channel $\mathcal{N} \in C(\mathcal{H}_A, \mathcal{H}_B)$ with a Stinespring dilation of the form of Eq. 4.62, the *complementary channel* $\mathcal{N}^c \in C(\mathcal{H}_A, \mathcal{H}_E)$ is defined as

$$\mathcal{N}^c(\rho) = \text{Tr}_A(U_{AE}(\rho \otimes \mathbb{I}_E)U_{AE}^\dagger). \quad (4.67)$$

Recall that the quantum capacity $Q(\mathcal{N})$ of a channel $\mathcal{N}$ characterizes how well quantum information may be transmitted in an asymptotic setting, and is defined according to [139–142] :

$$Q(\mathcal{N}) = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{\rho} I_c(\rho, \mathcal{N}^{\otimes n}), \quad (4.68)$$

where I_c denotes the coherent information of $\mathcal{N}$:

$$I_c(\rho, \mathcal{N}) = H(B)_{\mathcal{N}(\rho)} - H(E)_{\mathcal{N}^c(\rho)}. \quad (4.69)$$

For certain channels, the formula for quantum capacity may be greatly simplified. A channel is then *degradable* if there exists a degrading channel $\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_E)$ such that $\mathcal{N}^c = \mathcal{D} \circ \mathcal{N}$, so that $\mathcal{D}$ acting on the output of the channel reproduces the final state. A channel is *antidegradable* if there exists a channel $\mathcal{A} \in C(\mathcal{H}_E, \mathcal{H}_B)$ such that $\mathcal{N} = \mathcal{A} \circ \mathcal{N}^c$, so that $\mathcal{A}$ produces the final state of the input system given the final state of the environment. If $\mathcal{N}$ is antidegradable then it is straightforward to show that $Q(\mathcal{N}) = 0$ via information-theoretic arguments⁵. On the other hand, if $\mathcal{N}$ is degradable then the quantum capacity admits a single-letter formula, i.e. $Q(\mathcal{N}) = I_c(\rho, \mathcal{N})$ [144]. In the remainder of this chapter, we will discuss characterizing the quantum capacity of quantum channels with access to correlated auxiliary noise.

⁵It is less clear how this fact relates to the no-cloning theorem, c.f. Ref. [143]

4.6.1 The capacity of quantum bit flip channel with correlated auxiliary noise

In this section, we derive the quantum capacity of a quantum bit flip channel given access to correlated auxiliary noise. In Section 4.4.2 it was demonstrated that the capacity of a classical channel improves when the receiver has access to an auxiliary variable that is classically correlated with the variable describing the environment. In the case of transmitting quantum information, it is natural to consider a scenario where the receiver has access to a system that contains quantum correlations (i.e. entanglement) with the environment.

We first review the quantum bit flip channel acting on a single qubit. The channel $\mathcal{N}(\rho) = (1-p)\rho + pX\rho X$ introduces a bit flip (in the computational basis) with probability p , and may be written in the form of Eq. 4.62 as

$$\mathcal{N}(\rho) = \text{Tr}_E (U(\rho \otimes |\psi\rangle\langle\psi|_E)U^\dagger). \quad (4.70)$$

where $U := \text{CNOT}_{EA} \in \text{U}(\mathcal{H}_{AE}, \mathcal{H}_{AB})$ is the Controlled-NOT gate acting as $\text{CNOT}_{AB} := (\mathbb{I}_B \otimes |0\rangle\langle 0|_A + X_B \otimes |1\rangle\langle 1|_A$, and $|\psi\rangle = \sqrt{1-p}|0\rangle + \sqrt{p}|1\rangle$. This channel has quantum capacity

$$Q(\mathcal{N}) = 1 - h_2(p). \quad (4.71)$$

Note that this channel has a classical capacity of 1 bit, since the sender may transmit classical information perfectly using states encoded in the eigenbasis of X . Thus, there is no potential for improving the classical capacity by introducing correlated auxiliary noise.

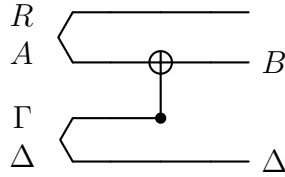


Figure 4.6: Circuit for a quantum bit flip channel with correlated auxiliary noise.

We now introduce a quantum bit flip channel with correlated auxiliary noise. This channel is constructed by extending the state $|\psi\rangle_\Gamma$ of Eq. 4.70 to $\mathcal{H}_{\Gamma\Delta}$ and then providing the state of $\mathcal{H}_\Delta$ to the receiver, i.e. we define a channel of the form

$$\mathcal{M}(\rho) = \text{Tr}_\Gamma (\text{CNOT}_{\Gamma A}(\rho \otimes \psi_{\Gamma\Delta})\text{CNOT}_{\Gamma A}), \quad (4.72)$$

$$|\psi\rangle_{\Gamma\Delta} = \sum_{i \in \{0,1\}} \lambda_i |v_i\rangle |w_i\rangle, \quad (4.73)$$

where $\lambda_i \in \mathbb{R}$ and $\{|v_i\rangle\}_i$ and $\{|w_i\rangle\}_i$ orthonormal systems on $\mathcal{H}_\Gamma$ and $\mathcal{H}_\Delta$ respectively. This channel is depicted in Fig. 4.6. Choosing a purification of $\rho \in \mathcal{D}(\mathcal{H}_A)$ to be $|\phi_{RA}\rangle$ for some space $\mathcal{H}_R$, the initial state of the complete system is

$$\rho_{RA\Gamma\Delta} = |\phi_{RA}\rangle\langle\phi_{RA}| \otimes |\psi\rangle\langle\psi|_{\Gamma\Delta}, \quad (4.74)$$

where we have assumed that the system $\mathcal{H}_{\Gamma\Delta}$ is initially in a pure state in Schmidt form as given in Eq. 4.73. We may describe the evolution of the initial state of $\mathcal{H}_A$ into a state of $\mathcal{H}_{B\Delta}$ according to the map

$$\mathcal{M}(\rho_A) = \text{Tr}_\Gamma (\text{CNOT}_{\Gamma A}(\rho_A \otimes |\psi_{\Gamma\Delta}\rangle\langle\psi_{\Gamma\Delta}|) \text{CNOT}_{\Gamma A}) \quad (4.75)$$

Writing $|v_i\rangle = c_{0i}|0\rangle + c_{1i}|1\rangle$, then the effect of applying $\mathcal{M}$ to an initial $\rho \in \mathcal{D}(\mathcal{H}_A)$ is

$$\mathcal{M}(\rho) = \text{Tr}_\Gamma \text{CNOT}_{\Gamma A} \left(\rho \otimes \sum_{i,j \in \{0,1\}} \lambda_i \lambda_j^* |v_i\rangle\langle v_j| \otimes |w_i\rangle\langle w_j| \right) \text{CNOT}_{\Gamma A} \quad (4.76)$$

$$= \text{Tr}_\Gamma \left(\sum_{i,j,k,\ell \in \{0,1\}} \lambda_i \lambda_j^* c_{ki} c_{\ell j}^* X^k \rho X^\ell \otimes |k\rangle\langle\ell| \otimes |w_i\rangle\langle w_j| \right) \quad (4.77)$$

$$= \sum_{i,j,k \in \{0,1\}} \lambda_i \lambda_j^* c_{ki} c_{kj}^* X^k \rho X^k \otimes |w_i\rangle\langle w_j| \quad (4.78)$$

$$= \rho \otimes \sum_{i,j \in \{0,1\}} \lambda_i \lambda_j^* c_{0i} c_{0j}^* |w_i\rangle\langle w_j| + X \rho X \otimes \sum_{i,j \in \{0,1\}} \lambda_i \lambda_j^* c_{1i} c_{1j}^* |w_i\rangle\langle w_j| \quad (4.79)$$

$$= q_0 (\rho \otimes |s_0\rangle\langle s_0|) + q_1 (X \rho X \otimes |s_1\rangle\langle s_1|), \quad (4.80)$$

where $|s_k\rangle = q_k^{-1/2} \sum_{i \in \{0,1\}} \lambda_i c_{ki} |w_i\rangle$ is a normalized state with inverse prefactor

$$q_k = \sum_i \lambda_i^2 |c_{ki}|^2. \quad (4.81)$$

Here, q_1 describes the total probability of a bit flip occurring. By the Schmidt decomposition and definition of $|v_i\rangle$ we have that $q_0 + q_1 = 1$, meaning that line 4.80 is a valid state. Evidently, if $\langle s_0 | s_1 \rangle = 0$ then ρ_B is perfectly recoverable by applying a controlled- X operation to $\mathcal{H}_B$ and local operations on $\mathcal{H}_\Delta$, in which case $\mathcal{M}$ may be represented as a mixed unitary channel as characterized in Ref. [125].

To derive the capacity of this channel, we will instead consider a dephasing channel with correlated auxiliary noise (equivalent to the previously discussed channel up to local

unitaries which cannot affect the capacity). This channel acts according to a Stinespring representation $\mathcal{M}(\rho) = \text{Tr}_\Gamma(V\rho V^\dagger)$ for an isometry $V \in \text{U}(\mathcal{H}_A, \mathcal{H}_{B\Delta})$,

$$V|i\rangle = |i\rangle \otimes \sum_{j,k \in \{0,1\}} \lambda_j c_{kj} (-1)^{ik} |k\rangle \otimes |w_j\rangle \quad (4.82)$$

for a state $|\psi\rangle_{\Gamma\Delta}$ introduced in Eq. 4.73. From Eq. 4.82 we explicitly calculate

$$\mathcal{M}^c(\rho) = \begin{pmatrix} q_0 & \gamma(\rho_{00} - \rho_{11}) \\ \gamma^*(\rho_{00} - \rho_{11}) & q_1 \end{pmatrix}, \quad (4.83)$$

where $\gamma := \sum_{k \in \{0,1\}} \lambda_k^2 c_{0k} c_{1k}^*$ and $\rho_{ij} = \langle i|\rho|j\rangle$ are components of the input density matrix. Letting Ψ be the maximally dephasing channel $\Psi(\rho) = \sum_i \rho_{ii} |i\rangle\langle i|$, then evidently

$$\mathcal{M}^c = \mathcal{M}^c \circ \Psi \circ \text{Tr}_\Delta \circ \mathcal{M} \quad (4.84)$$

and so $\mathcal{M}$ is degradable with degrading channel $\mathcal{M}^c \circ \Psi \circ \text{Tr}_\Delta$. We note that Eq. 4.80 is in the form of a *flagged* quantum channel characterized in Ref. [145], though their sufficient conditions for the degradability of such a channel are not useful here. Since $\mathcal{M}$ is equivalent to the quantum bit flip channel with correlated noise up to unitary operations acting on B (and the maximization of Eq. 4.69 is unaffected by local unitaries on B), this implies that the quantum bit flip channel with correlated noise and resource state $|\psi\rangle_{\Delta\Gamma}$ is degradable.

Following Refs. [20, 146, 147], the coherent information of this channel may be maximized using states that are diagonal in the eigenbasis of X , an argument which we restate here for completeness. From Eq. 4.83 it follows that $\mathcal{M}^c(Z\rho Z) = \mathcal{M}^c(\rho)$, and so $H(\mathcal{M}^c(Z\rho Z)) = H(\mathcal{M}^c(\rho))$. Meanwhile, $\mathcal{M}(Z\rho Z) = (Z \otimes \mathbb{I}_\Delta)\mathcal{M}(\rho)(Z \otimes I_\Delta)$. Since H is invariant up to unitary operations, this means

$$I_c(Z\rho Z, \mathcal{M}) = I_c(\rho, \mathcal{M}). \quad (4.85)$$

By concavity of the coherent information of degradable channels [144], it follows that

$$I_c\left(\frac{1}{2}(\rho + Z\rho Z), \mathcal{M}\right) \geq \frac{1}{2}[I_c(\rho, \mathcal{M}) + I_c(Z\rho Z, \mathcal{M})] = I_c(\rho, \mathcal{M}). \quad (4.86)$$

This means that the coherent information of the channel with respect to any state ρ is no larger than that with respect to $\Psi(\rho)$, and is therefore maximized by considering only diagonal input states of the form $x|0\rangle\langle 0| + (1-x)|1\rangle\langle 1|$ (or equivalently, an entangled reference state $x^{1/2}|00\rangle + (1-x)^{1/2}|11\rangle$ on the system RB).

Substituting an input state $\sigma = x|+\rangle\langle+| + (1-x)|-\rangle\langle-|$ diagonal in the eigenbasis of X , the entropy of the output of the channel (Eq. 4.80) becomes

$$H(\mathcal{M}(\sigma)) = H(\sigma \otimes (q_0|s_0\rangle\langle s_0| + q_1|s_1\rangle\langle s_1|)) \quad (4.87)$$

$$= h_2(x) + h_2(\lambda_0^2). \quad (4.88)$$

Meanwhile, the eigenvalues of $\mathcal{M}^c(\sigma)$ are given by

$$\lambda_{\pm}(x) = \frac{1}{2} \left[1 \pm (1 - 4q_0(1 - q_0) + 4|\gamma|^2(2x - 1)^2)^{1/2} \right], \quad (4.89)$$

which depends on both the input state and $|\psi\rangle_{\Gamma\Delta}$. The capacity of the channel is therefore the difference of the previous two equations,

$$Q(\mathcal{M}) = \max_{x \in [0,1]} h_2(x) + h_2(\lambda_0^2) - h_2(\lambda_+(x)). \quad (4.90)$$

The above relation depends on both the input state and $|\psi\rangle_{\Gamma\Delta}$. This formula recovers several expected results: If $|\psi\rangle_{\Gamma\Delta}$ is maximally entangled, $Q(\mathcal{M}) = 1$. This follows since maximal entanglement for $|\psi\rangle$ implies $\lambda_0 = \lambda_1 = 1/\sqrt{2}$. In this case, $q_0 = 1/2$ and the maximum in Eq. 4.90 is achieved by $x = 1/2$. Thus, having a maximal entanglement between the noise environment Γ and the auxiliary environment Δ allows us to perfectly transmit quantum information through the quantum bit flip channel with CAN. Moreover, if $|\psi\rangle_{\Gamma\Delta}$ is separable, we recover $q_1 \in \{|c_{10}|^2, |c_{11}|^2\}$ (both of which represent the probability of a bit flip being applied) and so $Q(\mathcal{M}) = 1 - h_2(q_1)$. This is the quantum capacity for a standard bit flip channel, which is what we expect.

4.6.2 Towards information-theoretic bounds on the helpfulness of UCAN

In the previous section, we introduced a version of the quantum bit flip channel with correlated auxiliary noise that can potentially improve the quantum capacity of the channel. We will now explore how significantly correlated auxiliary noise can improve channel capacities. Recalling Eq. 4.63, we will fix a unitary interaction $U \in \mathcal{U}(\mathcal{H}_{A\Gamma})$ between a system A and an environment Γ , and also fix a state $|\psi\rangle_{\Gamma\Delta} \in \mathcal{H}_{\Gamma\Delta}$. The pair $(U, \psi_{\Gamma\Delta})$ gives rise to a channel with correlated noise $\mathcal{M} \in \mathcal{C}(\mathcal{H}_A, \mathcal{H}_{B\Delta})$ and a channel without correlated noise, $\mathcal{N} \in \mathcal{C}(\mathcal{H}_A, \mathcal{H}_B)$ defined as

$$\mathcal{M}(\rho) = \text{Tr}_{\Gamma}((U \otimes \mathbb{I}_{\Delta})(\rho \otimes |\psi\rangle\langle\psi|_{\Gamma\Delta})(U^{\dagger} \otimes \mathbb{I}_{\Delta})) \quad (4.91)$$

$$\mathcal{N}(\rho) = \text{Tr}_{\Gamma}(U(\rho \otimes \sigma_{\Gamma})U^{\dagger}) \quad (4.92)$$

where $\sigma_\Gamma = \text{Tr}_\Delta |\psi\rangle\langle\psi|_{\Gamma\Delta}$ is the reduced state of Γ . One significant way in which correlated auxiliary noise might improve channel capacity is via *activation*: When $Q(\mathcal{N}) = 0$ while $Q(\mathcal{M}) > 0$. At the same time, we would like to be able to characterize the capacity of the channels that satisfy this property. We first note that $Q(\mathcal{M}) \geq Q(\mathcal{N})$, which follows since one may always discard the system Δ from the output of $\mathcal{M}$ and be left with the channel $\mathcal{N}$. Otherwise, we can observe several facts about the (anti)-degradability of these channels:

1. If $\mathcal{M}$ is antidegradable, then $\mathcal{N}$ is also antidegradable. This follows since, if $\mathcal{M}$ is antidegradable with an anti-degrading map $\mathcal{A}$ such that $\mathcal{M} = \mathcal{A} \circ \mathcal{M}$, then the anti-degrading map for $\mathcal{N}$ is $(\text{Tr}_\Delta \circ \mathcal{A})$.
2. If $\mathcal{N}$ is degradable, $\mathcal{M}$ is degradable (compare to Ref. [20, Lemma 20]). This follows since, if $\mathcal{D}$ is the degrading map for $\mathcal{N}$ such that $\mathcal{N}^c = \mathcal{D} \circ \mathcal{N}$, then the degrading map for $\mathcal{M}$ is given by $(\mathcal{D} \otimes \mathbb{I}_\Delta)$.

However, the reverse of either of the two statements above does not necessarily hold: In particular, if we can identify a case when $\mathcal{N}$ is antidegradable while $\mathcal{M}$ is not, then we will be able to demonstrate a scenario in which correlated auxiliary noise activates a given noise channel. The results of the previous section already provide an example of this: Letting $U = \mathbb{I}_B \otimes |0\rangle\langle 0|_\Gamma + X_B \otimes |1\rangle\langle 1|_\Gamma$ be a controlled- X gate and choosing $|\psi\rangle_{\Gamma\Delta} = (|00\rangle + |11\rangle)/\sqrt{2}$, we see that $\mathcal{N}$ is a kind of *maximally dephasing channel* with $Q(\mathcal{N}) = 0$, while $Q(\mathcal{M}) = 1$. But ideally, we could find a more general, information-theoretic characterization of this activation behavior.

The analysis in Sec. 4.6.1 yielded Eq. 4.90 for the the quantum capacity of the bit flip channel with correlated auxiliary noise, and quantum capacities have also been derived for a family of channels whose isometric extension is SWAP^γ [20]. However, the relationship between the quantum capacity and the correlations (or entanglement) present in the environment is obscure. We can compare Eq. 4.90 to the classical capacity of the classical bit flip channel with correlated auxiliary noise (Theorem 5), in which the channel capacity is directly related to the correlations between the environment noise and auxiliary noise.

The above comparison suggests that we might search for a characterization of the quantum capacity of a quantum channel with correlated auxiliary noise that depends explicitly on a measure of entanglement in $|\psi\rangle_{\Gamma\Delta} \in \mathcal{H}_{\Gamma\Delta}$, the state of the environment. The following lemma upper bounds the coherent information of an ordinarily antidegradable channel that has been supplemented with correlated auxiliary noise:

Lemma 9. Let $\mathcal{M} : \mathcal{L}(\mathcal{H}_A) \rightarrow \mathcal{L}(\mathcal{H}_{B\Delta})$ be a channel with unitary extension

$$\mathcal{M}(\rho) = \text{Tr}_\Gamma \left((U_{\mathcal{N}} \otimes \mathbb{I}_\Delta)(\rho \otimes |\psi\rangle\langle\psi|_{\Gamma\Delta})(U_{\mathcal{N}}^\dagger \otimes \mathbb{I}_\Delta) \right) \quad (4.93)$$

with $|\psi\rangle\langle\psi|_{\Gamma_0\Delta}$ chosen such that $\mathcal{N}(\rho) = \text{Tr}_\Delta(\mathcal{M}(\rho))$ is *antidegradable*. Let $|\phi\rangle_{RA}$ be a purification of $\rho \in \mathcal{D}(\mathcal{H}_A)$ and define

$$|\xi\rangle_{RB\Gamma\Delta} = (\mathbb{I}_R \otimes U_{\mathcal{N}} \otimes \mathbb{I}_\Delta)(|\phi\rangle_{RA} \otimes |\psi\rangle_{\Gamma\Delta}) \quad (4.94)$$

Then, the coherent information of $\mathcal{M}$ is bounded according to

$$I_c(R)_{B\Delta}_\xi \leq \frac{1}{2}[\text{H}(\Delta|\Gamma)_\xi + \text{H}(\Delta|B)_\xi]. \quad (4.95)$$

Proof. let $\mathcal{A} : \mathcal{L}(\mathcal{H}_\Gamma) \rightarrow \mathcal{L}(\mathcal{H}_{B'})$ be the antidegrading map for $\mathcal{N}$ such that $\mathcal{A} \circ \mathcal{N}^c = \mathcal{N}$. Define the state $\chi_{RB'} = (\mathbb{I}_R \otimes \mathcal{A} \circ \mathcal{N}^c)(\phi_{RA})$, where ϕ_{RA} is a purification of ρ . Then,

$$I_c(R)_{B\Delta}_\xi = \frac{1}{2}[I(R:B\Delta)_\xi - I(R:\Gamma)_\xi] \quad (4.96)$$

$$\leq \frac{1}{2}[I(R:B\Delta)_\xi - I(R:B')_\chi] \quad (4.97)$$

$$= \frac{1}{2}[\text{H}(B\Delta)_\xi - \text{H}(RB\Delta)_\xi - \text{H}(B')_\chi + \text{H}(RB')_\chi] \quad (4.98)$$

$$= \frac{1}{2}[\text{H}(B\Delta)_\xi - \text{H}(\Gamma)_\xi - \text{H}(B)_\xi + \text{H}(RB)_\xi] \quad (4.99)$$

$$= \frac{1}{2}[\text{H}(B\Delta)_\xi - \text{H}(\Gamma)_\xi - \text{H}(B)_\xi + \text{H}(\Gamma\Delta)_\xi] \quad (4.100)$$

$$= \frac{1}{2}[\text{H}(\Delta|B)_\xi + \text{H}(\Delta|\Gamma)_\xi]. \quad (4.101)$$

Line 4.96 uses the formula from [148, p. 11.6.6], and line 4.97 is the data processing inequality (Eq. 4.4). Line 4.99 uses antidegradability of $\mathcal{N}$, essentially

$$\text{H}(RB')_\chi = \text{H}(RB')_{(\mathbb{I}_R \otimes \mathcal{A} \circ \mathcal{N}^c)(\psi)} = \text{H}(RB)_{(\mathbb{I}_R \otimes \mathcal{N})(\psi)} = \text{H}(RB)_\xi \quad (4.102)$$

with a similar equality for $\text{H}(B')_\chi - \text{H}(B)_\xi$ following immediately. Lines 4.99-4.100 use the fact that entropy is equal across any bipartition of a pure state (Eq. 4.65). $\square$

This lemma relates the coherent information of an antidegradable channel that has been provided correlated auxiliary noise to the initial entanglement provided between the

noisy environment Γ and the system Δ containing correlated noise. However, we have yet to determine whether this provides a bound for the capacity of the channel, as the coherent information in this case may turn out to be nonadditive in general. Future work will be necessary to determine the relationship between the capacity of a quantum channel with correlated auxiliary noise and the amount of quantum correlations accessible to the receiver.

Chapter 5

Bounds and Guarantees for learning and entanglement

5.1 Chapter overview

As we saw in the previous chapter, information theory may be used to derive bounds for classical hypothesis testing. We will now discuss how information theory also provides tools to predict the performance of a learning algorithm on a given dataset. For instance, the accuracy of learning an unknown parameter can be upper bounded by reducing the learning task to hypothesis testing for a discrete random variable, with Fano's inequality then stating that a small conditional entropy between a learner's observations and the unknown parameter is necessary for successful estimation. In Sec. 5.3, we extend this relationship by demonstrating that a small conditional entropy is also sufficient for successful learning, thereby establishing an information-theoretic lower bound on the accuracy of a learner. This connection between information theory and learning suggests that we might similarly apply quantum information theory to characterize learning tasks involving quantum systems. Observing that the fidelity of a finite-dimensional quantum system with a maximally entangled state (the singlet fraction, introduced in Sec. 5.4) generalizes the success probability for estimating a discrete random variable, in Sec. 5.5 we introduce an entanglement manipulation task for infinite-dimensional quantum systems that similarly generalizes classical learning. In Sec. 5.6 we derive information-theoretic bounds for succeeding at this task in terms of the maximal singlet fraction of an appropriate finite-dimensional discretization. As classical learning is recovered as a special case of this task, our analysis suggests a deeper relationship at the interface of learning, entanglement, and

information.

The contents of this chapter are reproduced directly from Ref. [3], with minor adjustments to the structure of the contents. I was the sole author of these results.

5.2 Introduction

A key challenge in applying learning algorithms is knowing how much data is needed to successfully learn from observations. To this end, information theory plays an essential role in understanding the potential performance of a classical machine learning algorithm on some given dataset. As Shannon theory [115] is best suited to describing discrete variables, a typical approach to characterize the learning of an unknown parameter involves discretization. For instance, if we wish to characterize a learner’s ability to determine an unknown parameter to within ϵ accuracy, we might divide the parameter space into a set of non-overlapping balls with radius ϵ , and then consider the problem of determining which ball contains the unknown parameter. Fano’s inequality, which characterizes how much mutual information between the learner’s observations and the target ϵ -ball is necessary for success, then lower bounds the difficulty of this *hypothesis testing* task [127, 149]. This technique has been applied to establish lower bounds on the error of learning algorithms involving sparse vectors (e.g. *compressed sensing* [150–153]) for classical data gathered from both classical [154–156] and quantum [157] sources.

While information theory has been successfully used to lower bound learning error, it is rarely applied to guarantee the performance of a learning algorithm. In this chapter, we present such an information-theoretic guarantee: Specifically, we show that in a best-case scenario the performance of an optimal learner improves with decreasing conditional entropy between the given data and a particular discretization of the unknown parameter space. In other words, we study the *onset of learning* by using classical Shannon theory to show that there exists a sufficient amount of information contained in training data such that some unknown parameter may be learned to ϵ -accuracy.

Similarly to how Shannon theory characterizes correlations between random variables and provides bounds for hypothesis testing, quantum information theory characterizes correlations in quantum states and provides bounds for certain quantum information processing tasks. One quantity characterizing the quantum correlations contained in a state is its overlap with a maximally entangled state, maximized with respect to local operations. This quantity, the *maximal singlet fraction*, relates to the state’s distillable entanglement [158], entanglement of formation [159], negativity [160], and usefulness as a resource for teleportation [161]. Moreover, the maximal singlet fraction is an information-theoretic quantity

that reduces to the optimal success probability for hypothesis testing as a special case [162]. Given this relationship between quantum correlations in a finite-dimensional system and classical correlations among discrete random variables, might we find an entanglement manipulation task for continuous variable quantum systems that similarly generalizes the notion of classical learning? Moreover, given the established relationship between Shannon theory (for discrete variables) and classical learning (of continuous parameters), might we be able to use the finite-dimensional singlet fraction to characterize the ability to succeed at this continuous variable task?

In this chapter we consider such a task, study how it may be understood as a quantum generalization of learning, and demonstrate information-theoretic bounds and guarantees for success. While prior works have applied statistical learning theory [163–167] or information theory [86, 168, 169] to characterize error in quantum machine learning tasks, these prior tasks still involve extracting *classical* information from quantum (and/or classical) systems. Our work moves beyond the typical setting of quantum machine learning to consider an intrinsically quantum task: A learner receives part of an (infinite-dimensional) bipartite system, and then applies local operations to maximize the fidelity of the final state with a highly entangled state of the bipartite system. We derive upper and lower bounds on the learner’s error in this task for a best-case and worst-case initial state, respectively.

5.3 Bounds and guarantees for classical learning

An essential task in the theory of machine learning is to understand how well an algorithm might perform on some given learning problem. We now formalize the learning problem that we will study, before connecting classical learning to a more general quantum information processing task. We begin with preliminaries: Given a set $\mathcal{X}$, we define the set of probability densities $\mathcal{P}(\mathcal{X})$. Then, for a probability density $p \in \mathcal{P}(\mathcal{X})$ we let $\mathbb{E}_p[f(x)]$ denote the expected value of f with respect to p and define the *indicator function* $\mathbb{I}\{x \in S\}$ which is equal to one if $x \in S \subseteq \mathcal{X}$ and zero otherwise. We write the probability of some event $\omega \subseteq \mathcal{X}$ with respect to p as $\Pr_p(\omega) := \mathbb{E}_p[\mathbb{I}\{x \in \omega\}]$.

To define our learning problem, we assume that nature has chosen some fixed *target parameter* $\alpha \in A$, where A is a bounded subset of some metric space X with metric $d : X \times X \rightarrow \mathbb{R}^+$. We choose X to be a metric space so that we can judge estimates for α in terms of the distance d (for simplicity, the reader may consider X to be a normed vector space, e.g. $d(x, y) = \|x - y\|_2$ for $x, y \in A \subset \mathbb{R}^p$). We then assume that *observations* are described by a random variable B taking values in a set $\mathcal{B}$ according to probability $p_{B|A}$ conditional on the choice of α . We need not specify additional properties of $\mathcal{B}$ for now, but

as mentioned before the observations B might represent a training dataset for supervised machine learning, so that $\mathcal{B} = \mathcal{X}^n \times \mathcal{Y}^n$ is a sequence of n pairs of data points from $\mathcal{X}$ and labels from $\mathcal{Y}$ and $B = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Equivalently, the distribution $p_{B|A}$ may be considered as the output of a stochastic map $\mathcal{N}$ acting on p_A . Given observations B , the learner tries to estimate α using an *estimator* $\hat{\alpha} : \mathcal{B} \rightarrow A$. The learner's estimate is scored according to its distance from the true α . We will specify a scoring function later, but the essence is that the learner tries to minimize $d(\hat{\alpha}(B), \alpha)$.

The problem we study here differs from some other common models of learning. This learning task mostly resembles parameter estimation, and differs from *prediction* tasks common in machine learning where the learner's task is to construct a model that approximates an unknown target function with low error, with respect to the distribution of data. Probably Approximately Correct (PAC) learning is an example of a formal learning model for prediction tasks [170]. In PAC learning, an unknown function $f_\alpha : \mathcal{X} \rightarrow \mathcal{Y}$ is prepared and the goal of the learner is to output another function $h : \mathcal{X} \rightarrow \mathcal{Y}$ such that the error $\Pr_{x \sim p_X}(h(x) \neq f_\alpha(x))$ is small, with high probability. For instance, in a linear least squares problem the learner receives observations $B = \{(x_1, f_\alpha(x_1) + \epsilon_1), \dots, (x_n, f_\alpha(x_n) + \epsilon_n)\}$ with $x_i \sim p_X(x)$ and zero-mean error ϵ_i , and attempts to approximate $f_\alpha(x) := \langle \alpha, x \rangle$ with small error with respect to p_X . Instead, in our learning setting the learner's goal is to output $\hat{\alpha}$ close to α , a task which may generally be harder than PAC learning of f_α .

5.3.1 Classical learning lower bound via Fano method

We first review a well-known lower bound on the error of a learning task, based on Fano's inequality. This bound works by considering a discretized version of the learning task presented above, such that learning α within ϵ -error is harder than some hypothesis testing problem, which in turn has a success probability upper bounded by Fano's inequality. One kind of discretization that achieves this is an ϵ -packing:

Definition 10. ϵ -packing: $\mathcal{V} := \mathcal{V}(\epsilon, A, d) \subset A$ is an ϵ -packing of A in d if for every $\alpha_v, \alpha_{v'} \in \mathcal{V}$, $d(\alpha_v, \alpha_{v'}) \geq \epsilon$.

Since A is bounded, $|\mathcal{V}| < \infty$. We will assess the error of the learner's estimate with some non-decreasing loss function $\ell : \mathbb{R}^+ \rightarrow \mathbb{R}$, where smaller ℓ means a better estimate. Continuing with the previous example with $A \subset \mathbb{R}^p$, choosing $d(\alpha, \alpha') := \|\alpha - \alpha'\|_2$ and $\ell(x) = x^2$ corresponds minimizing squared error. Then, achieving optimal performance on the learning task turns out to be no harder than performing classical hypothesis testing for the nearest point to α in a 2ϵ -packing of A , as the following bound from the literature shows:

Proposition 11 (Classical lower bound, [149]). Fix $\epsilon > 0$ and let $\mathcal{V} := \mathcal{V}(2\epsilon, A, d)$ be an ϵ -packing of A in d . Define a random variable V uniformly distributed on $\{1, 2, \dots, |\mathcal{V}|\}$, and $\hat{V} : \mathcal{B} \rightarrow \{1, \dots, |\mathcal{V}|\}$ as an estimator for the element of $\mathcal{V}$ closest to α given observations B :

$$\hat{V}(r) = \{v : \arg \min_{\alpha_{v'} \in \mathcal{V}} d(\alpha_{v'}, r) = \alpha_v\}. \quad (5.1)$$

Then for a non-decreasing $\ell : \mathbb{R}^+ \rightarrow \mathbb{R}$ the following *minimax* lower bound holds:

$$\min_{\hat{\alpha}} \max_{\alpha \in A} \mathbb{E}_{p_{B|\alpha}} [\ell(d(\hat{\alpha}(B), \alpha))] \geq \ell(\epsilon) \left(\frac{H(V|B) - 1}{\log |\mathcal{V}|} \right). \quad (5.2)$$

Proof. For completeness, we reproduce the typical minimax bound that uses mutual information to lower bound the error of the best estimator on the worst-case distribution, adapted from Ref. [171, Chapter 7] and originally formulated in Ref. [149].

The random variable $\mathcal{B}$ is distributed according to $p_{B|\alpha}$ conditioned on $\alpha \in A$. We define an estimator $D : \mathcal{B} \rightarrow A$ for $\alpha \in A$, and write $\hat{\alpha} = D(\alpha)$. First, the total error is lower bounded by considering only the error arising from estimates that are at least ϵ -far from α , and assigning such estimates their minimal possible error of $\ell(\epsilon)$ (since ℓ is non-decreasing). Let $p_{\hat{\alpha}|\alpha}$ be the conditional distribution of $\hat{\alpha} = D(B)$ induced via the estimator $D : \mathcal{B} \rightarrow \mathcal{A}$ acting on $B \sim p_{B|\alpha}$, and henceforth denote $\hat{\alpha} := D(B)$. Defining $\mathbb{B}_\epsilon(\alpha) := \{\alpha' \in A : d(\alpha, \alpha') < \epsilon\}$ and its complement $\mathbb{B}_\epsilon^c(\alpha) = A \setminus \mathbb{B}_\epsilon(\alpha)$, then for any fixed α we have

$$\mathbb{E}_{p_{B|\alpha}} [\ell(d(\hat{\alpha}, \alpha))] = \int_A d\hat{\alpha} p_{\hat{\alpha}|\alpha}(\hat{\alpha}|\alpha) \ell(d(\alpha, \hat{\alpha})) \quad (5.3)$$

$$= \int_{\mathbb{B}_\epsilon(\alpha)} d\hat{\alpha} p_{\hat{\alpha}|\alpha}(\hat{\alpha}|\alpha) \ell(d(\alpha, \hat{\alpha})) + \int_{\mathbb{B}_\epsilon^c(\alpha)} d\hat{\alpha} p_{\hat{\alpha}|\alpha}(\hat{\alpha}|\alpha) \ell(d(\alpha, \hat{\alpha})) \quad (5.4)$$

$$\geq \ell(\epsilon) \int_{\mathbb{B}_\epsilon^c(\alpha)} d\hat{\alpha} p_{\hat{\alpha}|\alpha}(\hat{\alpha}|\alpha) \quad (5.5)$$

$$:= \ell(\epsilon) \Pr_{p_{\hat{\alpha}|\alpha}} (d(\alpha, \hat{\alpha}) \geq \epsilon) \quad (5.6)$$

The first line uses several basic properties of the Markov chain $A \rightarrow B \rightarrow D(B)$:

$$\mathbb{E}_{p_{B|\alpha}}[f(D(\beta), \alpha)] = \int_B d\beta p_{B|A}(\beta|\alpha) f(D(\beta), \alpha) \quad (5.7)$$

$$= \int_B d\beta p_{B|A}(\beta|\alpha) f(D(B), \alpha) \quad (5.8)$$

$$= \int_B d\beta p_{B|A}(\beta|\alpha) \underbrace{\int_A d\hat{\alpha} p_{\hat{A}|B}(\hat{\alpha}|\beta) f(D(\beta), \alpha)}_{=1} \quad (5.9)$$

$$= \underbrace{\int_B d\beta p_{B|A}(\beta|\alpha)}_{=1} \int_A d\hat{\alpha} p_{\hat{A}|A}(\hat{\alpha}|\alpha) f(\hat{\alpha}, \alpha) \quad (5.10)$$

$$= \int_A d\hat{\alpha} p_{\hat{\alpha}|\alpha}(\hat{\alpha}|\alpha) f(\hat{\alpha}, \alpha) \quad (5.11)$$

where we have applied the chain rule, conditional independence of $\hat{A}$ from A given B , and conditional independence of B from $\hat{A}$ given A . Consider $\hat{V} : A \rightarrow \mathcal{V}(2\epsilon, A, d)$ that returns the closest point in a 2ϵ -packing $\mathcal{V}(2\epsilon, A, d)$ to an estimate $\hat{\alpha} := D(B) \in A$. Writing $[\mathcal{V}] := \{1, \dots, |\mathcal{V}|\}$ and indexing every element $\alpha_v \in \mathcal{V}$ by $v \in [\mathcal{V}]$, we define

$$\hat{V}(\hat{\alpha}) = \arg \min_{v \in S_{\mathcal{V}}} d(\hat{\alpha}, \alpha_v). \quad (5.12)$$

Since $[\mathcal{V}]$ indexes a 2ϵ -packing, $d(\hat{\alpha}, \alpha_v) < \epsilon$ implies $\hat{V}(\hat{\alpha}) = v$ and conversely $\hat{V}(\hat{\alpha}) \neq v$ implies $d(\hat{\alpha}, \alpha_v) \geq \epsilon$. So, for any $v \in S_{\mathcal{V}}$ we have

$$\Pr_{p_{\hat{\alpha}|\alpha_v}}(\hat{V}(\hat{\alpha}) \neq v) \leq \Pr_{p_{\hat{\alpha}|\alpha_v}}(d(\hat{\alpha}, \alpha_v) \geq \epsilon) \quad (5.13)$$

Considering the most difficult element $\alpha \in A$ to learn leads to the following chain of inequalities, which holds for any prior probability distribution p_V on $[\mathcal{V}]$:

$$\max_{\alpha \in A} \Pr_{p_{\hat{\alpha}|\alpha}}(d(\hat{\alpha}, \alpha) \geq \epsilon) \geq \max_{v \in [\mathcal{V}]} \Pr_{p_{\hat{\alpha}|\alpha_v}}(d(\hat{\alpha}, \alpha_v) \geq \epsilon) \quad (5.14)$$

$$\geq \mathbb{E}_{p_V} \Pr_{p_{\hat{\alpha}|\alpha_v}}(d(\hat{\alpha}, \alpha_v) \geq \epsilon) \quad (5.15)$$

$$= \sum_{v \in [\mathcal{V}]} \Pr_{p_{\hat{\alpha}|\alpha_v}}(d(\hat{\alpha}, \alpha_v) \geq \epsilon) p_V(v) \quad (5.16)$$

$$\geq \sum_{v \in [\mathcal{V}]} \Pr_{p_{\hat{\alpha}|\alpha_v}}(\hat{V}(\hat{\alpha}) \neq v) p_V(v) \quad (5.17)$$

$$= \Pr_{p_V}(\hat{V}(\hat{\alpha}) \neq V) \quad (5.18)$$

Combining Eqs. 5.13 and 5.17 and minimizing over estimates $\hat{\alpha}$ gives

$$\min_{\hat{\alpha}} \max_{\alpha \in A} \mathbb{E}_{p_{\hat{\alpha}|\alpha}} [\ell(d(\hat{\alpha}(B), \alpha))] \geq \min_{\hat{\alpha}} \max_{\alpha \in A} \ell(\epsilon) \Pr_{p_{\hat{\alpha}|\alpha}} (d(\hat{\alpha}, \alpha) \geq \epsilon) \quad (5.19)$$

$$\geq \min_{\hat{\alpha}} \ell(\epsilon) \Pr_{p_V} (\hat{V}(\hat{\alpha}) \neq V) \quad (5.20)$$

$$:= \ell(\epsilon) \min_{\hat{V}: \mathcal{B} \rightarrow [|\mathcal{V}|]} \Pr_{p_V} (\hat{V}(B) \neq V) \quad (5.21)$$

where we have observed that the optimal discrete estimator $\hat{V}$ acting on observations in $\mathcal{B}$ performs equally well as the the optimal discretized estimator $\hat{\alpha}$ applied to those same observations. □

This bounds the accuracy of an optimal estimator outputting the hardest-to-learn element in A . Thus, the bound describes the best learning performance (with respect to learning algorithms) in a worst-case scenario. Some intuition behind this bound: we can tune the bound's sensitivity to error by choosing a larger or smaller ϵ . If ϵ is made very large, then $|\mathcal{V}| = 1$ is the largest packing of A , in which case the corresponding hypothesis testing task is trivial but the error of the estimate $\hat{\alpha}$ might be intolerably large. On the other hand, a small ϵ results in a large $|\mathcal{V}|$ and the error lower bound becomes vacuous. In general, this bound states that a mutual information of roughly $I(V:B) \approx H(V)$ bits between the observations B and the index of the ϵ -ball containing α is necessary for *any* estimator to succeed, but does not provide any statement as to whether such a condition is sufficient for success. Ref. [172] derived a similar bound to Proposition 11 for when observations B consist of quantum states, by reducing the corresponding learning problem to quantum hypothesis testing on an appropriately defined discretization.

5.3.2 Classical learning upper bound

In the previous section we saw how a learning problem may be discretized into a hypothesis testing problem for determining a discrete random variable V such that about $H(V)$ bits of mutual information $I(V:B)$ between the learner's observations and V are necessary to succeed. Here we will derive an amount of mutual information that is sufficient for an optimal learner to learn *something* about the desired variable. The guarantee works by discretizing the learning task differently than before, such that learning within ϵ -error is harder than a corresponding classical hypothesis testing problem, which in turn has a success probability lower bounded according to Eq. 4.6. In this case, the relevant discretization involves an ϵ -net:

Definition 12. ϵ -net: $\mathcal{W} := \mathcal{W}(\epsilon, A, d) \subset A$ is an ϵ -net of A in d if for every $\alpha \in A$, there exists $\alpha_w \in \mathcal{W}$ such that $d(\alpha, \alpha_w) \leq \epsilon$.

To make our discretization arguments using the ϵ -net, we will introduce the notion of an ϵ -covering partition, which divides the subset A into regions consisting of all points that are closest to any given $\alpha_w \in \mathcal{W}$:

Definition 13 (ϵ -covering partition). Let A be a bounded subset of a metric space and $\mathcal{W} := \mathcal{W}(\epsilon, A, d)$ be an ϵ -net of A in d . The set of closest points in $\mathcal{W}$ to r is

$$\hat{W}(r) = \{w : \arg \min_{\alpha_{w'} \in \mathcal{W}} d(\alpha_{w'}, r) = \alpha_w\} \quad (5.22)$$

Then we define an ϵ -covering partition of A with respect to $\mathcal{W}$ as

$$C(A, \mathcal{W}) = \{R_w : w = 1, \dots, |\mathcal{W}|\}, \quad (5.23)$$

$$\text{where } R_w = \{r \in A : \hat{W}(r) \ni w\}. \quad (5.24)$$

$C(A, \mathcal{W})$ forms a partition of A , since for all $R_w, R_{w'} \in C(A, \mathcal{W})$, $R_w \cap R_{w'} = \emptyset$ for $w \neq w'$ and $\cup_{w=1 \dots |\mathcal{W}|} R_w = A$ with ties broken arbitrarily, and the topology (e.g. openness) of any given subset R_w may also be assigned arbitrarily. An important property of $C(A, \mathcal{W})$ is that any estimator that outputs w for all $\alpha \in R_w$ is guaranteed to be ϵ -accurate. In this setting, the learner attempts to maximize a non-increasing *score* $s : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ that potentially depends nonlinearly on the distance d . For example, choosing $s(t) = c - t$ (with c being some large-enough constant like the diameter of A) means that maximizing $s(d(x, y))$ corresponds to minimizing the distance between elements x and y (larger s means a better estimator). We may now state the guarantee for the above learning problem:

Proposition 14 (Classical learning upper bound). Let $\mathcal{W} := \mathcal{W}(\epsilon, A, d)$ be an ϵ -net of A in d and define a random variable W uniformly distributed on $\mathcal{W}(\epsilon, A, d)$, and let $s : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a non-increasing strictly positive function used to score the accuracy of the estimator. Then we have,

$$\max_{\hat{\alpha}} \max_{\alpha \in A} \log \mathbb{E}_{p_{B|\alpha}} [s(d(\hat{\alpha}(B), \alpha))] \geq \log s(\epsilon) - H(W|B). \quad (5.25)$$

Proof. First, for any fixed $\alpha \in A$ the learner's expected score for sampling $\hat{\alpha} := \hat{\alpha}(B)$ with respect to the induced distribution $p_{\hat{\alpha}|\alpha}$ is no less than the minimum expected score

considering only estimates $\hat{\alpha}$ that are ϵ -close to α . Then, we have

$$\mathbb{E}_{p_{\hat{\alpha}|\alpha}}[s(d(\alpha, \hat{\alpha}))] = \mathbb{E}_{p_{\hat{\alpha}|\alpha}}[s(d(\alpha, \hat{\alpha}))(\mathbb{I}\{d(\alpha, \hat{\alpha}) \leq \epsilon\} + \mathbb{I}\{d(\alpha, \hat{\alpha}) > \epsilon\})] \quad (5.26)$$

$$\geq \mathbb{E}_{p_{\hat{\alpha}|\alpha}}[s(d(\alpha, \hat{\alpha}))\mathbb{I}\{d(\alpha, \hat{\alpha}) \leq \epsilon\}] \quad (5.27)$$

$$\geq s(\epsilon) \Pr_{p_{\hat{\alpha}|\alpha}}(d(\alpha, \hat{\alpha}) \leq \epsilon). \quad (5.28)$$

Line 5.27 requires $s(t) \geq 0$ while line 5.28 used the fact that s is non-increasing. Defining an ϵ -net $\mathcal{W}(\epsilon, A, d)$ with elements α_w indexed by $w \in [|\mathcal{W}|] := \{1, \dots, |\mathcal{W}(\epsilon, A, d)|\}$ we have

$$\max_{\alpha} \Pr_{p_{\hat{\alpha}|\alpha}}(d(\alpha, \hat{\alpha}) \leq \epsilon) \geq \max_{w \in [|\mathcal{W}|]} \Pr_{p_{\hat{\alpha}|\alpha_w}}(d(\alpha_w, \hat{\alpha}) \leq \epsilon) \quad (5.29)$$

$$\geq \mathbb{E}_{w \in [|\mathcal{W}|]} \Pr_{p_{\hat{\alpha}|\alpha_w}}(d(\alpha_w, \hat{\alpha}) \leq \epsilon) \quad (5.30)$$

The first inequality follows from $\mathcal{W}(\epsilon, A, d) \subset A$, and the second follows since the best-case estimator always outperforms the average-case estimator. For some distribution on $[|\mathcal{W}|]$, the RHS of Eq. 5.30 might result in a trivial bound since some distributions on $[|\mathcal{W}|]$ will not contain any w such that $d(\alpha_w, \hat{\alpha}) \leq \epsilon$. We may account for this possibility by evaluating the final expression for the uniform distribution on $\mathcal{W}$, i.e.

$$\max_{\alpha} \Pr(d(\alpha, \hat{\alpha}) \leq \epsilon) \geq \frac{1}{|\mathcal{W}|} \sum_{w \in [|\mathcal{W}|]} \Pr_{p_{\hat{\alpha}|\alpha_w}}(d(\alpha_w, \hat{\alpha}) \leq \epsilon). \quad (5.31)$$

Finally, we substitute the continuous estimator for $\hat{\alpha}$ for a discrete estimator $\hat{W} : A \rightarrow \mathcal{W}(\epsilon, A, d)$ to recover a multi-hypothesis testing scenario. The discrete estimator is given by any $w \in [|\mathcal{W}|]$ satisfying

$$\hat{W}(\hat{\alpha}) = \arg \min_{w \in [|\mathcal{W}|]} d(\hat{\alpha}, \alpha_w) \quad (5.32)$$

For some fixed α , suppose w is an index for which $d(\alpha, \alpha_w) \leq \epsilon$ (such an index must exist since we are working with an ϵ -net). Then, $\hat{W} = w$ implies $d(\hat{\alpha}, \alpha_w) \leq \epsilon$. This is opposite to the implication in Eq. 5.13, due to the possibility of overlap between the ϵ -balls composing a covering, so that several elements of $\mathcal{W}(\epsilon, A, d)$ may be ϵ -close to $\hat{\alpha}$. Intuitively, the optimal estimator $\hat{W}$ has a lower probability of success because, for some fixed α_w , $\hat{W}$ must output a unique solution from the ϵ -covering partition, whereas several elements of $\mathcal{W}$ may be ϵ -close to α_w . For any random variable W taking values in $[|\mathcal{W}|]$, the optimal estimator then obeys

$$\max_{\hat{\alpha}} \Pr(d(\hat{\alpha}, \alpha_w) \leq \epsilon) \geq \max_{\hat{W}: A \rightarrow [|\mathcal{W}|]} \Pr(\hat{W}(\hat{\alpha}) = w). \quad (5.33)$$

Finally, from the definition of conditional min-entropy we have

$$-\log \max_{\hat{W}} \Pr(\hat{W} = W) = H_{\min}(W|B) \leq H(W|B). \quad (5.34)$$

Combining Eqs. 5.30, 5.31, 5.33, and 5.34, and applying the appropriate maximizations and logarithm (imposing strict positivity on s if necessary), we recover the proposition. $\square$

In contrast with Proposition 11, this bound describes the optimal performance on a learning algorithm (with respect to estimators $\hat{\alpha}$) in a best-case scenario, i.e. the easiest-to-learn α . We might describe this as the onset of learning: Whenever this bound is nontrivial, then it must be true that *something* may be learned given the right observations B . The logarithm on the LHS of Proposition 14 (compared to Proposition 11) appears because Eq. 4.6 is loose, compared to Fano's inequality which is sharp [131]. This is easily rectified by substituting a tight bound (Ref. [130], Theorem 1) at the expense of simplicity.

This classical upper bound provides insight into the relationship between learning and classical correlations. In Proposition 14, a learner is successful if they are able to output $\hat{\alpha}$ within ϵ of the target variable α when provided observations $B|\alpha$. The maximum probability that this is satisfied (with respect to learning algorithms $\hat{\alpha}$) is,

$$\max_{\hat{\alpha}} \Pr_{p_{\hat{\alpha}|\alpha}}(d(\hat{\alpha}, \alpha) \leq \epsilon) := \max_{\hat{\alpha}} \mathbb{E}_{p_{\hat{\alpha}|\alpha}} [\mathbb{I}\{d(\hat{\alpha}, \alpha) \leq \epsilon\}] = \max_{\hat{\alpha}} \mathbb{E}_{p_{\hat{\alpha}|\alpha}} [\mathbb{I}\{\hat{\alpha} \in \mathbb{B}_{\epsilon}(\alpha)\}], \quad (5.35)$$

where $\mathbb{B}_{\epsilon}(\alpha) = \{\alpha' \in A : d(\alpha, \alpha') \leq \epsilon\}$ is the intersection of the ϵ -ball centered at α with A . The final expression can be interpreted as the expected value of a witness for classical correlation between two continuous variables $\alpha, \hat{\alpha}$, i.e. the classical learner is successful if their estimate $\hat{\alpha}$ results in a large expected value of the operator $\mathbb{I}\{\hat{\alpha} \in \mathbb{B}_{\epsilon}(\alpha)\}$.

5.4 Maximal singlet fraction

We now formalize the task of singlet fraction maximization. Consider a quantum system RA in Figure 5.1b initially containing a pure (possibly entangled) state $|\psi\rangle_{RA} \in \mathcal{H}_{RA}$, where Hilbert spaces $\mathcal{H}_R, \mathcal{H}_A$ have dimension $\dim(\mathcal{H}_R) = \dim(\mathcal{H}_A) := d$. Then, the system A evolves via channel $\mathcal{N} \in C(\mathcal{H}_A, \mathcal{H}_B)$, and the objective of the learner is to apply an *estimator channel* $\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$ (with $\dim(\mathcal{H}_{\hat{A}}) = \dim(\mathcal{H}_A)$) in order to maximize the overlap of the final state with a maximally entangled state. The *singlet fraction* is defined as this overlap, up to the dimensionality of the system:

$$q(R|B)_{\rho} := d \langle \phi | (\mathbb{I}_R \otimes \mathcal{D})(\rho_{RB}) | \phi \rangle_{R\hat{A}}, \quad (5.36)$$

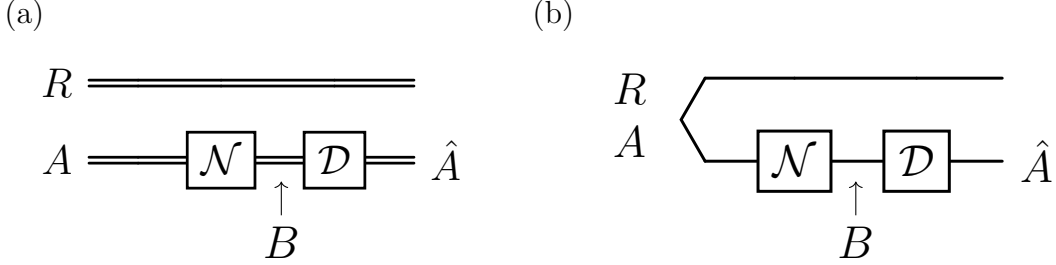


Figure 5.1: Schematic comparison of classical multiple hypothesis testing versus maximizing singlet fraction. **(a)** Letting double lines denote discrete classical random variables and boxes denote stochastic maps on the respective distribution, classical hypothesis testing involves a random variable A taking values $a \in \mathcal{A}$ sampled according to $p_A \in \mathcal{P}(\mathcal{A})$, which results in observations B according to $p_{B|A} = \mathcal{N}(p_A)$. The learner samples $\hat{A}(B)$ from $\mathcal{D}(p_{B|A})$ with the goal of maximizing $\Pr(\hat{A} = A)$ with respect to $\mathcal{D}$. The reference system plays no role in the classical case, as classical information may be freely copied from A to R . **(b)** In the quantum setting, single lines denote Hilbert spaces with dimension equal to the size of the alphabet of the corresponding random variable (e.g. $\mathcal{H}_A = \mathbb{C}^{|\mathcal{A}|}$). An initial state $|\psi\rangle_{RA} \in \mathcal{H}_{RA}$ is evolved by a channel $\mathcal{N} \in \mathcal{C}(\mathcal{H}_A, \mathcal{H}_B)$, and the goal of the learner is to maximize the singlet fraction $q(R|B) = |\mathcal{A}| \langle \phi | (\mathbb{I} \otimes \mathcal{D}) \rho_{RB} | \phi \rangle$ with respect to channels $\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$, where $|\phi\rangle \in \mathcal{H}_{R\hat{A}}$ is a maximally entangled state and $\rho_{RB} = (\mathbb{I} \otimes \mathcal{N})(|\psi\rangle\langle\psi|)$. In both cases we consider only the non-asymptotic, or *one-shot* setting: Noisy channel coding [115] does not apply because there is no source coding for A , and Schumacher coding [173] does not apply for analogous reasons.

where

$$\rho_{RB} = (\mathbb{I}_R \otimes \mathcal{N})(|\psi\rangle\langle\psi|_{RA}) \quad (5.37)$$

$$|\phi\rangle_{RA} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} |a_R\rangle \otimes |a_A\rangle \quad (5.38)$$

are the state on RB after the evolution $\mathcal{N}$ and a canonical maximally entangled state respectively. Unlike the classical quantity $p(A|\hat{A})$, $q(R|B)_\rho$ takes values in the interval $[0, d]$. The singlet fraction can be used to characterize the ability of $\mathcal{D}$ to recover quantum information from noise: If the input state were a singlet, $|\psi\rangle_{RA} = |\phi\rangle_{RA}$, then by the Choi-Jamiołkowski correspondence [174, 175] the maximal $q(R|B)_\rho = d$ is achieved if and only if $\mathcal{D} \circ \mathcal{N} = \mathbb{I}$, and if $\mathcal{D} \circ \mathcal{N}$ is close to $\mathbb{I}$ (in diamond norm) then $q(R|B)_\rho$ will be close to maximal [176]. However, moving forward we will not require the input state $|\psi\rangle_{RA}$ to be maximally entangled.

We can express the maximal value of $q(R|B)_\rho$ in terms of an information-theoretic quantity called the *conditional min-entropy*. Given a finite-dimensional bipartite quantum system $\mathcal{H}_A \otimes \mathcal{H}_B$, the conditional min-entropy is defined with respect to a state $\rho \in D(\mathcal{H}_{AB})$ as:

$$H_{\min}(A|B)_\rho = - \inf_{\sigma \in D(\mathcal{H}_B)} D_{\max}(\rho || \mathbb{I} \otimes \sigma) \quad (5.39)$$

where $D_{\max}(\rho || Q) = \inf\{\lambda \in \mathbb{R} : \rho \leq 2^\lambda Q\}$ for a positive operator $Q \geq 0$ (e.g. [177]). For any state, it holds that $H_{\min}(A|B)_\rho \leq H(A|B)_\rho$. $H_{\min}$ has an operational interpretation for both classical distributions and quantum states given by [162]: By choosing $\mathcal{H}_A = \mathbb{C}^{|A|}$ and $\mathcal{H}_B = \mathbb{C}^{|B|}$ and defining a diagonal state p with elements $p_{AB}(a, b)$ on its diagonal, we have

$$2^{-H_{\min}(A|B)_p} = \max_{\hat{A}: B \rightarrow A} p(A|\hat{A}(B)) \quad (5.40)$$

$$2^{-H_{\min}(A|B)_\rho} = \max_{\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} q(R|B)_\rho. \quad (5.41)$$

Noting that $q(R|B) \leq 1$ for separable states (and therefore classical distributions represented by p) [35, Example 6.16], we recover the statement of Eq. 4.6 about optimal success probability for classical hypothesis testing from Eq. 5.40

Just as the error of classical hypothesis testing is both upper and lower bounded in terms of the conditional Shannon entropy, the maximal singlet fraction may be bounded in terms of quantum entropies. The lower bound on $q(R|B)_\rho$ for the optimal estimator becomes¹ [162]:

$$\max_{\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} q(R|B)_\rho = 2^{-H_{\min}(R|B)_\rho} \geq 2^{-H(R|B)_\rho}. \quad (5.42)$$

Furthermore, we may upper bound $q(R|B)_\rho$ using an inequality that is superficially similar to Fano's inequality. The following represents a family of such inequalities:

Proposition 15 (Quantum Fano inequality [178]). For any pure state $|\psi\rangle \in \mathcal{H}_{RA}$ and any $\rho \in D(\mathcal{H}_{RB})$ with $\dim(B) = \dim(A) = d$, define $p = \langle \psi | \rho | \psi \rangle$. Then,

$$H(RB)_\rho \leq h_2(p) + (1 - p) \log(d^2 - 1) \quad (5.43)$$

¹While the guarantee for classical hypothesis testing is a special case of the conditional min-entropy in the same way that linear programming is a special case of semi-definite programming, it is interesting that the former was derived in Ref. [130] by different means entirely.

This inequality was originally derived in Ref. [178] to relate the *entropy exchange* (or entropy of environment plus reference system) to the *entanglement fidelity* (overlap of the noisy state $\rho = (\mathbb{I}_R \otimes \mathcal{N})(|\psi\rangle\langle\psi|)$ with initial state $|\psi\rangle$). While that choice leads to a relatively tight inequality that is relevant for, e.g. quantum error correction, it does not provide a bound on singlet fraction. Consider instead the following instantiation of Proposition 15:

$$H(RB)_\rho \leq h_2\left(\frac{q(R|B)_\rho}{d}\right) + \left(1 - \frac{q(R|B)_\rho}{d}\right) \log(d^2 - 1) \quad (5.44)$$

where $\rho = \rho_{RB}$ is defined in Eq. 5.37. In this setting, the above inequality provides an analogous bound for maximizing singlet fraction as Fano's inequality provides for classical hypothesis testing. $H(RB)_\rho$ appears here rather than $H(R|B)_\rho$ since this bound is derived from a spectral decomposition of the bipartite system RB , which does not relate to the conditional entropy in any general sense.

Between the tasks of classical hypothesis testing and maximizing singlet fraction is *quantum hypothesis testing*, which describes the scenario in which A represents a discrete random variable and B represents a quantum system (e.g. [179, 180]). In this scenario, conditional on $A = a$ the learner receives a quantum state $\sigma_B^a \in \mathcal{D}(\mathcal{H}_B)$ and the learner's task is to determine a by performing measurements on the quantum system. This scenario can be characterized via a joint classical-quantum system AB prepared in a state $\rho = \sum_a p_A(a) |a\rangle\langle a| \otimes \sigma_B^a$, such that $H_{\min}(A|B)_\rho$ determines the learner's optimal probability of identifying A using measurements on system B [162].

5.4.1 Reduction from singlet fraction to classical hypothesis testing

We briefly review how maximizing singlet fraction reduces to the objective of classical hypothesis testing. Our main tool will be a *completely dephasing map* $\Delta \in \mathcal{C}(\mathcal{H})$ defined with respect to a fixed basis $\{|i\rangle\} \subset \mathcal{H}$ by its action $\Delta(\rho) = \sum_i |i\rangle\langle i| \langle i|\rho|i\rangle$. Applying this map to both subsystems of the state ρ_{RB} reduces the singlet fraction to the probability of success for a classical estimator:

$$\langle \phi_{R\hat{A}}, (\mathbb{I} \otimes \mathcal{D}) \circ (\Delta \otimes \Delta)(\rho_{RB}) \rangle = \langle (\Delta \otimes \mathbb{I})(\phi_{R\hat{A}}), (\mathbb{I} \otimes \mathcal{D} \circ \Delta)(\rho_{RB}) \rangle \quad (5.45)$$

$$= \frac{1}{d} \sum_{r,b} p_{RB}(r, b) \langle r | \mathcal{D}(|b\rangle\langle b|) | r \rangle \quad (5.46)$$

where $p_{RB}(r, b) := \langle rb | \rho | rb \rangle$ and we have used the self-adjointness of Δ , $\langle X, \Delta(Y) \rangle = \langle \Delta(X), Y \rangle$. Fixing the values r, b and maximizing the final system over estimator channels gives,

$$\max_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \langle r | \mathcal{D}(|b\rangle\langle b|) | r \rangle = \max_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \langle r | \Delta \circ \mathcal{D} \circ \Delta(|b\rangle\langle b|) | r \rangle := \max_{\hat{A}: \mathcal{B} \rightarrow \mathcal{A}} \Pr(\hat{A}(B) = r | B = b). \quad (5.47)$$

Writing $\Pr(\hat{A}(B) = r | B = b) := p_{\hat{A}|B}(r|b)$ describing the conditional probability of a measurement outcome r , we recover

$$\sum_{r,b} p_{RB}(r, b) p_{\hat{A}|B}(r|b) = \sum_{r,b} p_{RB}(r, b) p_{\hat{A}|RB}(r|r, b) = \sum_r p_{R\hat{A}}(r, r) = \Pr_{p_R}(\hat{A} = R), \quad (5.48)$$

where we have used conditional independence of $\hat{A}$ and R given B . Combining our results,

$$\max_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} d \langle \phi_{R\hat{A}}, (\mathbb{I} \otimes \mathcal{D}) \circ (\Delta \otimes \Delta)(\rho_{RB}) \rangle = \max_{\hat{A}: \mathcal{B} \rightarrow \mathcal{A}} \Pr_{p_R}(\hat{A} = R). \quad (5.49)$$

We see that a singlet fraction maximization problem is naturally associated with, or *reduces to*, a classical hypothesis testing problem via application of a completely dephasing channel. Moving forward, when we introduce our quantum information processing task that reduces to classical learning it will be this kind of reduction that we seek.

5.5 A quantum generalization of learning

A primary challenge for quantum machine learning is to identify tasks for which quantum information processing might provide an advantage over classical methods. Accordingly, a large body of work has been devoted to proposing and studying settings where quantum computers could offer an advantage in learning from classical data [13, 18, 34, 181–183] and for characterizing quantum states and processes [184–189]. Some of these tasks involve data that are provided in superposition and may be manipulated coherently by the learner, such as the Harrow–Hassidim–Lloyd algorithm for linear systems of equations [190]. In particular, the model of quantum PAC learning [21] generalizes PAC learning by providing a learner access to observations (training data) in coherent superposition generated via an unknown concept (target function), but otherwise the learner’s goal is still to approximate the unknown concept to high accuracy with respect to the distribution of observations. Quantum learners in the quantum PAC setting have been shown to provide advantages in sample complexity for specific learning tasks [22, 23], though this advantage can disappear under more general conditions [24].

In order to reason about the relationship between entanglement and learning, we first consider an example of a formal learning scenario similar to the classical learning task introduced above. The framework of *exact learning* involves an unknown concept described by a discrete variable, and the learner must identify this variable *exactly* with high probability given a set of observations [191]. The generalization to quantum exact learning [192, 193], in which the learner has access to observations described by a quantum state, will help us reason about how we might generalize our classical learning task involving continuous variables to a quantum task involving infinite dimensional states.

5.5.1 An example: Beyond quantum exact learning

Fundamentally, despite providing a learner with quantum data, the quantum PAC and quantum exact learning settings described above still involve extracting classical information about an unknown target function. This limitation suggests a more general setting in which the target functions themselves are treated as *quantum data* that exist in coherent superposition. To demonstrate this concept in a limited (i.e. finite dimensional) setting, we will consider a generalization of classical (and quantum) exact learning that mirrors the generalization of classical learning that we will soon introduce.

We review the setting of classical exact learning in more detail. In this setting, without loss of generality, a concept (boolean function with bitstring inputs) c_A is sampled from a concept class $\mathcal{C} \subseteq \{c_a : \{0, 1\}^n \rightarrow \{0, 1\}\}$ with some probability $p_A(a)$. Once $A = a$ is fixed, a learner is given oracle access for evaluating $c_a(x)$ for input bitstrings $x \in \{0, 1\}^n$, and tries to determine c_a with high probability with respect to a distribution of concepts c . Specifically, a successful exact learner outputs a hypothesis $h : \{0, 1\}^n \rightarrow \{0, 1\}$ such that $h(x) = c(x)$ for all $x \in \{0, 1\}^n$ with probability at least $2/3$. Considering Figure 5.1a, this scenario is modelled by introducing a random variable A taking values $a \in \mathcal{A} := \{1, \dots, |\mathcal{C}|\}$ with probability $p_A(a)$, and then describing the learner's m queries via a random variable B sampled from $p_{B|A} = \mathcal{N}(p_A)$ conditioned on A :

$$B = \{(x_1, c_A(x_1)), \dots, (x_m, c_A(x_m))\}, \quad (5.50)$$

where each x_i is sampled from some distribution p_X , which can be completely described by $\mathcal{N}$. For consistency of notation, we describe the joint distribution of (A, B) as a bipartite classical probability distribution (diagonal density operator) of the form

$$p_{AB} = \sum_{a \in \mathcal{A}} p_A(a) |a\rangle \langle a|_A \otimes p_B^a \quad (5.51)$$

where $p_B^a = \sum_{b \in \mathcal{B}} p_{B|A}(b|a)|b\rangle\langle b|_B$ is a conditional distribution. In this case, the learner's strategy to identify c_a given $\{(x_i, c_a(x_i))\}_{i=1}^m$ is equivalent to finding an estimator $\hat{A} = \hat{A}(B)$ that maximizes $\Pr_{p_A}(\hat{A} = A)$. Then, for any fixed choice $A = a$ of target function, sampling the random variable B corresponds to preparing a size- m classical dataset of $(x, c_a(x))$ pairs, and we may view exact learning as a kind of classical hypothesis testing for the random variable A .

Quantum exact learning generalizes exact learning by providing samples in superposition via an oracle $|x, 0\rangle \rightarrow |x, c_A(x)\rangle$, while the concept c_A is still described by a (classical) random variable. Retaining all of the notation from classical exact learning, this scenario is characterized by a channel $\mathcal{N} \in C(\mathcal{H}_A, \mathcal{H}_B)$ composed of oracles and classical sampling from p_X that prepares a classical-quantum state of the form

$$\sum_{a \in \mathcal{A}} p_A(a) |a\rangle\langle a|_A \otimes |\psi_x^a\rangle\langle \psi_x^a|_B \quad (5.52)$$

where $|\psi_x^a\rangle = \sum_{i=1}^m \sqrt{p_X(x_i)} |x_i, c_a(x_i)\rangle \in \mathcal{H}_B$ is a coherent superposition of examples x_i along with the concept c_a evaluated for that input. The goal of a quantum exact learner is to identify c_a , which they can accomplish by learning some channel $\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$ acting on $\mathcal{H}_B$ that outputs an estimate $\hat{A}$ maximizing $\Pr_{p_A}(\hat{A} = A)$. From the arguments of Sec. 5.4.1, for each exact learning problem defined by priors p_A and conditionals p_B^a in Eq. 5.51, there is a quantum exact learning problem that is at least as difficult, which may be constructed by choosing states $|\psi_x^a\rangle\langle \psi_x^a|_B$ such that $\Delta(|\psi_x^a\rangle\langle \psi_x^a|_B) = p_B^a$.

The generalization from classical to quantum exact learning immediately suggests a further generalization of quantum exact learning in which the concepts c_A are themselves in superposition. This task should be at least as hard as quantum exact learning, and should reduce to quantum exact learning via application of a completely dephasing channel Δ . Referring to Fig. 5.1b, suppose that RA initially contains a state $|\psi_{RA}\rangle := \sum_{a \in \mathcal{A}} \sqrt{p_A(a)} |a\rangle_R \otimes |a\rangle_A$. Constructing $\mathcal{N}$ from oracles and classical sampling as before, we prepare the state

$$\sum_{a \in \mathcal{A}} \sqrt{p_A(a)} |a\rangle_R \otimes |\psi_x^a\rangle_B. \quad (5.53)$$

In this case, what should the goal be for a learner who has the ability to apply local operations to the system B ? Based on the discussion of Sec. 4.3, the most natural choice is for the learner to learn a channel $\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$ that maximizes the overlap of the final state $(\mathbb{I} \otimes \mathcal{D})(\sigma_{RB})$ with a maximally entangled state in $R\hat{A}$. Based on the arguments of Sec. 5.4.1, this choice to maximize the singlet fraction $q(R|B)_\rho$ results in a task that

is at least as hard as quantum exact learning, and recovers the setting of quantum exact learning by application of the completely dephasing map on register A .

5.5.2 The entanglement fraction task

We are now ready to characterize a task which generalizes the setting of classical learning. Recall that in Sec. 4.3 we saw how classical hypothesis testing (involving discrete random variables) can be understood as a special case the more general task of achieving the maximal singlet fraction (involving finite dimensional Hilbert spaces and reference systems). In Sec. 5.3, we saw that hypothesis testing provides bounds (via discretization arguments) for the error of classical learning. In Sec. 5.5.1 we observed how exact learning (involving classical or quantum information) can be recovered as a special case of a task involving the maximal singlet fraction. We will now combine all of these ideas to introduce a quantum information processing task that generalizes the notion of classical learning, and furthermore we will show that the error of an optimal learner for this task may be bounded using tools from quantum information theory for finite-dimensional systems.

We briefly review quantum information theory in infinite-dimensional Hilbert spaces, overloading notation from the previous sections in the process. $\mathcal{H}$ denotes a separable Hilbert space of states, with inner product denoted $\langle \Psi | \Phi \rangle$ for $|\Psi\rangle, |\Phi\rangle \in \mathcal{H}$. $\mathcal{B}(\mathcal{H})$ denotes bounded linear operators on $\mathcal{H}$, and $\mathcal{T}(\mathcal{H}) \subseteq \mathcal{B}(\mathcal{H})$ denotes the trace class, self-adjoint operators on $\mathcal{H}$. The subset $\mathcal{D}(\mathcal{H}) \subset \mathcal{T}(\mathcal{H})$ of unit trace operators denotes the *density operators*. We will sometimes denote a pure state in density operator form as $|\Psi\rangle\langle\Psi| := \Psi \in \mathcal{D}(\mathcal{H})$ when there is no risk of confusion, and we will tend to use uppercase (e.g. Φ, Σ) for states of an infinite dimensional Hilbert space and lowercase (e.g. ϕ, ρ, σ) for states of a finite dimensional space. We use bra-ket notation to denote both states in $\mathcal{H}$ as well as (unbounded) linear functionals, e.g. sets of operators $\{|\alpha\rangle\}_{\alpha \in A}$ satisfying $\langle \alpha | \alpha' \rangle = \delta(\alpha, \alpha')$. By contrast, a basis $\{|i\rangle\}_i \in \mathcal{H}$ for separable $\mathcal{H}$ will be indexed with natural numbers. Frequently we will be working with states $|\Psi\rangle \in \mathcal{H}$ that are completely specified by the sets of coefficients $\Psi(x) := \langle x | \Psi \rangle$ for elements x of a (Borel) subset $\mathcal{X} \subset \mathbb{R}^d$, in which case we write $\Psi \in L^2(\mathcal{X})$ to denote square-integrable functions with respect to the measure $\mathbb{I}\{x \in \mathcal{X}\}$. A state $\rho \in \mathcal{D}(\mathcal{H})$ with $\mathcal{H} = L^2(\mathcal{X})$ in a continuous variable quantum system is naturally associated with a (positive) integral kernel operator $\rho : L^2(\mathcal{X}) \rightarrow L^2(\mathcal{X})$ of the form $(\rho\Psi)(x) = \int_{x' \in \mathcal{X}} dx' \rho(x, x') \Psi(x')$ with $\langle x | \rho | x' \rangle = \rho(x, x')$ defined for all $x, x' \in \mathcal{X}$ [194].

We will also be working with channels acting on separable Hilbert spaces. For now, we simply define the set of channels $\mathcal{C}(\mathcal{H}, \mathcal{H}')$ to be bounded linear maps $\mathcal{N} : \mathcal{B}(\mathcal{H}) \rightarrow \mathcal{B}(\mathcal{H}')$

that act on a state $\Sigma \in \mathcal{D}(\mathcal{H})$ according to

$$\mathcal{N}(\Sigma) = \text{Tr}_{\mathcal{E}}(U(\Sigma \otimes \Psi)U^\dagger), \quad (5.54)$$

where $\mathcal{H}', \mathcal{E}$ are separable Hilbert spaces, $U \in \mathcal{U}(\mathcal{H} \otimes \mathcal{E}, \mathcal{H}' \otimes \mathcal{E})$ is an isometry, and $\Psi \in \mathcal{D}(\mathcal{E})$ is a pure state. Further treatment of channels is delayed to Sec. 5.6.

Recall that the goal of a classical learner was to receive observations B (which depended on α) and then output an estimate $\hat{\alpha}$ that is classically correlated with α . Analogously, in the entanglement fraction task a learner applies local operations to a system B to maximize a type of quantum correlation between the resulting system $\hat{A}$ and some reference system R . Consider separable Hilbert spaces $\mathcal{H}_R, \mathcal{H}_A, \mathcal{H}_{\hat{A}}$, which may be taken to be $L^2(A)$ without loss of generality, and some additional separable Hilbert space $\mathcal{H}_B$. Fix a channel $\mathcal{N} \in \mathcal{C}(\mathcal{H}_A, \mathcal{H}_B)$, and consider all channels $\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$ acting on B . The quantum systems (A , B , and $\hat{A}$) are each analogous to (continuous) random variables (α , B , and $\hat{\alpha}$) in the classical learning setting: a quantum system A evolves into a quantum system B of “observations” via the map $\mathcal{N}$, and then the learner prepares the state of system $\hat{A}$ by applying $\mathcal{D}$ to system B . Unlike before, the initial state of the system is a potentially entangled state over the joint system RA . Given the system RA prepared in some initial state $|\Psi\rangle_{RA} \in \mathcal{H}_{RA}$,

$$|\Psi\rangle_{RA} = \int_A dr \int_A d\alpha \Psi_{RA}(r, \alpha) |r\rangle_R \otimes |\alpha\rangle_A \quad (5.55)$$

then system B is prepared by applying $\mathcal{N}$ to produce the state $\Sigma_{RB} = (\mathbb{I} \otimes \mathcal{N})(|\Psi\rangle\langle\Psi|_{RA})$. The goal of the learner is to find the $\mathcal{D}$ that maximizes the overlap of $(\mathbb{I} \otimes \mathcal{D})(\Sigma_{RB}) \in \mathcal{D}(\mathcal{H}_{R\hat{A}})$ with a particular (unnormalized) entangled state $|\Phi_\epsilon\rangle_{R\hat{A}} \in \mathcal{H}_{R\hat{A}}$, defined as

$$|\Phi_\epsilon\rangle_{R\hat{A}} = \int_A dr \int_{\mathbb{B}_\epsilon(r)} d\hat{\alpha} |r\rangle_R \otimes |\hat{\alpha}\rangle_{\hat{A}} \quad (5.56)$$

where $\mathbb{B}_\epsilon(r) := \{r' \in A : d(r, r') \leq \epsilon\}$ denotes an ϵ -ball centered at r (restricted to A). Thus, the learner’s goal is to maximize the entanglement fraction of the final state on $R\hat{A}$, which we might denote

$$Q_\epsilon(R|B)_\Sigma := \langle\Phi_\epsilon|(\mathbb{I} \otimes \mathcal{D})(\Sigma_{RB})|\Phi_\epsilon\rangle. \quad (5.57)$$

Before proceeding, we justify the unusual choice of entangled state for this scenario. A significant body of work has been devoted to characterizing entanglement in infinite dimensional systems [195–198]. In particular, many information-theoretic quantities in

finite-dimensional systems have been at least partially extended to the infinite dimensional case [199–202], and such analyses could conceivably be extended to define a notion of maximal singlet fraction in infinite dimensions. As there is no well-defined notion of a maximally entangled state in a bipartite infinite-dimensional Hilbert space (barring the use of potentially unbounded operators [203]), there is no obvious choice of entangled state with which to compute the fidelity of $(\mathbb{I} \otimes \mathcal{D})(\Sigma_{RB})$. However, we have chosen fidelity with the entangled state Eq. 5.56 with the goal that this task to reduce to classical learning as a special case, via an argument involving dephasing that is analogous to the relationship between maximal singlet fraction and classical hypothesis testing provided in Sec. 5.5.3. In the next section we will demonstrate that the choice of entangled state $|\Phi_\epsilon\rangle$ satisfies this goal, and in this sense we will argue that above entanglement fraction task is a generalization of learning.

5.5.3 The relationship to classical learning

Just as optimal performance in classical hypothesis testing can be considered a special case of maximal singlet fraction in the presence of completely dephasing noise, we will now demonstrate how classical learning arises as a special case of the entanglement fraction maximization defined above. The technique of Sec. 5.4.1 of applying a completely dephasing channel does not generalize to our purposes. The issue is that, while one could define a completely dephasing channel with respect to a countable basis $\{|i\rangle\} \subset \mathcal{H}$ (e.g. Ref. [204]), this approach will not recover statements about probability distributions over parameters $\alpha \in A$ that resemble classical learning of continuous parameters in any clear way. The basic problem is that any such basis is countable in $\mathcal{H}$, but our learning task involves statements about probability densities $\mathcal{P}(A)$. There is no such basis for $\mathcal{H}$ with components labelled by $\alpha \in A$ that is not also overcomplete [205]. Here, we will use an approach involving unbounded operators acting on $\mathcal{H}$, though we will need to also employ several basic concepts from measure theory.

For $\mathcal{H} = L^2(A)$, denote the *expectation value* $\langle X \rangle_\rho$ of a self-adjoint linear operator $X \in \mathcal{L}(\mathcal{H})$ with respect to a state $\rho \in D(\mathcal{H})$. Computed with respect to the set of linear functionals $\{|\alpha\rangle\}_{\alpha \in A}$ that form a resolution of the identity, we have that $\langle X \rangle_\rho = \int_A \int_A d\alpha d\alpha' \langle \alpha | X | \alpha' \rangle \rho(\alpha, \alpha') \in \mathbb{R}$. Then, we define the (unbounded) map $\Delta_A : D(\mathcal{H}) \rightarrow \mathcal{L}(\mathcal{H})$, the image of which has an expectation value that satisfies

$$\langle \Delta_A(\rho) \rangle_\sigma = \int_A d\alpha \rho(\alpha, \alpha) \sigma(\alpha, \alpha). \quad (5.58)$$

for $\rho, \sigma \in D(\mathcal{H})$. If we associate the input state ρ with a probability density $p \in L^1(A)$

according to $p(\alpha) := \rho(\alpha, \alpha) = \langle \alpha | \rho | \alpha \rangle$ and associate σ with $q \in L^1(A)$ similarly, then Eq. 5.58 corresponds to the probability density for measurements of ρ and σ to agree. While the coefficients $\langle \alpha | \rho | \alpha \rangle$ corresponding to a set of operators $\{|\alpha\rangle\langle\alpha|\}_{\alpha \in A}$ do not correspond to outcomes of a valid measurement ($\{|\alpha\rangle\langle\alpha|\}_{\alpha \in A}$ is not a POVM), a valid POVM can be constructed by integrating the operators $|\alpha\rangle\langle\alpha|$ with respect to a partition of Borel subsets of A . In this way, the quantities we will derive may be interpreted in terms of a measurement with respect to $\alpha \in A$ while not themselves corresponding to any such measurement. Defining a probability density $p_R(r) = \int_A d\alpha |\Psi(r, \alpha)|^2$ and writing the state

$$|\Psi^r\rangle_A = \frac{1}{\sqrt{p_R(r)}} \int_A d\alpha \Psi(r, \alpha) |\alpha\rangle \in L^2(A), \quad (5.59)$$

we find

$$\langle \Delta_A(\Phi_\epsilon) \rangle_{\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}(\Psi_{RA})} = \int_A dr p_R(r) \int_{\mathbb{B}_\epsilon(r)} d\hat{\alpha} \langle \hat{\alpha} | \mathcal{D} \circ \mathcal{N}(\Psi^r) | \hat{\alpha} \rangle \quad (5.60)$$

$$= \int_A dr p_R(r) \Pr_{p_{\hat{A}|R}}(d(\hat{\alpha}, r) \leq \epsilon), \quad (5.61)$$

where we have identified a probability density $p_{\hat{A}|R}(\hat{\alpha}|r) = \langle \hat{\alpha} | \mathcal{D} \circ \mathcal{N}(\Psi^r) | \hat{\alpha} \rangle$ for an outcome $\hat{\alpha}$ with respect to the collection of operators $\{|\hat{\alpha}\rangle\langle\hat{\alpha}|\}_{\alpha \in A}$ applied to a state $\Sigma_{\hat{A}}^r := \mathcal{D} \circ \mathcal{N}(\Psi^r) \in \mathcal{D}(\mathcal{H}_{\hat{A}})$. Substituting r for α , we recover the statement of Eq. 5.35 relating classical learning to the preparation of a variable $\hat{\alpha}$ (given observations in system B) that is highly correlated with the target parameter α via optimization over channels $\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$. Indeed, the operator $\Delta_A(\Phi_\epsilon)$ is essentially the indicator function $\mathbb{I}\{\hat{\alpha} \in \mathbb{B}_\epsilon(r)\}$ of Eq. 5.35 that is nonzero whenever the learner's estimate $\hat{\alpha}$ is ϵ -close to a classical reference variable r (which may be chosen arbitrarily close to α in the classical case).

The goal of the entanglement fraction task is for the learner to maximize $Q_\epsilon(R|B)_\Sigma$ with respect to channels $\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$. We may understand the best-case performance that a learner might achieve at this task by considering the maximum of $Q_\epsilon(R|B)_\Sigma$ with respect to states $|\Psi_{RA}\rangle \in \mathcal{H}_{RA}$. We will therefore compare maximization with respect to $|\Psi\rangle_{RA} \in \mathcal{H}_{RA}$ and estimator channels to the maximum over *parameters* $\alpha \in A$ and stochastic maps in the classical case (e.g. Proposition 14). Since Eq. 5.61 no longer explicitly depends on the

system A , optimization over initial states and channels $\mathcal{D}$ recovers

$$\sup_{|\Psi\rangle_{RA} \in \mathcal{H}_{RA}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \log \int_A dr p_R(r) \Pr_{p_{\hat{A}|R}}(d(\hat{\alpha}, r) \leq \epsilon) \quad (5.62)$$

$$= \sup_{p_R \in \mathcal{P}(A)} \sup_{\hat{\alpha}: B \rightarrow \hat{A}} \log \int_A dr p_R(r) \Pr_{p_{B|R}}(d(\hat{\alpha}(B), r) \leq \epsilon) \quad (5.63)$$

$$:= \text{ess sup}_{\alpha \in A} \sup_{\hat{\alpha}: B \rightarrow \hat{A}} \log \Pr_{p_{B|\alpha}}(d(\hat{\alpha}(B), \alpha) \leq \epsilon) \quad (5.64)$$

where the ess sup indicates that the supremum is to be taken with respect to subsets $\{\alpha\} \subseteq A$ having nonzero measure with respect to any probability measure defined on A . This constraint is arguably more reasonable than a taking a supremum over all $\alpha \in A$ by avoiding pathological scenarios, for instance, distributions of the unknown target parameter α that are everywhere discontinuous. Note that the system B implicitly represents a quantum system, with $\hat{\alpha}$ being some composition of quantum channel followed by measurement - we discuss this limitation further in Sec. 5.7.

We have shown that maximizing the entanglement fraction (the overlap with the particular entangled state $|\Phi_\epsilon\rangle$) reduces to maximizing classical correlations (the expected value of $\Delta_A(\Phi_\epsilon)$) that correspond to classical learning. Specifically, the choice of ϵ in $|\Phi_\epsilon\rangle$ corresponds to the error probability $\Pr(d(\alpha, \hat{\alpha}) \geq \epsilon)$ in some continuous parameter learning task, and $|\Phi_\epsilon\rangle\langle\Phi_\epsilon|$ is essentially a witness for quantum correlations that reduces the operator $\mathbb{I}\{\hat{\alpha} \in \mathbb{B}_\epsilon(\alpha)\}$ in Eq. 5.35 used to quantify classical correlations in the classical setting.

5.6 Bounds for the entanglement fraction task

We are interested in upper and lower bounds on the error of the entanglement fraction task, but we must first impose an additional constraint on the problem. The state $|\Phi_\epsilon\rangle$ is both theoretically challenging to work with due to its possibly infinite Schmidt rank, and practically challenging since preparation of such a state in a lab may be infeasible. Instead, we restrict our analysis to a more feasible scenario in which the entanglement in $|\Phi_\epsilon\rangle$ has been bounded by applying some bipartite projector $\Pi \in \mathcal{L}(\mathcal{H}_{R\hat{A}})$. This choice also allows us to interpret the entanglement fraction in terms of teleportation directly.

It is helpful to refer to the analogous classical learning guarantee: Recall that Proposition 14 was proven by covering the space A with balls having radius no greater than ϵ , and then showing that predicting the target parameter α to ϵ accuracy is easier than succeeding

at a multi-hypothesis testing task for determining which ball has the closest center to the target parameter α . The latter task may be understood as determining membership of α to an element of an ϵ -covering partition of A . Analogously, the technique here involves partitioning of a state space into a finite number of regions, and then working with classes of states defined piece-wise with respect to these regions. In this case, optimizing singlet fraction with respect to this class of states takes on a relationship to the entanglement fraction task that is similar to the relationship between hypothesis testing and classical learning.

We return to the discussion of quantum channels, following the treatment of Ref. [206, Chapter 6]. For separable Hilbert spaces $\mathcal{X}, \mathcal{Y}$, the partial trace $\text{Tr}_{\mathcal{Y}} : \mathcal{X} \otimes \mathcal{Y} \rightarrow \mathcal{X}$ is the unique operator satisfying $\text{Tr}(\text{Tr}_{\mathcal{Y}}(A)X) = \text{Tr}(A(X \otimes \mathbb{I}))$ for all $X \in \mathcal{B}(\mathcal{X})$ and $A \in \mathcal{T}(\mathcal{X} \otimes \mathcal{Y})$. For a linear map $\mathcal{N} : \mathcal{B}(\mathcal{H}) \rightarrow \mathcal{B}(\mathcal{H}')$, then $\mathcal{N}$ is a channel, denoted $\mathcal{N} \in \mathcal{C}(\mathcal{H}, \mathcal{H}')$, if and only if for all $X \in \mathcal{T}(\mathcal{H})$, we have $\mathcal{N}(\rho) = \text{Tr}_{\mathcal{E}}(U(X \otimes \Sigma)U^\dagger)$ for some separable Hilbert space $\mathcal{E}$, some (pure) state $\Sigma \in \mathcal{D}(\mathcal{E})$, and unitary operator $U \in \mathcal{U}(\mathcal{H} \otimes \mathcal{E}, \mathcal{H}' \otimes \mathcal{E})$. An alternative characterization of quantum channels may be given in terms of completely positive unital linear maps. An operator $\mathcal{M} : \mathcal{B}(\mathcal{H}') \rightarrow \mathcal{B}(\mathcal{H})$ is positive if it sends positive operators to positive operators, and is completely positive if the extension $\mathcal{M}_n$ of $\mathcal{M}$ to $\mathcal{B}(\mathcal{H} \otimes \mathbb{C}^n)$ given by $\mathcal{M}_n(X \otimes Y) = \mathcal{M}(X) \otimes Y$ is positive. We say that the operator is unital if $\mathcal{M}(\mathbb{I}) = \mathbb{I}$. Then, $\mathcal{N}$ is a channel if there is a completely positive, unital $\mathcal{N}^* : \mathcal{B}(\mathcal{H}') \rightarrow \mathcal{B}(\mathcal{H})$ such that $\text{Tr}(\mathcal{N}(X)Y) = \text{Tr}(X\mathcal{N}^*(Y))$ for all $X \in \mathcal{T}(\mathcal{H})$ and $Y \in \mathcal{B}(\mathcal{H})$. We omit any further characterization of these maps, including the celebrated results of Stinespring [136] and Kraus [207], and direct the interested reader to Ref. [206].

We once again consider the compact subset of a metric space $A \subset \mathbb{R}^p$ representing the space of parameters α that a learner attempts to estimate in the classical setting. Instead of a probability distribution over A , we will consider continuous-variable states defined with respect to this parameter space. We will consider pure states as square-integrable functionals $A \rightarrow \mathbb{C}$ defined at each point in A , thereby identifying each state $|\Psi\rangle \in \mathcal{H}$ with $\Psi \in L^2(A)$ and fixing our Hilbert space to be $\mathcal{H} \sim L^2(A)$. We define a partition of the subset A as a set $\{A_1, \dots, A_N\}$ of subsets that are non-intersecting and whose union recovers A : $\bigcup_i A_i = A$ and $A_i \cap A_j = \emptyset$ whenever $i \neq j$, for $i = 1, \dots, N$. For each subset in the partition, we define a space of states $\mathcal{H}^{(i)}$ containing all square-integrable functions having support only on A_i :

$$\mathcal{H}^{(i)} \sim L_i^2(A_i) := \{\Psi \in L^2(A) : \Psi(\alpha) = 0 \text{ for all } \alpha \notin A_i\}, \quad (5.65)$$

where L_i^2 may be thought of a L^2 with respect to a uniform measure having support only on A_i . Each $L_i^2(A_i)$ is embedded in $L^2(A)$ in a straightforward way, such that $\langle \Psi, \Phi \rangle = 0$

for $\Psi \in \mathcal{H}^{(i)}, \Phi \in \mathcal{H}^{(j)}$ whenever $i \neq j$ (the inner product is computed with respect to the measure $\mathbb{I}\{\alpha \in A_i\}$ on $\mathbb{R}^p$). As such, we write²

$$\bigoplus_{i=1}^N \mathcal{H}^{(i)} = \mathcal{H}, \quad \mathcal{H}^{(i)} \perp \mathcal{H}^{(j)} \text{ for } i \neq j. \quad (5.66)$$

We will define a set of pure states in $\mathcal{H}$ that are piecewise constant within a set of subspaces $P = \{\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(N)}\}$. This forms a finite-dimensional subspace of $\mathcal{H}$ which we define as

$$\mathcal{H}_P = \{|\Psi\rangle \in \mathcal{H}, \Psi(x) = \Psi(x') \text{ for all } x, x' \in A_i, i = 1, \dots, N\}. \quad (5.67)$$

Letting $|i\rangle \in \mathcal{H}$ be constant on $\mathcal{H}^{(i)}$ and zero elsewhere, every element of $\mathcal{H}_P$ may be written:

$$|\Psi\rangle = \sum_{i=1}^N c_i |i_P\rangle. \quad (5.68)$$

By identifying $|i_P\rangle$ with a piecewise constant function on $L_i^2(A)$, it follows that the expression on the right is bounded and normalizable, hence $|\Psi\rangle\langle\Psi|$ is a valid state. To construct bounds for the entanglement fraction task, we will need to reason about finite-dimensional subspaces of infinite-dimensional $\mathcal{H}$. For this, the following will be useful.

Lemma 16. Let $\mathcal{H}, \mathcal{H}'$ be separable Hilbert spaces and define finite-dimensional subspaces $\mathcal{K} \subset \mathcal{H}, \mathcal{K}' \subset \mathcal{H}'$. Let $\Pi_{\mathcal{K}} : \mathcal{H} \rightarrow \mathcal{H}$ be a projector onto the subspace $\mathcal{K}$ of an orthogonal decomposition $\mathcal{H} = \mathcal{K} \oplus \mathcal{K}^\perp$. Let $\mathcal{M} \in \mathcal{C}(\mathcal{K}, \mathcal{K}')$ be a fixed channel. Then, for any $X_H \in \mathcal{T}(\mathcal{H})$ such that $\Pi_{\mathcal{K}} X_H \Pi_{\mathcal{K}} := X_K \in \mathcal{T}(\mathcal{K})$, there exists some $\mathcal{N} \in \mathcal{C}(\mathcal{H}, \mathcal{H}')$ such that

$$\mathcal{N}(X_H) = \mathcal{M}(X_K) \quad (5.69)$$

Proof. We invoke the unitary representation of a channel $\mathcal{M}$. Assuming for now that $\dim(\mathcal{K}) = \dim(\mathcal{K}')$, there exists some choice of separable Hilbert space $\mathcal{E}$, unitary $U \in \mathcal{U}(\mathcal{K} \otimes \mathcal{E})$, and $\Omega \in \mathcal{D}(\mathcal{E})$ such that

$$\mathcal{M}(X_K) = \text{Tr}_{\mathcal{E}}(U(X_K \otimes \Omega)U^\dagger), \quad (5.70)$$

²For our setting it does not matter whether each subset A_i in the partition of A is open and/or closed. For a sense of closure, the reader may consider $L_i^2(A_i)$ to be defined with respect to the *closure* of A_i . This does not affect any of our statements, as $L_i^2(A_i)$ forms an equivalence class of functions on $A \subset \mathbb{R}^p$ up to a (well-behaved) intersection of the closures of A_i and A_j having dimension at most $p - 1$.

Then we choose an embedding $V \in \mathcal{U}(\mathcal{H} \otimes \mathcal{E})$ of U , e.g. $V = U \oplus W$ for some $W \in \mathcal{U}(\mathcal{K}^\perp \otimes \mathcal{E})$. Then, we find

$$\text{Tr}_{\mathcal{E}}(V(X_H \otimes \Omega)V^\dagger) := \mathcal{N}(X_H) \quad (5.71)$$

defines a valid quantum channel in $\mathcal{C}(\mathcal{H})$, and by construction $\mathcal{N}(X_H) = \mathcal{M}(X_K)$. If $\dim(\mathcal{K}) \neq \dim(\mathcal{K}')$, we instead choose a separable Hilbert space $\mathcal{E}'$ such that U is a unitary map from $\mathcal{K} \otimes \mathcal{E} \rightarrow \mathcal{K}' \otimes \mathcal{E}'$. The desired result follows instead for $\mathcal{M}(X_K) = \text{Tr}_{\mathcal{E}'}(U(X_K \otimes \Omega)U^\dagger)$. $\square$

Recall that for an ϵ -net we define $\hat{W} : A \rightarrow A$ to be the function that returns the index of some closest point in $\mathcal{W}(\epsilon, A, d)$ to r :

$$\hat{W}(r) \in \{w : \alpha_w = \arg \min_{\alpha_{w'} \in \mathcal{W}} d(r, \alpha_{w'})\}, \quad (5.72)$$

Since $\mathcal{W}$ is finite a minimum exists and any ties are broken in an arbitrary but consistent way.³

We now state an upper and lower bound on the entanglement fraction task, via reduction to the maximal singlet fraction in finite dimensions. Essentially the claim is that, a learner who acts locally via an *optimal* channel $\mathcal{D}$ on system B of a state σ_{RB} to maximize overlap with $\Pi|\Phi_\epsilon\rangle$ will, in the *best case* with respect to the initial state of RA , have a performance guarantee in terms of the conditional entropy of an appropriately chosen finite-dimensional system:

Theorem 17 (Entanglement fraction guarantee). Let $\mathcal{H}_R, \mathcal{H}_A, \mathcal{H}_B, \mathcal{H}_{\hat{A}} \simeq L^2(A)$ be separable Hilbert spaces $\mathcal{N} \in \mathcal{C}(\mathcal{H}_A, \mathcal{H}_B)$ be a fixed channel. Define the (unnormalized) state $|\Phi_\epsilon\rangle_{R\hat{A}} \in \mathcal{H}_{R\hat{A}}$ and $|\Psi\rangle_{RA} \in \mathcal{H}_{RA}$ according to:

$$|\Phi_\epsilon\rangle_{R\hat{A}} = \int_A dr \int_{\mathbb{B}_\epsilon(r)} d\hat{\alpha} |r\rangle_R \otimes |\hat{\alpha}\rangle_{\hat{A}} \quad (5.73)$$

$$|\Psi\rangle_{RA} = \int_A dr \int_A d\alpha \Psi_{RA}(r, \alpha) |r\rangle_R \otimes |\alpha\rangle_A \quad (5.74)$$

where $\mathbb{B}_\epsilon(r) := \{r' \in A : d(r, r') \leq \epsilon\} \cap A$ is the ϵ -ball centered at r . Then, letting $\Sigma_{RB} = (\mathbb{I} \otimes \mathcal{N})(|\Psi\rangle\langle\Psi|_{RA})$, we have

$$\sup_{|\Psi\rangle \in \mathcal{H}_{RA}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \log \langle \Phi | \Pi_{\mathbb{W}^2} (\mathbb{I} \otimes \mathcal{D}) (\Sigma_{RB}) \Pi_{\mathbb{W}^2} | \Phi \rangle_{R\hat{A}} \geq -H(R|B)_\rho \quad (5.75)$$

³For our analysis it should be that the set of possible values for $\hat{W}(r)$ is measure zero in A with respect to the Lebesgue measure anyways.

where $\Pi_{\mathbb{W}^2}$ is the projector onto the subset $\{(r, \alpha) \in A \times A : \hat{W}(r) = \hat{W}(\alpha)\}$, and $H(R|B)_\rho$ is computed with respect to $\rho_{RB} = (\mathbb{I} \otimes \mathcal{N})(|\phi\rangle\langle\phi|_{RA})$, the outcome of applying $\mathcal{N}$ to one half of a maximally entangled state $|\phi\rangle \in \mathbb{C}^{|\mathcal{W}|} \otimes \mathbb{C}^{|\mathcal{W}|}$.

Proof. The basic idea may be understood as a communication protocol between Alice (system RA) and Bob (system RB) through a noisy channel $\mathcal{N}$. While the parties may optimize over infinite-dimensional states and decoding channels respectively, they instead opt to use only finite dimensions - this results in a lower bound for their performance. The remaining details are in the specific sets of finite dimensional states (or channels, respectively) that Alice and Bob elect to use, which are informed by the ϵ -covering partition chosen to establish the information-theoretic bound of Proposition 14. The key details are as follows: states that are piecewise-constant on $C(A, \mathcal{W})$ are a subset of states in $\mathcal{H}_{RA}$, the set of channels $\mathcal{D}$ acting on $\mathcal{L}(\mathcal{H}_B)$ may be restricted to a subset acting nontrivially only on specific finite-dimensional subspaces, and the linearity of any fixed $\mathcal{N} \in C(\mathcal{H}_A, \mathcal{H}_B)$ restricts the dimension of the image of some subset of $D(\mathcal{H}_B)$ in a predictable way. Each of these properties is applied to bound either the infimum or supremum appearing in Theorem 17, until the problem is completely described by states and channels acting on embedded finite-dimensional subspaces. The technical details are in setting up the discretizations of the relevant Hilbert spaces and picking specific operators and states to achieve the desired bound. In Sec. 5.5.3 we have seen how this guarantee relates to classical learning, in that Theorem 17 resembles the classical learning guarantee of Proposition 14 except for the presence of a score function s .

For each $w \in [|\mathcal{W}|]$ we define $\mathbb{W}(w) := \{r : r \in A, \hat{W}(r) = w\} \subset A$ as the set of points in A for which the closest element of $\mathcal{W}$ has index w . By construction, $\bigcup_{w \in [|\mathcal{W}|]} \mathbb{W}(w) = A$ and $\mathbb{W}(w) \cap \mathbb{W}(v) = \emptyset$ for $w \neq v$, i.e. $\mathbb{W} := \{\mathbb{W}(w) : w \in [|\mathcal{W}|]\}$ is an ϵ -covering partition of A (c.f. Definition 13 of the main text). It will be convenient to use notation for describing elements of $L^2(A)$ in terms of the linear forms $\{|x\rangle : x \in \mathcal{X} \subseteq A\}$, e.g.

$$|w\rangle := \int_{\mathbb{W}(w)} dx |x\rangle. \quad (5.76)$$

In particular, we are interested in two subspaces:

1. $\mathcal{H}_{\mathbb{W}} \subset \mathcal{H}_{\hat{A}}$, defined as the subspace of states which are piecewise constant on the ϵ -covering partition $\mathbb{W}$ of A (see Eq. 5.67). These states have the form

$$|\psi\rangle_{\hat{A}} = \sum_{w=1}^{|\mathcal{W}|} \psi_w |w\rangle \quad (5.77)$$

2. $\mathcal{H}_{\mathbb{W}^2} \subset \mathcal{H}_{RA}$, defined as the subspace of bipartite states of the form

$$|\psi\rangle_{RA} = \sum_{w=1}^{|\mathcal{W}|} \psi_w |w\rangle_R \otimes |w\rangle_A. \quad (5.78)$$

Given that $\mathcal{H}_{\mathbb{W}} \simeq \mathbb{C}^{|\mathcal{W}|}$ and is therefore closed, we may partition $\mathcal{H}_{\hat{A}}$ into orthogonal subspaces, $\mathcal{H}_{\hat{A}} = \mathcal{H}_{\mathbb{W}} \oplus \mathcal{H}_{\mathbb{W}}^\perp$. Letting $\Pi_{\mathbb{W}}$ and $\Pi_{\mathbb{W}}^\perp$ be orthogonal projectors onto these respective subspaces, we have that $\Pi_{\mathbb{W}}|\psi\rangle = |\psi\rangle$ for $|\psi\rangle \in \mathcal{H}_{\mathbb{W}}$. We will similarly define a projector that preserves states in $\mathcal{H}_{\mathbb{W}^2}$. As A is a compact subset of $\mathbb{R}$, we may partition $A \times A$ into a finite number of bounded regions. We consider the subset of $A \times A$ containing coordinate pairs that share a closest point in an ϵ -net $\mathcal{W}(\epsilon, A, d) := \mathcal{W}$. We call this subset $K \subset A \times A$, defined as

$$K := \{(r, \alpha) \in A \times A : \hat{W}(r) = \hat{W}(\alpha)\}. \quad (5.79)$$

Define a projector $\Pi_{\mathbb{W}^2}$ onto the subspace of $\mathcal{H}_R \otimes \mathcal{H}_{\hat{A}}$ spanned by linear forms $|r\rangle \otimes |\alpha\rangle$ with $(r, \alpha) \in K$, i.e. all r, α that share a nearest point in $\mathcal{W}$. $\Pi_{\mathbb{W}^2}$ may be written in this way as

$$\Pi_{\mathbb{W}^2} = \int dr \int d\alpha \mathbb{I}\{\hat{W}(r) = \hat{W}(\alpha)\} |r\rangle\langle r| \otimes |\alpha\rangle\langle\alpha| \quad (5.80)$$

$$= \int dr \int d\alpha \sum_{w=1}^{|\mathcal{W}|} \mathbb{I}\{\hat{W}(r) = \hat{W}(\alpha) = w\} |r\rangle\langle r| \otimes |\alpha\rangle\langle\alpha| \quad (5.81)$$

$$= \int dr \int d\alpha \sum_{w=1}^{|\mathcal{W}|} \mathbb{I}\{\hat{W}(r) = w\} \mathbb{I}\{\hat{W}(\alpha) = w\} |r\rangle\langle r| \otimes |\alpha\rangle\langle\alpha| \quad (5.82)$$

$$= \sum_{w=1}^{|\mathcal{W}|} \int dr \int d\alpha \mathbb{I}\{\hat{W}(r) = w\} \mathbb{I}\{\hat{W}(\alpha) = w\} |r\rangle\langle r| \otimes |\alpha\rangle\langle\alpha| \quad (5.83)$$

$$= \sum_{w=1}^{|\mathcal{W}|} \int_{\mathbb{W}(w)} dr \int_{\mathbb{W}(w)} d\alpha |r\rangle\langle r| \otimes |\alpha\rangle\langle\alpha| \quad (5.84)$$

We then find,

$$\Pi_{\mathbb{W}^2}|\Phi_\epsilon\rangle = \sum_{w=1}^{|\mathcal{W}|} \int_{\mathbb{W}(w)} dr \int_{\mathbb{W}(w)} d\alpha \int_{\mathbb{B}_\epsilon(r)} d\alpha' |r\rangle \delta(\alpha - \alpha') |\alpha\rangle \quad (5.85)$$

$$= \sum_{w=1}^{|\mathcal{W}|} \int_{\mathbb{W}(w)} dr \int_{\mathbb{W}(w)} d\alpha |r\rangle |\alpha\rangle \quad (5.86)$$

$$= \sum_{w=1}^{|\mathcal{W}|} |w\rangle \otimes |w\rangle \quad (5.87)$$

$$:= |\mathcal{W}|^{1/2} |\phi\rangle_{R\hat{A}}, \quad (5.88)$$

which follows since $\mathbb{W}(w) \cap \mathbb{B}_\epsilon(r) = \emptyset$ whenever $\hat{W}(r) \neq w$, and

$$|\phi\rangle_{RA} := \frac{1}{|\mathcal{W}|^{1/2}} \sum_{w=1}^{|\mathcal{W}|} |w\rangle \otimes |w\rangle \quad (5.89)$$

is a maximally entangled state in $\mathcal{H}_{\mathbb{W}^2}$. With these operators in place, we find

$$\sup_{|\Psi\rangle \in \mathcal{H}_{RA}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \langle \Phi_\epsilon | \Pi_{\mathbb{W}^2} (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\Psi\rangle \langle \Psi|) \Pi_{\mathbb{W}^2} |\Phi_\epsilon\rangle \quad (5.90)$$

$$= \sup_{|\Psi\rangle \in \mathcal{H}_{RA}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} |\mathcal{W}| \langle \phi | (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\Psi\rangle \langle \Psi|) | \phi \rangle_{R\hat{A}} \quad (5.91)$$

$$\geq \sup_{|\psi\rangle \in \mathcal{H}_{\mathbb{W}^2}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} |\mathcal{W}| \langle \phi | (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\psi\rangle \langle \psi|) | \phi \rangle_{R\hat{A}} \quad (5.92)$$

$$\geq \sup_{|\psi\rangle \in \mathcal{H}_{\mathbb{W}^2}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_{\mathbb{W}})} |\mathcal{W}| \langle \phi | (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\psi\rangle \langle \psi|) | \phi \rangle_{R\hat{A}} \quad (5.93)$$

In Line 5.91 we have substituted Eq. 5.89, and Line 5.92 follows since $\mathcal{H}_{\mathbb{W}^2} \subset \mathcal{H}_{RA}$. For Line 5.93 we apply Lemma 16. Since $|\psi\rangle \in \mathcal{H}_{\mathbb{W}^2}$ is finite-dimensional, we may consider a restriction of $\mathcal{N}$ to $\mathcal{T}(\mathcal{H}_{\mathbb{W}})$ for some finite dimensional subspace $\mathcal{H}_{\mathbb{W}}^A \subset \mathcal{H}_A$, the image of which also lies in some finite-dimensional subspace (e.g. $\text{span}\{\mathcal{N}(|i\rangle\langle j|)\} \subset \mathcal{T}(\mathcal{H}_B)$, where $\{|i\rangle\} \subset \mathcal{H}_{\mathbb{W}}^A$ are an orthonormal basis), and this image has dimension no greater than $|\mathcal{W}|^2$. We then identify some $\mathcal{H}_{\mathbb{W}}^B \subset \mathcal{H}_B$ with dimension $\dim \mathcal{H}_{\mathbb{W}}^B = \dim \mathcal{H}_{\mathbb{W}} = |\mathcal{W}|$ such that $\mathcal{N}(\mathcal{T}(\mathcal{H}_{\mathbb{W}}^A))$ is contained in $\mathcal{T}(\mathcal{H}_{\mathbb{W}}^B)$. In other words, for $\mathcal{H}_B = \mathcal{H}_{\mathbb{W}}^B \oplus (\mathcal{H}_{\mathbb{W}}^B)^\perp$ and $X \in \mathcal{T}(\mathcal{H}_{\mathbb{W}}^A)$, and we can embed $\mathcal{N}(X) \in \mathcal{T}(\mathcal{H}_{\mathbb{W}}^B)$ into $X_B \in \mathcal{T}(\mathcal{H}_B)$ such that $\mathcal{N}(X) = \Pi_{\mathbb{W}'} X_B \Pi_{\mathbb{W}'}$ for an orthogonal projector $\Pi_{\mathbb{W}'}$ onto $\mathcal{H}_{\mathbb{W}}^B$. Finally, choosing $\mathcal{H}_{\mathbb{W}}^{\hat{A}} \subset \mathcal{H}_{\hat{A}}$ with

$\dim \mathcal{H}_{\mathbb{W}}^{\hat{A}} = \dim \mathcal{H}_{\mathbb{W}}$ and $\mathcal{H}_{\hat{A}} = \mathcal{H}_{\mathbb{W}}^{\hat{A}} \oplus (\mathcal{H}_{\mathbb{W}}^{\hat{A}})^{\perp}$, by Lemma 16 for any channel $\mathcal{D}' \in C(\mathcal{H}_{\mathbb{W}}^B, \mathcal{H}_{\mathbb{W}}^{\hat{A}})$ which has an extension $\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_{\hat{A}})$ such that $\mathcal{D}'(\mathcal{N}(X)) = \mathcal{D}(X_B)$, which implies the final inequality. Since $\dim \mathcal{H}_{\mathbb{W}}^{\hat{A}} = \dim \mathcal{H}_{\mathbb{W}}^B = \dim \mathcal{H}_{\mathbb{W}}^{\hat{A}}$, all of these subspaces are isomorphic to $\mathcal{H}_{\mathbb{W}} = \mathbb{C}^{|\mathbb{W}|}$ and we drop the superscripts for clarity. However, it is implicit that the finite-dimensional channels $\mathcal{D}$ are embedded in such a way to act only on finite-dimensional subsystems of infinite-dimensional states.

We now apply the definition of conditional min-entropy (Eq. 4.6) for channels $\mathcal{D} \in C(\mathcal{H}_{\mathbb{W}})$, we find

$$\sup_{|\Psi\rangle \in \mathcal{H}_{RA}} \sup_{\mathcal{D} \in C(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \langle \Phi_{\epsilon} | \Pi_{\mathbb{W}^2} (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\Psi\rangle\langle\Psi|) \Pi_{\mathbb{W}^2} | \Phi_{\epsilon} \rangle \quad (5.94)$$

$$= \sup_{|\psi\rangle \in \mathcal{H}_{\mathbb{W}^2}} \left(\sup_{\mathcal{D} \in C(\mathcal{H}_{\mathbb{W}})} |\mathcal{W}| \langle \phi | (\mathbb{I} \otimes \mathcal{D}) (\rho_{RB}) | \phi \rangle_{R\hat{A}} \right) \quad (5.95)$$

$$= \sup_{|\psi\rangle \in \mathcal{H}_{\mathbb{W}^2}} -H_{\min}(R|B)_{\rho} \quad (5.96)$$

$$\geq \sup_{|\psi\rangle \in \mathcal{H}_{\mathbb{W}^2}} -H(R|B)_{\rho} \quad (5.97)$$

with both entropies being taken with respect to $\rho_{RB} = (\mathbb{I} \otimes \mathcal{N})(|\psi\rangle\langle\psi|)$. This supremum is lower bounded by choosing any state in $\mathcal{H}_{\mathbb{W}^2}$, and in particular we choose $|\phi\rangle$ to recover the statement of the Theorem. $\square$

Theorem 17 demonstrates how we can upper bound the error of the entanglement fraction task for a best-case scenario in terms of the conditional entropy of a particular finite-dimensional quantum state. We now discuss how we may lower bound the error of this task for a worst-case scenario in a similar way. A key difference from the previous analysis is that the worst case entanglement fraction with respect to states $|\Psi\rangle_{RA} \in \mathcal{H}_{RA}$ may be made arbitrarily small by preparing a *separable* state of the form $|\Psi\rangle_{RA} = |\Psi_R\rangle \otimes |\Psi_A\rangle$. Since the learner's local operation $\mathcal{D}$ cannot generate entanglement between R and $\hat{A}$, the resulting entanglement fraction $Q_{\epsilon}(R|B)_{\Sigma}$ may approach zero. Therefore we may only derive a non-trivial error lower bound if the space of input states is somehow constrained.

We will therefore only consider input states that have sufficiently large overlap with a maximally entangled state in some finite-dimensional subspace of $\mathcal{H}_{RA}$. This guarantees that, neglecting the effect of $\mathcal{N}$, there is some fixed amount of entanglement initially present in the system. Briefly, for a state $|\Psi\rangle_{RA}$ we define $q_k(\Psi_{RA})$ to be the overlap of $|\Psi\rangle$ with $\frac{1}{\sqrt{k}} \sum_{i=1}^k |i_R\rangle \otimes |i_A\rangle$ for some choice of bases $\{|i_R\rangle\}_i \subset \mathcal{H}_R$ and $\{|i_A\rangle\}_i \subset \mathcal{H}_A$. To choose an appropriate k for the task, we will once again consider a kind of discretization on our

Hilbert space. As before, let $\mathcal{V} = \mathcal{V}(\epsilon, A, d)$ be an ϵ -packing of A in d . Letting $|v\rangle$ be constant within $\mathbb{B}_\epsilon(\alpha_v)$ and zero elsewhere, we can define a singlet-like state on $\mathcal{H}_{RA}$ as

$$|\phi\rangle_{RA} = \frac{1}{|\mathcal{V}|^{1/2}} \sum_{v=1}^{|\mathcal{V}|} |v\rangle \otimes |v\rangle \quad (5.98)$$

Given that this is a maximally entangled state in $\mathbb{C}^{|\mathcal{V}| \times |\mathcal{V}|}$, a natural choice for the maximal singlet fraction of the initial state $|\Psi\rangle_{RA}$ is to choose a subspace of dimension $k = |\mathcal{V}|$. As before, we consider overlap with a state $\Pi|\Phi_\epsilon\rangle$ of bounded entanglement. The following then provides a upper bound on the entanglement fraction:

Theorem 18 (Entanglement fraction bound, informal). Let $|\Phi_\epsilon\rangle$, $\mathcal{D}$, and $\mathcal{N}$ be as defined in Theorem 17, and define $q_k(\Psi)$ to be the singlet fraction of a state in $\mathcal{H}_{RA}$ computed in any $(k \times k)$ -dimensional subspaces of $\mathcal{H}_{RA}$. Define a subset $K := \{(r, \alpha) \in A \times A : \hat{V}(r) = \hat{V}(\alpha)\} \subset A \times A$, where $\hat{V}$ returns the nearest point to r in the ϵ -packing $\mathcal{V}$, and let $\Pi \in \mathcal{B}(\mathcal{H}_{RA})$ be a projector onto the subspace of $\mathcal{H}_{RA}$ spanned by unbounded linear functionals $|r\rangle \otimes |\alpha\rangle$ with $(r, \alpha) \in K$. Then,

$$\inf_{\substack{|\Psi\rangle_{RA} \in \mathcal{H}_{RA} \\ q_{|\mathcal{V}|}(\Psi)=1}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} (1 - |\mathcal{V}|^{-1} \langle \Phi_\epsilon | \Pi(\mathbb{I} \otimes \mathcal{D})(\Sigma) \Pi | \Phi_\epsilon \rangle) \geq \frac{H(RB)_\rho - 1}{\log(|\mathcal{V}|^2 - 1)} \quad (5.99)$$

where $\Sigma = (\mathbb{I} \otimes \mathcal{N})(|\Psi\rangle\langle\Psi|)$ and $\rho = (\mathbb{I} \otimes \mathcal{N})(|\phi\rangle\langle\phi|)$ for $|\phi\rangle$ defined in Eq. 5.98.

Proof. Similarly to Theorem 17, the LHS of Eq. 5.99 is analogous to $\Pr(d(\alpha, \hat{\alpha}) \geq \epsilon)$ in Proposition 11, without any reference to the loss function ℓ . The bounded expression (restricted to $[0, 1]$) can be interpreted as a normalized error, which becomes larger as the maximal entanglement fraction decreases. Applying an overall factor of $|\mathcal{V}|$ (corresponding to the chosen packing $\mathcal{V}$), Theorem 18 provides a lower bound for the error of the entanglement fraction task.

As mentioned already, the minimum of $\langle \Phi_\epsilon | (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N})(\Psi) | \Phi_\epsilon \rangle$ with respect to $|\Psi\rangle \in \mathcal{H}_{RA}$ will be trivially small, since the system RA could always be prepared in a separable state. Instead, we will consider a constrained minimization in which the system RA is guaranteed to have some fixed amount of entanglement, and then ask what the optimal performance of a quantum learner is for the worst-case initial state? We will now derive a bound for this task in terms of the conditional Von Neumann entropy of the system RB using the quantum Fano inequality.

To this end, we will define a specific entangled state in infinite dimensions, and then impose that $|\Psi\rangle_{RA}$ has some minimum overlap with that state. Consider two separable

Hilbert spaces $\mathcal{H}_R$ and $\mathcal{H}_A$ with orthonormal bases $\{|j_R\rangle\}$ and $\{|j_A\rangle\}$ respectively. For any fixed $k \in \mathbb{N}$, we consider an orthonormal set

$$S = \{|i_R\rangle \otimes |j_A\rangle : i \in \mathcal{I}, j \in \mathcal{J}\} \quad (5.100)$$

which is specified by sets of indices $\mathcal{I}, \mathcal{J}$ of size $|\mathcal{I}| = |\mathcal{J}| = k$ and spans a k^2 -dimensional subspace of $\mathcal{H}_{RA}$. In this subspace there is a maximally entangled state $|\Omega^S\rangle \in \mathcal{H}_{RA}$ of the form

$$|\Omega^S\rangle := \frac{1}{\sqrt{k}} \sum_{j=1}^k |j_R\rangle \otimes |j_A\rangle. \quad (5.101)$$

The maximal singlet overlap of any $\Psi \in \mathcal{H}_{RA}$ with some maximally entangled state in $\text{span}(S)$ is then defined as

$$q_k(\Psi, S) = \sup_{\mathcal{M} \in \mathcal{C}(\mathcal{H}_A)} \langle \Omega^S | (\mathbb{I} \otimes \mathcal{M})(\Psi) | \Omega^S \rangle \quad (5.102)$$

From this, we may define a notion of singlet fraction for Ψ as the largest achievable singlet fraction, with respect to all subsets S :

$$q_k(\Psi) := \sup_{\substack{S \subset \mathcal{H}_{RA} \\ |S|=k^2}} q(\Psi, S), \quad (5.103)$$

where the supremum is to be taken over choices of bases for $\mathcal{H}_R$ and $\mathcal{H}_A$ as well as sets $\mathcal{I}$ and $\mathcal{J}$. In this way, $q(\Psi)$ describes the maximum overlap of Ψ with *some* maximally entangled state in any $k \times k$ -dimensional subspace of $\mathcal{H}_{RA}$.

To prove Theorem 18, we follow a similar procedure as for the proof of Theorem 17, namely we focus on states that are piecewise constant within the ϵ -balls that constitute an ϵ -packing $\mathcal{V} := \mathcal{V}(\epsilon, A, d)$. We begin by defining the function $\hat{V} : A \rightarrow A$ that identifies the ϵ -ball containing $r \in A$:

$$\hat{V}(r) \in \{v : \alpha_v = \arg \min_{\alpha_{v'} \in \mathcal{V}} d(r, \alpha_{v'})\}, \quad (5.104)$$

and as before, ties are irrelevant and may be broken in some consistent way. For each $\alpha_v \in \mathcal{V}$, we define $\mathbb{V}(v) := \mathbb{B}_\epsilon(\alpha_v)$ the ϵ -ball at α_v . Unlike the scenario of Theorem 17, the set of ϵ -balls composing $\mathcal{V}$ does not form a partition of A . An (unnormalized) state that is piecewise constant on $\mathbb{V}$ is

$$|v\rangle := \int_{\mathbb{V}(v)} dx |x\rangle. \quad (5.105)$$

We define the subspace $\mathcal{H}_{\mathbb{V}} \subset \mathcal{H}_{\hat{A}}$ containing the set of states which are piecewise constant on the ϵ -packing $\mathcal{V}$ of A , $|\psi\rangle_{\hat{A}} = \sum_{v=1}^{|\mathcal{V}|} \psi_v |v\rangle$. Similarly, $\mathcal{H}_{\mathbb{V}^2}$ is the space of bipartite states of the form $|\psi\rangle_{RA} = \sum_{v=1}^{|\mathcal{V}|} \psi_v |v\rangle_R \otimes |v\rangle_A$. In particular, the subspace $\mathcal{H}_{\mathbb{V}^2} \subset \mathcal{H}_{R\hat{A}}$ contains the maximally entangled state

$$|\phi\rangle_{R\hat{A}} = \frac{1}{|\mathcal{V}|^{1/2}} \sum_{v=1}^{|\mathcal{V}|} |v_R\rangle \otimes |v_{\hat{A}}\rangle \quad (5.106)$$

Again, we will use the projector $\Pi_{\mathbb{V}^2}$ onto the set $K \subset A \times A$ containing all coordinates in the same ϵ -ball:

$$K_{\mathbb{V}^2} := \{(r, \alpha) \in A \times A : \hat{V}(r) = \hat{V}(\alpha)\} \quad (5.107)$$

In this setting, we may find a non-trivial infimum by requiring Ψ to have a singlet fraction $q_k(\Psi) = F$ with some k -dimensional singlet. Intuitively, this means that the adversary initially prepares RA in an entangled state, without imposing any specifics on the basis or subspace in which that state is entangled.

Similarly to Eq. 5.89, we may compute

$$\Pi_{\mathbb{V}^2} |\Phi_{\epsilon}\rangle = \sum_{v \in \mathcal{V}} |v\rangle_R \otimes |v\rangle_{\hat{A}} := |\mathcal{V}|^{1/2} |\phi\rangle_{R\hat{A}} \quad (5.108)$$

For any fixed S , since $q_k(\Psi, S) = 1$ implies $q_k(\Psi) = 1$ we have the inclusion

$$\{|\Psi\rangle \in \mathcal{H}_{\mathbb{V}^2} : q_k(\Psi, S) = 1\} \subseteq \{|\Psi\rangle \in \mathcal{H}_{RA} : q_k(\Psi) = 1\}. \quad (5.109)$$

With this, the proof proceeds almost identically to that of Theorem 17:

$$\inf_{\substack{|\Psi\rangle \in \mathcal{H}_{RA} \\ q_k(\Psi)=1}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \langle \Phi_{\epsilon} | \Pi_{\mathbb{V}^2} (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\Psi\rangle \langle \Psi|) \Pi_{\mathbb{V}^2} | \Phi_{\epsilon} \rangle \quad (5.110)$$

$$\leq \inf_{\substack{|\psi\rangle \in \mathcal{H}_{\mathbb{V}^2} \\ q_k(\psi)=1}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \langle \Phi_{\epsilon} | \Pi_{\mathbb{V}^2} (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\psi\rangle \langle \psi|) \Pi_{\mathbb{V}^2} | \Phi_{\epsilon} \rangle \quad (5.111)$$

$$\leq \inf_{\substack{|\psi\rangle \in \mathcal{H}_{\mathbb{V}^2} \\ q(\psi, \mathcal{H}_{\mathbb{V}})=1}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_B, \mathcal{H}_{\hat{A}})} \langle \Phi_{\epsilon} | \Pi_{\mathbb{V}^2} (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\psi\rangle \langle \psi|) \Pi_{\mathbb{V}^2} | \Phi_{\epsilon} \rangle \quad (5.112)$$

$$= \inf_{\substack{|\psi\rangle \in \mathcal{H}_{\mathbb{V}^2} \\ q(\psi, \mathcal{H}_{\mathbb{V}})=1}} \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_{\mathbb{V}})} \langle \Phi_{\epsilon} | \Pi_{\mathbb{V}^2} (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\psi\rangle \langle \psi|) \Pi_{\mathbb{V}^2} | \Phi_{\epsilon} \rangle \quad (5.113)$$

$$\leq \sup_{\mathcal{D} \in \mathcal{C}(\mathcal{H}_{\mathbb{V}})} |\mathcal{V}| \langle \phi | (\mathbb{I} \otimes \mathcal{D} \circ \mathcal{N}) (|\phi\rangle \langle \phi|) | \phi \rangle \quad (5.114)$$

$$= q(R|B)_{\rho} \quad (5.115)$$

$$\leq |\mathcal{V}| \left(1 - \frac{H(RB)_{\rho} - 1}{\log(k^2 - 1)} \right) \quad (5.116)$$

In Line 5.111 we have used $\mathcal{H}_{V^2} \subset \mathcal{H}_{RA}$, and in Line 5.112 we have used the inclusion of Eq. 5.109. For Line 5.113 we apply Lemma 16 in an identical manner as in the proof of Theorem 17, and Line 5.114 follows by specifying $|\psi\rangle = |\phi\rangle$ as defined in Eq. 5.108, followed by the quantum Fano inequality (Proposition 15) in Line 5.116. Specifying $k = |\mathcal{V}|$ completes the proof.

□

5.7 Discussion

In this work, we have demonstrated how classical learning can be understood as a special case of a particular entanglement manipulation task in a continuous variable quantum system. The task that we have introduced is distinct from the typical setting of quantum machine learning, in that it involves coherent manipulation of quantum information at all stages. We derived upper and lower bounds for the error of this task via reduction to maximal singlet fraction in finite dimensions, in an analogous fashion to the reduction of classical learning to multi-hypothesis testing.

A central challenge remains in *applying* learning guarantees (such as Proposition 14) and analogous entanglement fraction guarantee (Theorem 17). Applying the former relies on upper bounding the conditional entropy $H(W|B)$ of a variable indexed by the ϵ -net for a particular distribution of observations $p_{B|A}$. For context, when a learner's observations are independent samples $B = X_1, \dots, X_n$, then minimax bounds of the form of Proposition 11 may be combined with the convexity of relative entropy to lower bound $H(V|B)$ in terms of marginal entropies [156]. Applying the entanglement fraction guarantee may require identifying a problem setup where similar simplifications hold.

There are key differences between the entanglement fraction bounds and their classical analogues, with significant technical challenges appearing in the former case. For instance, there is no obvious analogue for the loss function ℓ and score function s applied to the distance $d(\hat{\alpha}, \alpha)$ in the classical setting. One might introduce some unbounded operator $|L\rangle\langle L| \in \mathcal{L}(\mathcal{H}_{R\hat{A}})$ to fulfill an analogous role to the loss function ℓ , for instance defining $|L\rangle$ such that $\langle \alpha, \hat{\alpha} | L \rangle = \ell^{1/2}(d(\alpha, \hat{\alpha}))$ for $\alpha, \hat{\alpha} \in A$. However, the non-commutativity of quantum theory obscures how the expected value of such an observable is related to a state's overlap with $|\Phi_\epsilon\rangle$. Future work may explore alternative generalizations of classical learning that can better accommodate loss functions. In Sec. 5.5.3 we discussed how the entanglement fraction task reduced to a classical learning task, albeit one involving quantum input data. A reduction to the purely classical case might involve technical

hurdles. While applying the completely dephasing map Δ to the system RB in the finite-dimensional case (Eq. 5.45) is straightforward, a similar procedure could not be followed for the infinite-dimensional case due to the unboundedness of the map Δ_A . Equivalently, evaluating $\langle \Delta_A \circ (\mathbb{I} \otimes \mathcal{D}^*)(\Phi_\epsilon) \rangle_{(\mathbb{I} \otimes \mathcal{N})\Psi_{RA}}$ involves products of unbounded operators, which raises difficulties. In any case, we have imposed a reduction to classical learning by exploiting the commutativity of measurements with respect to a spectral measure constructed from linear functionals $\{|\alpha\rangle\}_{\alpha \in A}$, at the expense of constraining measurement statistics in a conjugate basis. A more complete theory should ideally involve genuinely non-classical measurement statistics, i.e. constraints on the measurements of incompatible observables – one reasonable place to start might be in the theory of learning with continuous variable quantum kernels, e.g. Ref [208].

In this chapter we have focused on a specific type of learning task involving parameter estimation, which differs from the task of predicting the value of a target function applied to a set of observations – another typical task in machine learning. Our analysis generalizes to the case when the observations are produced by a function f_α parameterized by $\alpha \in A$. However, in more general learning models the data may be produced by a process for which there is no explicit target function f_α (as is the case in agnostic PAC learning [209]). It remains an open question how we might view more general prediction tasks through a similar framework as the one developed here.

Chapter 6

Perturbative readout error mitigation for near-term quantum computers

6.1 Chapter overview

This chapter switches focus completely to consider classical techniques for mitigating errors in experiments run on real quantum devices. In particular, readout errors on near-term quantum computers can introduce error to the empirical probability distribution sampled from the output of a quantum circuit. These errors can be mitigated by classical post-processing given access to an experimental *response matrix* that describes the error associated with measurement of each computational basis state. However, the resources required to characterize the response matrix and to compute the corrected probability distribution scale exponentially in the number of qubits n .

In this chapter, we modify standard matrix inversion techniques using perturbative approximations with significantly reduced complexity and bounded error when the likelihood of high order bit flip events is strongly suppressed. Secs. 6.2-6.3 provide background context for the readout inversion problem. Given a characteristic error rate q , Sec. 6.4 introduces a method to recover the probability of the all-zeros bit string p_0 by sampling only a small subspace of the response matrix before inverting readout error resulting in a relative speedup of $\text{poly}(2^n/\binom{n}{w})$, which we motivate using a simplified error model for which the approximation incurs only $O(q^w)$ error for some integer w . Sec. 6.5 then provides a generalized technique to recover full output distributions with $O(q^w)$ error in the perturbative limit.

This chapter is reproduced directly from Ref. [4]. I originally developed the perturbative technique for all-zeros bitstrings appearing in Sec. 6.4. Andy C. Y. Li generalized and formalized this technique to recovering a complete distribution, and we wrote Sec. 6.5 together. I primarily wrote the remaining sections, with Gabe Perdue and Andy both contributing during the drafting of the manuscript.

6.2 Introduction

While quantum computing will potentially provide an exponential speed up in solving certain problems, noisy intermediate-scale quantum (NISQ) [210] devices are subject to high error rates that must be mitigated in order to extract useful information from the quantum processors. Readout error is unique among the standard sources of decoherence since it is well modelled by a *classical* stochastic process and is therefore entirely reversible by classical post-processing. In the simplest approach, the effects of readout error can be reversed by inverting a *response matrix* R that relates pre-measurement computational basis states to bitstrings sampled by the measurement, provided that R is nonsingular and accurately characterizes the readout error dynamics.

Previous works in readout error mitigation typically involve some variation of inverting a Markovian process [11, 211–216]. The goal of these post-processing techniques is to recover the entire probability mass function $p(x)$ over bitstrings $x \in \{0,1\}^n$. However, these techniques are generally not scalable as they require experimental characterization of a response matrix followed by an (approximate) matrix inversion step, requiring device time and computing resources that grow exponentially in n . Bayesian Iterative Unfolding [217, 218] avoids the latter hurdle by approximating the matrix inversion and readout rebalancing [219] improves on the accuracy of readout error correction for recovering high-weight bitstrings by biasing measurements based on some prior knowledge of the support of $p(x)$ on $\mathbb{R}^{2^n}$. However, unless error mitigation is applied to recover a specific observable [220–222] both techniques still generally require an exponentially large device time to characterize the response matrix. Furthermore, with limited exceptions (e.g., ref. [6]), few of these techniques have been specialized for the case where only a single bitstring probability is desired, which requires significantly fewer resources to mitigate readout error.

In this chapter, we present a perturbative technique for approximately correcting readout error on near-term quantum computers. Intuitively, the technique relies on an assumption that the likelihood of a readout error event involving many simultaneous bit flips (for instance, observing the bitstring 1111 after a computational basis measurement of the state $|0000\rangle$) is strongly suppressed in the number of simultaneous bit flips. This includes

scenarios for which the bit flips are weakly correlated between different qubits and the individual bit flip rates are small, which is often the case on existing devices [223].

We introduce our technique by considering the task of recovering the probability of the all-zeros bitstring p_0 and show that this may be accomplished using only a small submatrix of $\mathbf{R}$, and we provide numerical and theoretical evidence justifying this approximation. By tailoring experimental determination of $\mathbf{R}$ towards recovering a specific bitstring even in the presence of correlated readout errors, this approach offers a potential performance advantage over existing techniques designed to recover full distributions.

We then present the general technique to approximately recover the full empirical bitstring probability distribution by perturbatively expanding $\mathbf{R}$ in terms of a characteristic readout error rate q . This technique represents a middle ground between full matrix inversion of $\mathbf{R}$ and sparsity-based techniques. It is well-suited for mitigating readout error when both the strength of correlations between readout errors is known and the distribution $p(x)$ has non-trivial support on a large number of bitstrings (e.g. superpolynomial in n), such that the probability of observing each bitstring is influenced by the underlying probabilities for many other bitstrings that are close in Hamming distance. Both variants of our technique allow for probabilities to be *approximately* corrected with the benefit of greatly reduced error correction overhead, and are therefore especially well suited for experiments in which readout error is not the limiting factor in the accuracy of the sampled probabilities.

6.3 Inverting readout error

Given an n -qubit state represented by its density matrix ρ , the error in a projective measurement $\{|i\rangle\langle i|\}$ for $i = 0, \dots, 2^n - 1$ over the computational basis states can be modelled as a classical Markovian process [26], which is described by the equation

$$p' = \mathbf{R} p. \quad (6.1)$$

Here, $\mathbf{R}$ is a $2^n \times 2^n$ matrix with nonnegative entries whose columns sum to one (i.e. a left stochastic matrix), $p = \text{diag}(\rho)$ is a length- 2^n normalized array of probabilities measured in computational basis without measurement noise, and p' is the length- 2^n array of observed (erroneous) bitstring probabilities. The response matrix $\mathbf{R}$ may be defined elementwise in terms of transition likelihoods,

$$R_{ij} \equiv p(i|j) = p(i_1 \dots i_n | j_1 \dots j_n) \quad (6.2)$$

where $i, j \in \{0, 1\}^n$ are length- n bitstrings, and the notation i_k is understood to refer to the k -th bit of i . If we are provided with an invertible $\mathbf{R}$, a basic prescription for correcting readout error is to compute

$$p = \mathbf{R}^{-1}p' \quad (6.3)$$

In practice, $\mathbf{R}$ may be singular and a least squares approximation to the linear equation (6.1) may be used.

Even when $\mathbf{R}$ is invertible, computing Eq. 6.3 in a general setting requires two distinct, resource intensive steps: (i) measuring the complete response matrix of bitstring transition probabilities using a diagnostic experiment to determine $\mathbf{R}$, with time complexity $\mathcal{O}(2^n)$ and (ii) performing matrix inversion on $\mathbf{R}$, which can be as costly as $\mathcal{O}(2^{3n})$ ¹. Notably, recent works have explored more efficient sparsity-based techniques for mitigating readout error, for example inverting the response matrix in the subspace spanned by nonzero components of p' [224, 225].

A small infidelity in the readout error mitigation can usually be tolerated as a trade-off for an improved complexity scaling in many cases, for example, when the readout error is less significant compared to other sources of error such as decoherence. We now introduce heuristic techniques for reducing the resource requirements of both of these steps while incurring some small, controllable error that may be inferred under some mild assumptions about the structure of $\mathbf{R}$.

6.4 Recovering the probability of all-zeros bitstring

We first study a special case of our technique for complexity reduction when one only desires to determine the probability of the all-zeros bitstring $p_0 = \text{Tr}(|0^n\rangle\langle 0^n|\rho)$. This scenario is relevant for near-term algorithms such as quantum kernel methods [11, 12], dual-state purification [226], qubit assignment on hardware [227], and quantum circuit learning [75, 228]. In this context, Eq. 6.3 can be cast in the form of a dot product,

$$p_0 = r \cdot p' \quad (6.4)$$

where the vector $r \in \mathbb{R}^{2^n}$ is defined elementwise as $r_i = (\mathbf{R}^{-1})_{0i}$. One would expect that a simplified readout error mitigation can be performed to recover the observable p_0 with both

¹For simplicity, we will assume both inversion and multiplication of generic $M \times M$ matrices inversion proceeds with complexity $\mathcal{O}(M^3)$. Optimized algorithms like Strassen's algorithm reduce this complexity but these considerations will not affect the *relative* speedups that we present in this chapter.

tight error bounds and greater efficiency than for recovering the full distribution p , since only the subspace of $\mathbf{R}^{-1}$ that describes likely transitions into and out of 0^n is relevant. Assuming that the probability of a bitstring transition falls monotonically in the number of individual bits flipped, this subspace corresponds to the set of probabilities $(\mathbf{R})_{ij}$ for which i and j are low-weight strings.

Following this intuition, our proposed technique works by correcting p_0 using the inverse of a projection of $\mathbf{R}$ onto the subspace of low-weight basis vectors. The weight of a binary bitstring $x = x_1x_2 \dots x_n \in \{0, 1\}^n$ is defined as

$$w(x) = \sum_{i=1}^n x_i \quad (6.5)$$

which is the number of 1's appearing in x . We denote the set of all bitstrings with weight less than w as

$$S_w = \{x : x \in \{0, 1\}^n, w(x) \leq w\} \subseteq \{0, 1\}^n. \quad (6.6)$$

We then define the weight projection operator $P_w : \mathbb{R}^d \rightarrow \mathbb{R}^{|S_w|}$ that projects vectors onto the subspace spanned by basis vectors whose binary index is in S_w . This is equivalent to the action

$$P_w \hat{e}_j = \begin{cases} \hat{e}_j, & \text{if } j \in S_w, \\ 0, & \text{otherwise.} \end{cases} \quad (6.7)$$

where $\hat{e}_j$ is the j -th unit vector. Then, a $d \times d$ matrix A may be projected onto the same subspace by the operation $P_w A P_w$. Defining $\mathbf{R}_T = P_w \mathbf{R} P_w$ and the first row of its inverse as $(r_T)_i = (\mathbf{R}_T^{-1})_{0i}$ by analogy with Eq. 6.4, our goal is to demonstrate that for some choice of $w < n$ and mild assumptions about the structure of $\mathbf{R}$, we can compute

$$\tilde{p}_0 = r_T \cdot p'_T \quad (6.8)$$

as a close approximation to p_0 , where $p'_T = P_w p'$. Eq. 6.8 simply uses the first row r_T of the inverse of a truncated response matrix in place of r applied to a truncated observed probability vector p'_T . Applying Eq. 6.8 consumes a significantly smaller response matrix $\mathbf{R}_T$ which may be understood conceptually as the top-left submatrix of $\mathbf{R}$ with rows and columns resorted by index weight. $\mathbf{R}_T$ has dimensions $t_w \times t_w$ given by the sum over binomial coefficients

$$t_w = \sum_{j=0}^w \binom{n}{j} \quad (6.9)$$

and so the readout error correction for $\tilde{p}_0$ may be carried out by sampling $\mathbf{R}_T$ from a quantum processor with complexity $\mathcal{O}(t_w)$ and then computing $\mathbf{R}_T^{-1}$ with complexity $\mathcal{O}(t_w^3)$.

Here, we neglect the effects sampling error in both R and p' , which introduces a constant overhead for readout error correction as a function of the strength of noise on the device [221]. In the absence of statistical effects (such that Eq. 6.4 is satisfied) the error of our method is given as

$$|r_T \cdot p'_T - p_0| = |r_T \cdot p'_T - r \cdot p'| \quad (6.10)$$

This approach therefore introduces error from two different sources: the first kind of error results from computing p after discarding bit flip events involving more than w simultaneous relaxations and excitations, while the second kind of error results is due to the truncation approximation $R_T^{-1} - P_w R^{-1} P_w$. To motivate our technique, we proceed study situations for which this difference vanishes and provide the resulting bounds on $|r_T \cdot p'_T - r \cdot p'|$ for progressively looser restrictions on the structure of R .

6.4.1 Exact bounds for a relaxation-only model

It is convenient to use the convention that vectors and matrices be sorted according to the weight of the binary representation of the index, with indices of equal weight sorted arbitrarily. For example with $n = 3$, this has the effect of rearranging the vector of readout probabilities such that

$$p = (p_{000}, p_{001}, p_{010}, p_{100}, p_{110}, \dots)^T \quad (6.11)$$

We now consider an instructive toy model for readout error for which an analytical upper bound on the error $|p_0 - r \cdot p'|$ may be derived exactly. In this model, R is both overly simplified and trivially invertible, but the analysis will provide insight into approximations for situations where R has a more complex structure. The model for readout error that we study analytically is an extreme example of asymmetric readout error described by a response matrix of the form

$$R = \bigotimes_{k=1}^n Q_k \quad (6.12)$$

$$Q_k = \begin{pmatrix} 1 & q \\ 0 & (1 - q) \end{pmatrix} \quad (6.13)$$

for $0 \leq q < 0.5$. We can make very strong arguments about readout error arising from this model.

Proposition 19. Let R be defined as in Eq. 6.12. For any fixed projector P_w satisfying $P_w^2 = P_w$, define $R_T = P_w R P_w$. Then,

$$P_w (R^{-1}) P_w = (R_T)^{-1} \quad (6.14)$$

In other words, for this definition of R the inverse of the projected response matrix R_T is equal to a projection of R^{-1} . This is a straightforward property of the kinds of upper triangular matrices we are interested, but an intuitive proof is provided in Appendix B.1. The following theorem applies Proposition 19 to show that we can compute only a very small subspace of R and invert that subspace to apply readout error correction to recover p_0 with error that is exponentially suppressed in our choice of truncation weight w .

Theorem 20. Let R be defined as in Eq. 6.12. Then the error introduced by correcting readout error using a truncated response matrix is bounded by:

$$|r_T \cdot p'_T - r \cdot p'| \leq (2q)^{w+1} \quad (6.15)$$

where r_T and r are defined elementwise as $(r_T)_i = (R_T^{-1})_{0i}$ and $r_i = (R^{-1})_{0i}$.

The proof is given in Appendix B.2. We remark that this is the tightest possible bound given the structure assumed of R that does not incorporate additional information about the readout probability distribution over the truncated subspace. Theorem 20 makes a simple but powerful observation that given the restricted noise model we have considered, one can apply readout error using the inverse of a projected matrix R_T such that the truncation error introduced is exponentially suppressed in w . This bound becomes quite weak in the limit that $q \rightarrow 0.5$ since the dimension of the truncated matrix itself grows combinatorially in w . Conversely, for $q \ll 1$ such that $q^2 \rightarrow 0$, this result guarantees that a matrix projected onto the $w = 1$ weight subspace with size *linear in n* can recover the probability of 0 with an accuracy almost as good as using the exponentially large R . Since this result is based only on the structure of R and not the value of elements contained therein, we can immediately lift some of the restrictions on constructing R .

Corollary 21. Let R have a tensor structure composed of distinct individual qubit response matrices of the following form:

$$R = \bigotimes_{k=1}^n \begin{pmatrix} 1 & q_k \\ 0 & 1 - q_k \end{pmatrix} \quad (6.16)$$

for $0 \leq q_k < 0.5$. Then

$$|r_T \cdot p'_T - r \cdot p'| \leq (2q_{max})^{w+1} \quad (6.17)$$

where $q_{max} = \max_k \{q_k\}$.

This is shown in Appendix B.2. Corollary 21 expands on the intuition of Theorem 20: If k -th order simultaneous bit flip events are suppressed exponentially in k , then we need only sample a submatrix of R to perform good readout correction. We can further extend this line of reasoning to its practical limit in a somewhat less rigorous way. Suppose R is *any* response matrix that allows only for “relaxation” events, that is R may be defined elementwise as

$$R_{ij} = \begin{cases} p(i|j) & \text{for } w(i) < w(j) \text{ or } i = j, \\ 0 & \text{otherwise.} \end{cases} \quad (6.18)$$

If we assume that the probability of a relaxation event is suppressed exponentially in the number of simultaneous bit flips, i.e. $p(i|j) \leq \mathcal{O}(q^{w(j)-w(i)})$ for some characteristic rate q then the above bounds still hold in the approximate sense:

$$|r_T \cdot p'_T - r \cdot p'| \lesssim \mathcal{O}(2q^{w+1}) \quad (6.19)$$

This follows directly from Theorem 20; each entry with magnitude exactly $(1-q)^{w(x)}q^{w(y)}$ in the strictly upper triangular part of R can be replaced with approximate term with order $\mathcal{O}((1-q)^{w(x)}q^{w(y)})$. As this modification does not affect the structure of R , similar nilpotency and series expansion arguments that lead to Theorem 20 may be applied by substituting $q \rightarrow \mathcal{O}(q)$. In this situation R can no longer be decomposed and therefore R^{-1} can no longer be efficiently computed as $\bigotimes_k Q_k^{-1}$ using individual qubit response matrices $\{Q_k\}_{k=1}^n$. This extension also marks a departure from Corollary 21 by relaxing the assumption that R has a tensor structure, and therefore accommodates weakly correlated readout errors. Despite this structural change, the projected R_T^{-1} constructed from events with order no greater than w still serves as a useful surrogate for R^{-1} if only elements in the first row of R^{-1} are desired, and provides some justification for extending the reasoning of Eq. 6.8 to the more general case.

6.5 Perturbative mitigation for recovering full distributions

In the previous section, projecting R onto a subspace of low-weight indices was motivated by a model for readout error that penalizes the transition of high-weight bitstrings into 0^n . This reasoning can be generalized to recovering the full bitstring distribution p more efficiently, assuming an error model that penalizes transitions between any two bitstrings that differ by a large hamming weight. This is a practical model even when there is

correlated readout error between different qubits, provided the correlation strength is not comparable to the characteristic rate. This model is further motivated by the observation that correlations in readout error are likely to be strong only among qubits that are physically adjacent on a device, for example nearest neighbors on a two-dimensional grid of superconducting qubits [229].

To proceed, we assume there is a characteristic rate q that describes the probability of any given single bit flip event. For each $j = 1, \dots, n$, we define a sparse $2^n \times 2^n$ off-diagonal matrix R_j whose entries are of magnitude $\mathcal{O}(1)$. Then, without loss of generality, we define R with respect to a series structure such that

$$R = R_0 + \sum_{j=1}^{2^n} q^j R_j, \quad (6.20)$$

where each R_j contains a subset of elements of R according to some pairwise comparison function s ,

$$(R_j)_{nm} = \begin{cases} (R)_{nm} & \text{if } s(n, m) = j, \\ 0 & \text{otherwise.} \end{cases} \quad (6.21)$$

Here we will focus on the specific choice

$$s(n, m) = w(n \oplus m) \quad (6.22)$$

where w is the weight function of Eq. 6.5 and $\oplus$ denotes the bitwise modulo-2 sum. Then, R_0 consists of the diagonal matrix elements of R of all orders of q and R_j consists of the off-diagonal terms of the order q^j describing all likelihoods involving bit flips whose weights differ by j . Note that under this definition, R_0 will describe all bit flip events of even order in which the observed bitstring is identical to the prior bitstring, and similarly for $j = 1, \dots, n$ so that R_j does not necessarily characterize events involving exactly j bit flips. Even if R is not well modelled by this choice of decomposition, such as in cases involving strongly correlated readout errors, it still may be possible to define R_j to include all events with probability on order q^j . This will require significant prior knowledge about the scale of readout errors on the device and is out of scope for this work.

The form of Eq. 6.20 suggests that the R_j corresponding to small j will dominate the effects of dynamics error. Applying this intuition, we expand the inverse of Eq. 6.20 as a series,

$$R^{-1} = R_0^{-1} + \sum_{k=1}^{\infty} \left(- \sum_{j=1}^{2^n} q^j R_0^{-1} R_j \right)^k R_0^{-1}. \quad (6.23)$$

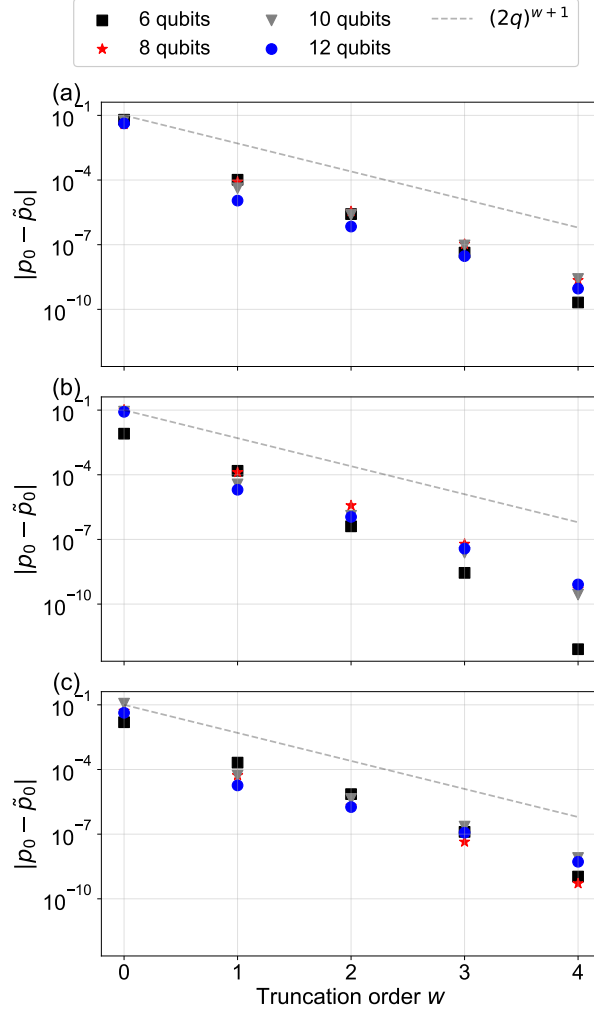


Figure 6.1: Performance of all-zeros readout error mitigation given by Eq. 6.8 compared for different prior distributions. (a) The Gaussian prior is centered at 0 with n -bit overflow for all bitstrings with value less than 2^{n-1} , i.e. $p_j \propto \exp((x_j - 0.5)^2/\sigma^2)$ where $x_j \equiv 2^{-n}((j + 2^{n-1}) \bmod 2^n)$ and $\sigma = 0.25$. This distribution is adversarial to recovering p_0 as it has significant support on the high-weight subspace. (b) The truncated Gaussian is given by the same distribution without overflow ($x_j = j \cdot 2^{-n}, \sigma = 0.25$) and renormalized, and the (c) uniform distribution is $p_j = 2^{-n}$. In all plots, $w = 0$ is defined to correspond to the uncorrected probability p'_0 . The dashed line indicates the bound of Eq. 6.19 derived for a relaxation-only model, which we observed was not violated even for the more general error model of Eq. 6.30.

Truncating both series to the order of q^w results in

$$\mathbf{R}^{-1} = \left[1 + \sum_{k=1}^w \left(- \sum_{j=1}^w q^j \mathbf{R}_0^{-1} \mathbf{R}_j \right)^k \right] \mathbf{R}_0^{-1} + \mathcal{O}(q^{w+1}). \quad (6.24)$$

Note that this expression contains some terms of order higher than q^w but the error introduced by the truncation remains bounded by $\mathcal{O}(q^{w+1})$. Applying this expression to p' , we arrive at an approximate probability distribution $\tilde{p}$ given by

$$\tilde{p} = \left[1 + \sum_{k=1}^w \left(- \sum_{j=1}^w \mathbf{R}_0^{-1} \mathcal{R}_j \right)^k \right] \mathbf{R}_0^{-1} p'. \quad (6.25)$$

where we have introduced $\mathcal{R}_j \equiv q^j \mathbf{R}_j$ as the matrix of elements describing order- j transitions sampled directly from an experimental response matrix (for which knowledge of the rate q is not strictly necessary). This result can be viewed as a generalization of the specialized task described in Eq. 6.8, which we discuss in Appendix B.3.

The implementation of the perturbative readout error mitigation to recover an empirical distribution over bitstrings sampled from a quantum computer subject to measurement error is described by the following pseudocode.

Algorithm 1: Perturbative mitigation for the full distribution

Input: $p', \{\mathcal{R}_j\}_{j=0,1,\dots,w}$
1 $\mathbf{S} \leftarrow - \sum_{j=1}^w \mathbf{R}_0^{-1} \mathcal{R}_j$;
2 $v \leftarrow \mathbf{R}_0^{-1} p'$;
3 $\tilde{p} \leftarrow v$;
4 **for** $k \leftarrow 1$ **to** w **do**
5 $v \leftarrow \mathbf{S} v$;
6 $\tilde{p} \leftarrow \tilde{p} + v$;
7 **end**

We note that ref. [230] also employed series approximations for computing $\mathbf{R}^{-1}$ but the implementation is otherwise unrelated to the technique described here. If $\mathbf{R}^{-1}$ exists, then the Neumann series introduced in Eq. 6.23 converges only if

$$\left\| \sum_{j=1}^w \mathbf{R}_0^{-1} \mathcal{R}_j \right\| < 1,$$

which determines whether Algorithm 1 can be applied in its given form. If this condition is met, then the error $\epsilon = \|p - \tilde{p}\|_2$ introduced by Algorithm 1 is concentrated in the $(w + 1)$ -th order terms, resulting in an approximate error given by $\epsilon \lesssim 2q^{w+1} + \mathcal{O}(q^{w+2})$ where we have applied the slightly stronger assumption that $\|\mathbf{R}_0^{-1}\mathbf{R}_j\| \leq 1$. To reach an accuracy with an error ϵ , we need to implement the algorithm with an order at least

$$w \geq \left\lceil \frac{\ln \epsilon^{-1} + \ln 2}{\ln q^{-1}} \right\rceil - 1. \quad (6.26)$$

The complexity of this technique is dominated by matrix products involving $\mathbf{S}$ in Algorithm 1. If $\mathbf{S}$ is sparse with s being roughly the fraction of elements that are non-zero, the algorithm requires w matrix product operations resulting in approximate time complexity given by

$$O(s_w w M^3). \quad (6.27)$$

We now compute the sparsity of $\mathcal{R}_w$ to determine the relative speedup of this technique over standard matrix inversion. The nonzero elements of $\mathcal{R}_j$ occur at all index pairs (x, y) satisfying $x \oplus y = j$, where $x, y \in B$ and $B = \{0, 1\}^n$. The number of pairs (x, y) that satisfy this condition is equivalent to the number of strings x satisfying $x = z \oplus y$, where $z \in B_j$ and $B_j = \{s : s \in B, w(s) = j\}$ is the set of all weight- j bitstrings. This number is given by $|B| \times |B_j|$, and so the sparsity factor s_w describing the number of nonzero terms in $\mathcal{R}_w$ may be computed directly as $2^{-2n} \times |B| \times |B_j|$, or

$$s_w = 2^{-n} \binom{n}{w}. \quad (6.28)$$

This sparsity is computed using the union of Hamming balls around each weight- k bitstring, i.e. our construction implicitly assumes readout errors that are only weakly correlated such that transitions between bitstrings with large Hamming distance are suppressed.

In the context of Eq. 6.27, s_w has the effect of replacing one term proportional to 2^n with a term proportional to $w \binom{n}{w}$. The core speed-up therefore comes generating $\tilde{p}$ in Algorithm 1 by a series of sparse matrix-vector products using matrices with at most s_w nonzero terms. If convergence of the Neumann series of Eq. 6.23 is not guaranteed, a perturbative technique may still be useful. In this case, an experimentalist would still measure the set $\{\mathcal{R}_j\}$ to a desired truncation point w , and then directly invert the resulting approximation to $\mathbf{R}$ to recover

$$\tilde{p} = \left(\sum_{j=1}^w \mathcal{R}_j \right)^{-1} p' \quad (6.29)$$

This may result in significant speedup compared to sampling the full $\mathbf{R}$, but incurs additional computational cost to compute a standard matrix inverse, and we explore this tradeoff in Appendix B.4. This approach might therefore be compatible with other techniques that avoid directly computing a matrix inverse (for instance, Bayesian iterative unfolding [217–219]) or with techniques for efficient inversion of sparse, banded matrices [231, 232].

This approach allows us to safely ignore the small response matrix elements corresponding to higher-order terms. In experiments, each column of $\mathbf{R}$ may be estimated by preparing the state corresponding to that column and then computing the output bitstring distribution for a computational basis measurement, resulting in an estimate for the matrix elements of the corresponding column of $\mathbf{R}$. By discarding the higher-order terms, this distribution can be reliably determined with a number of shots like $n_{\text{meas}} \sim 1/q^{2w}$. If a sufficient truncation order w is known either from previous experiments or knowledge of the hardware design (e.g. Ref. [223]), our technique saves resources by allowing the experimentalist to omit measurement of higher-order terms. Moreover, the resource requirement may be reduced further if only a few bitstrings of the output distribution are required. This includes, for example, the all-zeros bitstring case discussed in Sec. 6.4 and cases where only a specific qubit excitation sector is of interest due to symmetry. In these cases, we need only estimate columns of $\mathbf{R}$ by computing elements corresponding to the desired bitstring population (see Appendix B.3). This allows us to compute $\mathbf{R}$ using fewer experiments on quantum hardware.

6.6 Numerical experiments

We implemented the technique introduced in the preceding sections for a variety of prior distributions and readout error strengths. To avoid complications due to statistical uncertainty, we restrict ourselves to using p and $\mathbf{R}$ that were simulated to floating point precision without introducing any sampling error. We generated each $\mathbf{R}$ randomly in the following manner: We constructed a tensor product of the form

$$\mathbf{R} = \bigotimes_{k=1}^n \begin{pmatrix} 1 - \eta_k & \epsilon_k \\ \eta_k & 1 - \epsilon_k \end{pmatrix} \quad (6.30)$$

with $\epsilon_k, \eta_k \sim \text{Uniform}(0, q)$, and then we randomly permuted each weight- k subspace of $\mathbf{R}$ for $k = 1, \dots, n - 1$. The resulting matrix is not separable and therefore cannot be trivially inverted by inverting each term in the Kronecker product of Eq. 6.30. We applied

the resulting linear map to a prior distribution p to generate p' . Fig. 6.1 demonstrates the performance of Eq. 6.8 to correct for p_0 for different prior distributions. The error in the method is suppressed exponentially in the truncation order w , which is consistent with behavior that was analytically derived for a more restricted error model in Sec. 6.4. A similar exponential suppression of the error can be observed in experiments using response matrices measured on IBM QPUs (Appendix B.5.1). In Appendix B.5.2 we provide preliminary comparison of our method to the M3 technique of Ref. [224].

For Algorithm 1, we are interested in assessing the performance of the readout correction for recovering the full distribution p' compared to p . To compare the two distributions, we compute the trace distance (or L1 norm),

$$d(p, q) = \sum_{j \in \{0,1\}^n} |p_j - q_j|. \quad (6.31)$$

This has a useful interpretation in terms of computing expectation values of Hermitian operators. Let O be an operator with 2^n real entries on the diagonal bounded by $O_j \in [-1, 1]$. The expected value $\langle O \rangle = \text{Tr}(O\rho)$ corresponding to the corrected distribution $\tilde{p}$ is estimated using the quantity

$$E_O = \sum_{j \in \{0,1\}^n} O_j \tilde{p}_j. \quad (6.32)$$

We can then show that $d(p, \tilde{p})$ bounds the error incurred in E_O :

$$|\text{Tr}(O\rho) - E_O| \leq d(p, \tilde{p}), \quad (6.33)$$

and so d serves as a natural comparison between output distributions that will be postprocessed to compute observables. Fig. 6.2 shows the performance of Algorithm 1 for varying truncation orders, numbers n of qubits and increasing noise rates q . For modest q , the effects of readout error can be strongly suppressed using only a fraction of the resources required for inverting all of R . As n or q increases, the performance of the algorithm rapidly drops off as the series for truncated inverse requires significantly more terms to converge. Fig. 6.2 also provides a visual example of Algorithm 1 applied using a second order truncation, which will generally consume $\mathcal{O}(n^2)$ resources for sampling R_2 to recover $\tilde{p}$.

6.7 Discussion

We have proposed a technique for approximately correcting readout error in quantum computers requiring significantly less overhead than traditional matrix inversion techniques,

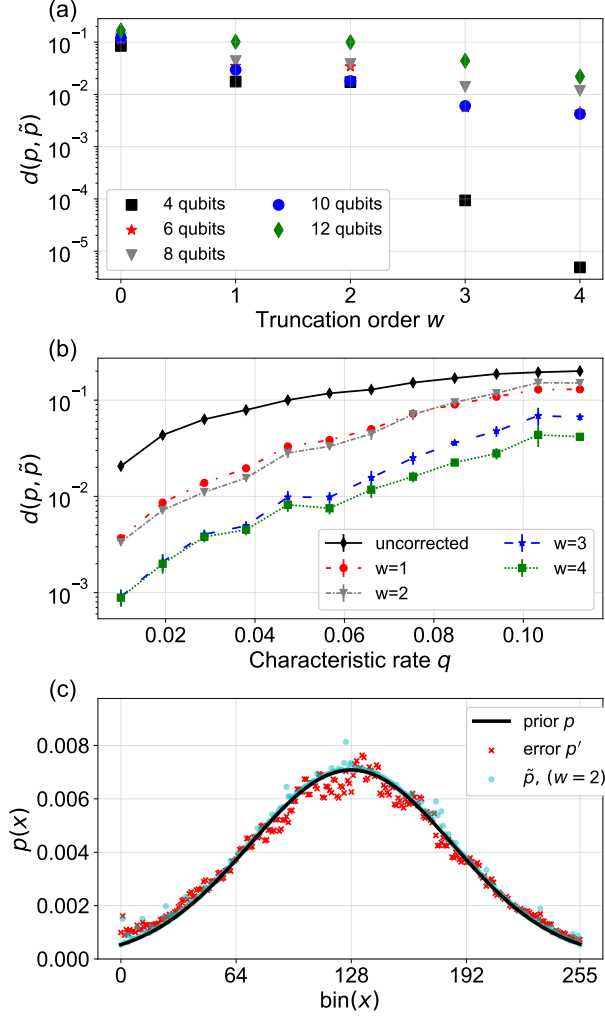


Figure 6.2: (a) Performance of Algorithm 1 as a function of truncation order w , for fixed $q = 0.05$. The elements of the initial distribution p were each drawn from $\text{Uniform}(0, 1)$ and then normalized. We define $w = 0$ to represent the uncorrected case corresponding to $d(p, p')$. (b) The performance of the algorithm ($n = 8$) diminishes with increasing q , which corresponds to the series approximation for R^{-1} diverging. We discuss the resulting limitations and workarounds for this behavior in Appendix B.4. (c) Visualization of applying correction to a Gaussian distribution for $n = 8, w = 2$ with a characteristic rate $q = 0.6$.

while still capturing enough of the readout error behavior to correct distributions with support on a large number of bitstrings that might present a challenge for sparsity-based techniques. Such approximations are beneficial when the error in the correction scheme becomes negligible compared to other sources of device noise, and so this technique may be useful for running quantum algorithms on near-term devices. We have justified the technique in the perturbative regime and also provided numerical evidence suggesting that this technique can be useful if either the characteristic error rate q or the number of qubits n remains small (e.g. nq does not grow too large - see Appendix B.4). Future work may further generalize the bounds we have derived and elucidate the non-perturbative regimes for which the errors in our methods remain well bounded.

Chapter 7

Qubit assignment using time reversal

7.1 Chapter overview

Continuing with the topic of Chapter 6, this chapter will discuss techniques for improving the performance of experiments run on near-term (i.e. not fault-tolerant) devices simply by using better qubits. As quantum computers with large numbers of qubits become increasingly available, experiments executed on a given device may not utilize all available qubits. In this case, the outcome of executing a quantum program will depend on the ability to efficiently select a subset of high-performing physical qubits. For any given quantum program and device there are many ways to assign physical qubits for execution of the program, and assignments will differ in performance due to the variability in quality across qubits and entangling operations on a single device. Evaluating the performance of each assignment using fidelity estimation introduces significant experimental overhead and will be infeasible for many applications, while relying on standard device benchmarks provides incomplete information about the performance of any specific program. Furthermore, the number of possible assignments grows combinatorially in the number of qubits on the device and in the program, motivating the use of heuristic optimization techniques.

This chapter presents a practical solution to the problem of qubit assignment by using simulated annealing with a cost function based on the Loschmidt Echo, a diagnostic that measures the reversibility of a quantum process. Secs. 7.2-7.3 provide background context for this problem, as well as as theoretical justification for this choice of cost function by demonstrating that the optimal qubit assignment coincides with the optimal qubit assignment based on state fidelity in the weak error limit. Sec. 7.4 provides experimental justification using diagnostics performed on Google’s superconducting qubit devices, and

establishes the performance of simulated annealing for qubit assignment using classical simulations of noisy devices as well as optimization experiments performed on a quantum processor.

This chapter is mostly reproduced from Ref. [5]. I was primarily responsible for the theory and ideas developed in this chapter, and for running and analyzing simulations and experiments appearing here and in Appendix C. Andy C. Y. Li proved Proposition 22 as it appears here, though I have included only my proof of the special case in Appendix C.1.1. Prasanth Shyamsundar proposed and developed early code implementations of simulated annealing for qubit assignment. Andy, Prasanth, and Gabe Perdue all contributed to drafting the manuscript that is reproduced here.

7.2 Introduction

With the increased availability of NISQ [25] devices there is a growing need for tools to efficiently deploy quantum circuits on hardware in a noise-aware manner. Part of this task is qubit assignment [233, 234], where the goal is to assign a logical circuit ¹ to the set of physical qubits on noisy hardware that maximizes the circuit performance. This task requires access to a performance metric for circuits implemented on hardware that is both efficient to evaluate and faithful to standard fidelity metrics, as well as a means of efficiently optimizing that performance metric with respect to the set of all possible qubit assignments.

Performing qubit assignment in a way that satisfies these requirements faces a number of challenges. For instance, an experimentalist might choose physical qubits according to which subset maximizes the fidelity of the output of the quantum program, but this introduces an experimental overhead that is exponential in the system size on near-term devices. On the other hand, choosing physical qubits based on standard device benchmarks such as randomized benchmarking [235, 236] or cross-entropy benchmarking [237, 238] can result in poor-performing assignments since these benchmarks capture average error behavior that may differ from the noise occurring in the context of a specific quantum program. Once a performance metric is chosen, there still remains the challenge of efficiently exploring

¹In the context of this chapter, a “logical circuit” is an abstraction used in the specification of a sequence of gates acting on qubits that have no correspondence to any physical device. Meanwhile, “physical qubits” exist on a specific device and are subject to connectivity constraints and errors. Qubit assignment is the process of selecting a correspondence between each logical qubit and some physical qubit that respects connectivity constraints. This differs from the logical-vs-physical distinction typically used in the context of quantum error correction (QEC).

a space of qubit assignments that grows combinatorially both in the size of the quantum program and the hardware device. Implementing qubit assignment techniques can improve device performance whenever the total number of qubits on the device is larger than the number required for an application. This scenario is typical for near and intermediate-term devices, particularly in the context of quantum error correction experiments [239–243].

To overcome these challenges we study the Loschmidt Echo [244], a tool for probing reversibility in quantum systems. We then demonstrate that this metric can be used with simulated annealing (SA) [28] to effectively perform qubit assignment on hardware. The large state space of potential hardware circuits combined with high variance in qubit error rates naturally leads to an uneven cost landscape, for which SA is particularly well suited. Combining SA with Loschmidt Echoes then provides a scalable technique for qubit assignment that does not rely on potentially inaccurate hardware diagnostics.

Prior works have used variants of the Loschmidt Echo to assess the performance of qubit assignments. Refs. [245, 246] introduced techniques for benchmarking quantum devices using “mirror circuits” composed entirely of Clifford operations, for which the fidelity of output state may be computed with the aid of efficient classical simulations of Clifford circuits [247, 248]. While this technique provides a benchmark for the performance of a set of qubits with respect to a general set of operations, we are interested in the ability of the device to accurately prepare some specific state $|\psi\rangle$. Furthermore, while it is possible to use mirror circuits to evaluate the performance associated with specific circuits, its application to large circuits is limited due to the lack of efficient classical simulation of general non-Clifford circuits. Moreover, we will demonstrate theoretically how generic benchmarks (which assess qubit and gate performance independently of circuit structure) have limited ability to predict the performance of a given qubit assignment. We will provide experimental evidence to support this claim for a specific choice of generic benchmark.

Similarly, this chapter differs from prior qubit assignment experiments [233, 249] since we do not rely on a set of gate fidelities characterizing hardware performance as a proxy for assessing the performance of a specific circuit on hardware. Therefore, while Ref. [233] used a hybrid algorithm involving SA to search over “sub-allocations” of progressively larger qubit subsets, in our approach we altogether avoid using a graph weighted by gate fidelities of (partial) circuits to anneal over.

Qubit assignment [250–254] (often referred to as qubit routing, qubit allocation, or quantum compilation) has been extensively studied as a tool for improving the performance of circuits executed on hardware. Qubit assignment typically includes modifying logical circuits to run on hardware when the gateset and connectivity constraints of the device do not match those of the logical program, typically with the goal of minimizing the number

of additional operations introduced to the program. For instance, Ref. [255] introduced a technique for circuit compilation using SA with a cost function based on CNOT count. However, in this chapter we simplify qubit assignment on hardware by only considering a fixed circuit that already satisfies hardware connectivity, with the goal of maximizing the performance of a quantum program with respect to the choice of hardware qubit subsets satisfying this connectivity.

This chapter proceeds as follows. In Sec. 7.3.1 we overview the problem of qubit assignment on a hardware device. In Sec. 7.3.2 we provide theoretical evidence that the Loschmidt Echo is useful for ranking qubit assignments when the goal is to maximize the fidelity of a prepared state, and demonstrate shortcomings of other existing methods based on benchmark data or random circuit fidelities in this same context. Specifically, we show that scoring qubit assignments based on gate fidelities taken from device benchmark data fails to capture general performance trends. In Sec. 7.3.3 we then develop a framework for performing qubit assignment using SA, and in Sec 7.4 we verify the performance of the Loschmidt echo and demonstrate optimization results that outperform competitive techniques on both simulated and experimental datasets.

7.3 Methods

7.3.1 Qubit assignment on hardware

We describe the task of qubit assignment in terms of a hardware topology and a set of gates [256]. We will provide a graph-based description of a quantum circuit as a sequence of gates applied to single qubits or pairs of qubits in a logical circuit, and a representation of qubit assignment as a transformation subject to connectivity constraints on a hardware device.

We describe the connectivity of a quantum device by an undirected graph $G_p = (V_p, E_p)$ where each vertex $i \in V_p$ describes a physical qubit and each edge $\{i, j\} \in E$ indicates that entangling operations can be executed between qubits i and j . We assume $|V_p| = N$ qubits are available on hardware and that the hardware supports a gateset $\mathcal{G}_p$ consisting of single- and two-qubit gates which allows for universal computation [257] and is a common choice on superconducting qubit platforms [258, 259]. We similarly define a logical graph $G_L = (V_L, E_L)$ and a logical circuit $\mathcal{C}_L$ over n qubits as a sequence of m gate operations $[g_1(v_1, e_1), \dots, g_m(v_m, e_m)]$, with $e_k \in E_L$, $v_k \in V_L$, and $g_k \in \mathcal{G}_L$ for $k = 1 \dots m$. Taking every gate g_k in $\mathcal{C}_L$ to act on either $(i, \emptyset)$ (single-qubit gate) or $(\emptyset, \{i, j\})$ (two-qubit gate),

we can summarize the logical circuit $\mathcal{C}_L \in (\mathcal{G}_L^m, V_L^m, E_L^m)$ according to three sequences, one for each gate description, one for each set of target qubits, and one for each set of target edges. For example a circuit might look like

$$\begin{aligned} \mathbf{g} &= [R_x(\theta), \text{CNOT}, X, \dots] \\ \mathbf{v}_L &= [\{3\}, \emptyset, \{3\}, \dots] \\ \mathbf{e}_L &= [\emptyset, \{1, 4\}, \emptyset, \dots] \end{aligned} \tag{7.1}$$

Let $U(\mathcal{C}_L)$ be the unitary representation of $\mathcal{C}_L$ acting on n unique qubits and let $U(\mathcal{C}_p)$ be the unitary representation of a hardware program $\mathcal{C}_p \in (\mathcal{G}_p^{m'}, V_p^{m'}, E_p^{m'})$ acting on up to $N \geq n$ qubits. Then defining the state $|\psi\rangle = U(\mathcal{C}_L)|0^n\rangle$ we are interested in searching over the range of a logical to physical circuit map of the form

$$M : (\mathcal{G}_L^m, V_L^m, E_L^m) \rightarrow (\mathcal{G}_p^{m'}, V_p^{m'}, E_p^{m'}) \tag{7.2}$$

subject to the constraint

$$|\psi\rangle\langle\psi| = \text{Tr}_A (U(\mathcal{C}_p)|0^N\rangle\langle 0^N|U(\mathcal{C}_p)^\dagger)$$

for some subsystem A consisting of up to $N - n$ qubits, in which case we will say that M satisfies $M(\mathcal{C}_L) = \mathcal{C}_p$. In general, the effects of noise on the quantum device will interfere with perfect realization of the unitary operation $U(\mathcal{C}_p)$. We therefore assume that there is a quantum channel $\mathcal{E}(\rho; \mathcal{C}_p)$ that describes the effect of $\mathcal{C}_p$ applied to an input state ρ on hardware. We will consider implementing a specific unitary U on noisy hardware and therefore restrict our attention to circuits $\mathcal{C}_p$ for which $U(\mathcal{C}_p) = U$ up to some permutation of subsystems and we will hereafter implicitly assume dependence on some hardware circuit $\mathcal{C}_p$ (and therefore some qubit assignment) using the notation $\mathcal{E}_U(\rho) = \mathcal{E}(\rho; \mathcal{C}_p)$.

7.3.2 Loschmidt Echoes for evaluating qubit assignments

We are interested in selecting the best qubits for preparing a state $|\psi\rangle = U|0\rangle$ over n qubits given a choice of N hardware qubits with constrained topology. Ideally we would be able to directly compare the state prepared by a noisy implementation $\mathcal{E}_U$ to the perfect preparation of $|\psi\rangle = U|0\rangle$. One way to characterize the hardware performance is to employ the fidelity function [260], which provides a measure of closeness between two quantum states ρ and σ given as $\mathcal{F}(\rho, \sigma) = \left(\text{Tr} \sqrt{\rho^{1/2} \sigma \rho^{1/2}}\right)^2$. Using $\mathcal{F}$ we can compare

the prepared state $\rho = \mathcal{E}_U(|0\rangle\langle 0|)$ (which implicitly depends on a hardware circuit $\mathcal{C}_p$) to an ideal state prepared by U in a noiseless environment:

$$F(\mathcal{C}_p) := \mathcal{F}(\rho, |\psi\rangle\langle\psi|) = \langle\psi|\rho|\psi\rangle \quad (7.3)$$

In general, evaluating F on a near-term device can incur significant resource overhead. Computing F using Direct Fidelity Estimation (DFE) [261, 262] can incur significant experimental overhead for each choice of qubits, and is therefore impractical to evaluate for qubit assignment on large quantum devices. Similarly, classical shadows [263] can be used to evaluate F in constant time, but evaluating fidelities in this way requires significant overhead and the ability to execute operations drawn randomly from the n -qubit Clifford group — a requirement that is out of reach for near-term devices due to the depth requirements of such circuits. We are therefore motivated to find a proxy for circuit fidelity that can be executed quickly and efficiently, at the expense of accurately characterizing the true state fidelity or process fidelity. Given the ability to implement unitaries U and V , the fidelity between the states $|\psi\rangle = U|0\rangle$ and $|\phi\rangle = V|0\rangle$ can be determined in constant time since the quantity

$$\mathcal{F}(|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi|) = |\langle\psi|\phi\rangle|^2 = |\langle 0|U^\dagger V|0\rangle|^2 \quad (7.4)$$

can be estimated as the empirical probability of observing the all-zeros bitstring 0^n at the output of a circuit $U^\dagger V$. By extension, to test the effects of noise we choose to evaluate a performance metric for $\mathcal{C}_p$ defined as

$$F_{\text{LE}}(\mathcal{C}_p) = \text{Tr}(|0\rangle\langle 0|\mathcal{E}_{U^\dagger} \circ \mathcal{E}_U|0\rangle\langle 0|) \quad (7.5)$$

where $\mathcal{E}_{U^\dagger}$ represents the process resulting from a noisy implementation of a unitary $U^\dagger$ on hardware. Eq. 7.5 computes the probability that a state prepared by implementing U can be successfully returned to $|0\rangle$ by implementing $U^\dagger$, though this will not provide an accurate estimate for the state fidelity F in general. This process is referred to as a Loschmidt Echo [27, 244], and serves as a general measure for reversibility in a quantum process (see Ref. [264] for a thorough review). In the special case where $\mathcal{E}_{U^\dagger}(\rho) = U^\dagger \rho U$ can be implemented with perfect fidelity, comparison to Eq. 7.3 shows that this procedure exactly recovers $F(\mathcal{C}_p)$. From this perspective, the application of $\mathcal{E}_{U^\dagger}$ can be understood as preparing an imperfect measurement of the projector $|\psi\rangle\langle\psi|$, a comparison that suggests that F_{LE} may be an effective tool for analyzing the implementation of U on hardware. Fig. 7.1 shows a circuit which estimates F_{LE} as the probability of sampling the all-zeros bitstring at the circuit output.



Figure 7.1: (left) Logical circuit for sampling $|\langle 0|U^\dagger U|0\rangle|^2$, which will equal unity in the absence of device error. (right) schematic representation for sampling $F_{\text{LE}}(\mathcal{C}_p) = \langle 0|\mathcal{E}_{U^\dagger} \circ \mathcal{E}_U(|0\rangle\langle 0|)|0\rangle$, where curved edges indicate a completely positive trace preserving (CPTP) map acting on n -qubit density operators.

While F_{LE} will generally not provide an estimate for the state fidelity F , we now show that given some assumptions about device noise, the optimal qubit assignment with respect to maximizing F_{LE} will result in a state with high fidelity as computed by Eq. 7.3. We first provide a concrete statement of this result by assuming a noise model consisting of one-qubit and two-qubit local depolarizing channels parameterized by the error rate ϵ_i for each single-qubit gate acting on qubit i and the error rate η_{ij} for each two-qubit gate acting on qubits $\{i, j\}$. This model is applied to the hardware implementation of each hardware circuit $\mathcal{C}'_p = (\mathbf{g}, \mathbf{v}', \mathbf{e}')$ with a fixed gate sequence $\mathbf{g}$ of length m . We then define the sets of optimal qubit assignments with respect to F and F_{LE} respectively as:

$$S^* = \{S' : S' = \arg \max_{\mathcal{C}'_p: U(\mathcal{C}'_p)=U} F(\mathcal{C}'_p)\} \quad (7.6)$$

$$S_{\text{LE}}^* = \{S' : S' = \arg \max_{\mathcal{C}'_p: U(\mathcal{C}'_p)=U} F_{\text{LE}}(\mathcal{C}'_p)\} \quad (7.7)$$

where the fixed gate sequence means that each maximum is computed with respect to all possible, valid assignments on the hardware device, i.e. sequences of qubits $\mathbf{v} \in (V')^m$ and sequences of edges $\mathbf{e} \in (E')^m$ subject to $V' \subset V_p$ and $E' \subset E_p$ for a device with connectivity described by (V_p, E_p) . Under the assumed noise model, we can then compare these optimal assignments in a low-error limit by the following proposition:

Proposition 22. For a graph $G_p = (V_p, E_p)$ describing hardware qubits, consider an error model consisting of Markovian noise channels, each characterized by an error rate ϵ , inserted after each gate. Suppose each gate U and its inverse implementation $U^\dagger$ are augmented by the same noise channel. Let the optimal qubit assignments S_{LE}^* and S^* be defined as in Eqs. 7.6-7.7. Then in the weak-error limit $\epsilon \ll 1$,

$$S_{\text{LE}}^* = S^* \quad (7.8)$$

We prove this claim for the special case of local depolarizing noise (Appendix C.1.1) and global depolarizing noise (Appendix C.1.2) (Ref. [5] contains a proof of this proposition

for more general noise models). In Appendix C.2 we show that F_{LE} is strongly biased by readout error compared to the true fidelity of the prepared state, and so it is necessary to implement readout error mitigation [211, 213, 218, 221, 224] and in particular, readout correction to recover the all-zeros bitstring [4] when selecting qubits based on this metric.

Under the assumption of a 2-local depolarizing error model discussed in Appendix C.1.1, it is natural to consider simpler metrics that might be used to estimate the performance a qubit assignment without the need for circuit-specific hardware experiments. For instance, the fidelity of random circuits exposed to discrete errors can be well approximated using device diagnostic data [265], and so in addition to computing F_{LE} and F for each qubit assignment, we also consider a heuristic that can be computed by combining the structure of the executed circuit with device benchmark data:

$$F_0(\boldsymbol{\epsilon}, \boldsymbol{\eta}) = \prod_{i \in V_p} (1 - \epsilon_i)^{n_i} \prod_{\{i,j\} \in E_p} (1 - \eta_{ij})^{n_{ij}} \quad (7.9)$$

where n_i counts the number of single qubit gates acting on qubit i , n_{ij} counts the two-qubit gates acting on edge $\{i, j\}$, the length- $|V_p|$ vector $\boldsymbol{\epsilon}$ contains weights for the graph vertices, and the length- $|E_p|$ vector contains weights for the graph edges. These weights describe error rates for single-qubit gates and two-qubit gates respectively. We omit a term characterizing measurement error from F_0 since we assume that readout error correction will be employed when estimating F and F_{LE} .

Assuming the 2-local depolarizing noise model, F_0 is the probability that an error does not occur during the execution of U on noisy hardware with intrinsic polarization error rates defined by $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$. This suggests that an experimentalist could perform device diagnostics such as randomized benchmarking (RB) [235, 236] or cross-entropy benchmarking (XEB) [237, 238] to estimate the gate error rates as $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\eta}}$ and then compute $F_0(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\eta}})$ to evaluate the performance of any given qubit assignment using Eq. 7.9.

This strategy has been used in previous works for qubit assignment [6, 233, 266] (the approaches of Refs. [267, 268] involved finding a shortest path on a graph with weights $\{\eta_{ij}\}$, though the operational meaning of a sum of error rates is not as clear compared to the product formula for F_0). However we argue that this strategy must be approached with caution since it neglect information about how device noise affects a specific circuit. Even in the scenario of a circuit exposed to depolarizing noise where an experimentalist is given perfect knowledge of depolarizing error rates, we show in Appendix C.1 that the optimal qubit assignment with respect to F_0 will not necessarily maximize F . Errors in experimental characterization of device performance (e.g. due to drift [269]) will then only compound this limitation. Our experimental results in Sec. 7.4.2 will demonstrate that F_0

is almost useless as a proxy for determining the relative performance of qubit assignments on actual hardware.

In our experiments we also consider performing qubit assignment with respect to an average Loschmidt survival rate $\langle F_{LE}^{rand} \rangle$ computed over a series of random circuits. Random circuits have been used to successfully evaluate the average fidelity of operations in theory and on quantum hardware [236, 265, 270–272], and therefore $\langle F_{LE}^{rand} \rangle$ will be sensitive to coherent errors that may not be captured by F_{LE} . Our results will show that optimizing a qubit assignment with respect to $\langle F_{LE}^{rand} \rangle$ will serve as a weak proxy for determining the optimal qubit assignment for a circuit U , unless U is itself a random circuit.

7.3.3 Simulated annealing for qubit selection

Simulated Annealing (SA) [28] is an optimization technique motivated by the behavior of physical systems to approach their ground state when cooled sufficiently slowly. We now describe the task of SA on finite sets [273] in the context of qubit assignment. Given a fixed logical circuit $\mathcal{C}_L$ to execute and the set $S = \{\mathcal{C}_p : M(\mathcal{C}_L) = \mathcal{C}_p\}$ of possible hardware circuits to evaluate, accompanied by a cost function $C : S \rightarrow \mathbb{R}$ to be minimized, our goal is to find the optimal subset $S^* \subseteq S$ corresponding to the minimum value of C . For hardware experiments we make a number of assumptions about M that guarantee that S will be finite. Namely, we restrict the sequence of gates $\mathcal{C}_p$ to be identical to the sequence of gates in $\mathcal{C}_L$ and fix the logical connectivity to be the same as the physical device connectivity, so that M can be understood as a one-to-one map between logical and physical qubits (see Appendix C.3). For a physical circuit $\mathcal{C}_p$ implementing a corresponding noisy process $\mathcal{E}_U$, we use the Loschmidt Echo performance metric of Sec. 7.3.2 to define a cost function as

$$C(\mathcal{C}_p) = 1 - F_{LE}(\mathcal{C}_p) \quad (7.10)$$

where F_{LE} is implicitly related to the set of qubits and entangling edges output by M . By analogy with annealing in physical systems, we define a temperature $T(t)$ at each iteration t of the algorithm that dictates how likely the annealer is to evaluate a worse state in the interest of exploration. Then, given a choice between a current physical circuit $\mathcal{C}_p$ and a new circuit $\mathcal{C}'_p$ at each step, the algorithm will probabilistically select $\mathcal{C}'_p$ depending on $C(\mathcal{C}'_p)$ and T : If $C(\mathcal{C}'_p) < C(\mathcal{C}_p)$ then the algorithm always chooses $\mathcal{C}'_p$. Otherwise if $C(\mathcal{C}'_p) > C(\mathcal{C}_p)$, SA will still allow for exploration of the state space with probability

$$P(\mathcal{C}_p \rightarrow \mathcal{C}'_p) = \exp\left(\frac{C(\mathcal{C}'_p) - C(\mathcal{C}_p)}{T(t)}\right) \quad (7.11)$$

and will therefore be more likely to explore disadvantageous configurations in the early stages of exploration to avoid becoming trapped in a local minimum. The implementation of SA for qubit assignment depends strongly on the set of new circuits that the algorithm can sample from at each step. Given a physical circuit $\mathcal{C}_p$, at any given iteration we define the neighborhood of circuits $S(\mathcal{C}_p)$ at S as the set of all circuits $\mathcal{C}'_p \neq \mathcal{C}_p$ that the algorithm may sample from. The definition of a neighborhood is important for understanding and tuning the convergence properties of finite SA. If the set of neighbors is too large, such as the extreme example $S_{rand}(\mathcal{C}_p) = S \setminus \{\mathcal{C}_p\}$ (i.e., S_{rand} contains all circuits in S except $\mathcal{C}_p$) then SA performs a random walk over S , and convergence to S^* requires the maximum number of iterations over all possible choices of $S(\mathcal{C}_p)$ [274]. Conversely if the neighborhood is too small, then SA may require a prohibitive runtime to reach S^* . For each experiment we will implicitly or explicitly define $S(\mathcal{C}_p)$ as it relates to the logical circuit connectivity. In Appendix C.3.1 we provide Algorithm 2 with pseudocode to implement SA for qubit assignment.

7.3.4 Logical and physical connectivity

For hardware experiments on a device containing a grid of qubits with nearest neighbor connectivity, we consider logical circuits with nearest neighbor line connectivity, thereby restricting S and $S(\mathcal{C}_p)$. Each circuit $\mathcal{C}_L = (\mathbf{g}, \mathbf{v}_L, \mathbf{e}_L)$ is defined over n logical qubits with edges between nearest neighbors:

$$\mathbf{v}_L = [v_1, \dots, v_n] \quad (7.12)$$

$$\mathbf{e}_L = [\{v_1, v_2\}, \dots, \{v_{n-1}, v_n\}] \quad (7.13)$$

where $v_k \in V_p$ for $k = 1 \dots n$ and $\{v_\ell, v_{\ell+1}\} \in E_p$ for $\ell = 1 \dots n - 1$ subject to the restriction that $\{v_1, \dots, v_n\}$ compose a length- n simple path on G_p . This restriction on $\mathcal{C}_L$ to satisfy the hardware connectivity graph (V_p, E_p) simplifies the implementation of M without sacrificing access to qubit assignments with higher fidelity: for any fixed $\mathbf{v}_L$ only two of the $n!$ orderings of $\{v_1, \dots, v_n\}$ will respect nearest neighbor connectivity on hardware, while the remainder will require additional overhead to implement SWAP networks that will reduce the fidelity of the program and are therefore ignored here. Given the correspondence between $\mathcal{C}_p$ and a simple path described by an array $\mathbf{v} \in V_p^n$, we explicitly define the neighborhood around $\mathcal{C}_p$ as

$$S(\mathcal{C}_p) = \{\mathbf{v}' : 0 < |\{v'_i\}_{i=1}^n - \{v_i\}_{i=1}^n| \leq k\} \cup \{R(\mathbf{v})\} \quad (7.14)$$

where R is the reversal operation sending $[x_1, \dots, x_n] \rightarrow [x_n, \dots, x_1]$, $|A - B|$ denotes a set difference over elements of A and B , and $k \in \mathbb{Z}$ is a tunable parameter that enforces

that two neighboring assignments differ by no more than k qubits and edges. By construction every state in the state space of length- n simple paths over G_p may be reached by performing updates to $\mathcal{C}_p$ according to $S(\mathcal{C}_p)$.

In simulations we also considered the more general case of unconstrained logical circuit connectivity thereby extending the state space to include qubit routing or *transpilation*, the task of introducing SWAP networks into a circuit so that circuit entangling operations respect the hardware connectivity. An n qubit circuit without connectivity constraints mapped onto N hardware qubits will result in a state space of size at least $N!/(N - n)!$, greatly complicating the optimization procedure compared to the cases of constrained circuits. The outcome of transpilation varies significantly depending on the initial qubit assignment [275] and may involve choosing (deterministically or stochastically) between SWAP networks of equal gate count without any consideration for the device noise. This hinders the ability of an SA algorithm to iteratively explore the full state space of qubit assignments using an “out-of-the-box” transpiler (see Appendix C.3) and emphasizes the need for noise-aware transpilation. In Appendix C.3.1 we provide Algorithm 3 that implicitly defines the neighborhood $S(\mathcal{C}_p)$ around $\mathcal{C}_p$ in the case of unconstrained logical connectivity.

7.4 Experiments

7.4.1 Simulations

To test our techniques in a controlled setting we constructed a connectivity graph $G_p = (V, E)$ encoding information about the (simulated) fidelities of single and two-qubit gates executed at each vertex and edge in G_p .² We used the same noise model described in Proposition 22 consisting of a series of symmetric depolarizing channels $\mathcal{D}_d$ acting on d -dimensional density matrices according to

$$\mathcal{D}_d(p) : \rho \rightarrow p \frac{I_d}{d} + (1 - p)\rho \quad (7.15)$$

which can be simulated on any subsystem of a state ρ using appropriately defined local Kraus operators. After assigning the weight ϵ_i to each qubit $i \in V$ and η_{ij} to each $\{i, j\} \in E$, we construct a noise model by inserting a depolarizing channel $\mathcal{D}_2(\epsilon_i)$ after each single-qubit operation on i and $\mathcal{D}_4(\eta_{ij})$ after each two-qubit operation acting on $\{i, j\}$. Circuit simulations were performed using the QISKIT [258] and CIRQ [259] software libraries.

²All code and data to reproduce the analyses in this chapter are available online [276].

We first tested the performance of Loschmidt Echoes in simulation by preparing noise graphs with a predetermined optimal qubit assignment. Each noise graph was uniquely determined by its connectivity and the arrays ϵ and η . Fig. 7.2a shows the results of computing F_{LE} and $F_0(\epsilon, \eta)$ for every possible qubit assignment on a 5×5 simulated grid. We find good agreement between the single-error extrapolation $F \approx \frac{1}{2}(F_{LE}F_0^{-1} + F_0)$ derived in Appendix C.1 to approximate F in the low-error regime, validating our analysis of the Loschmidt Echo.

We also considered the locality of the SA cost function, which determines how much the cost function can change at each iteration of annealing. Given a current qubit assignment (V, E) , a neighborhood size of k for the line connectivity state update rule enforces that any new assignment (V', E') satisfies $|V' - V| \leq k$ (for example $k = 0$ implies $|E \cap E'| = 4$ and $|V \cap V'| = 5$ for all $(V', E') \subset G_p$). The ability of SA to navigate the space of qubit assignments non-randomly depends on the locality of the cost function defined in Eq. 7.10 which we choose to measure using the quantity

$$\langle |F_{LE} - F'_{LE}| \rangle_{G_p} = \mathbb{E}_{C_p \in S} \mathbb{E}_{C'_p \in S(C_p)} |F_{LE} - F'_{LE}| \quad (7.16)$$

Eq. 7.16 describes the typical change in cost moving between assignments for a given definition of neighborhood $S(C_p)$. Fig. 7.2b shows that $\langle |F_{LE} - F'_{LE}| \rangle_{G_p}$ consistently decreases with k , indicating that the cost function C_{LE} is local in the space of qubit assignments.

We then tested the performance of SA for qubit assignment using Loschmidt Echoes. Fig. 7.3 demonstrates example performance of SA qubit assignment using F_{LE} as a cost function for a simulated noisy device where the logical and physical connectivity and gate-sets were equivalent, $C_L = C_p$ for all assignments. Our technique reliably finds high-performing and optimal assignments in noisy simulations. SA consistently outperforms a comparable baseline which keeps the highest performing assignment from a set of n_s random assignments, where n_s is the number of unique assignments attempted by the simulated annealer. We also considered SA for all-to-all logical connectivity circuits which were then transpiled onto nearest neighbor connectivity as a step in the qubit mapping M . We found SA to consistently outperformed random qubit assignment, demonstrating that SA for qubit assignment is a viable tool for the assignment of unconstrained logical connectivity circuits (Appendix C.3.1). However, the optimality guarantees for SA using a transpiler are somewhat weaker due to the mismatch between the size of the set of logical circuits and the set of all possible transpiled circuits, emphasizing the need for the development of noise-aware transpilers that can be integrated into a search over qubit assignments.

7.4.2 Hardware Experiments

We performed experiments using the 23-qubit Rainbow and 52-qubit Weber devices accessed via Google Quantum Cloud Services. Both devices are based on 2D grid nearest-neighbor connectivity (see Fig. C.3 in Appendix C.4). Since readout error significantly biases the estimation of F_{LE} we implemented readout error correction for all experiments (see Appendix C.2).

We first validated the use of the Loschmidt Echo metric F_{LE} introduced in Sec. 7.3.2 as a tool for determining the relative performance of qubit assignments. To test the performance of F_{LE} as a proxy for maximizing the final state fidelity we estimated both F_{LE} and the fidelity F for a subset of possible assignments on the Rainbow device. We executed three kinds of circuits. The first type of circuit prepares a Greenberger–Horne–Zeilinger (GHZ) state $|\psi_{\text{GHZ}}\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})$ using $n = 8$ qubits. F was computed using $O(n)$ experiments per qubit assignment using DFE [277] (see Appendix C.4.2). The second circuit was a SWAP network (SWAPnet): We first prepared an arbitrary single qubit state $|\psi_{\text{SWAPnet}}\rangle = \alpha|0\rangle + \beta|1\rangle$ and then moved the state to a different position on the hardware using 8 sequential pairs of $\sqrt{\text{iSWAP}}$ gates. The fidelity of the SWAPnet can be evaluated in constant time by performing a single-qubit projective measurement onto $|\psi_{\text{SWAPnet}}\rangle\langle\psi_{\text{SWAPnet}}|$. The final circuit we considered was a Clifford conjugation circuit (or mirror circuit in Refs. [245, 246]), which has the form

$$U = H^{\otimes n} C^\dagger P C H^{\otimes n} \quad (7.17)$$

where $C \in \text{Cl}(2^n)$ is a Clifford circuit over n qubits, $P = \sigma_{p_1} \otimes \cdots \otimes \sigma_{p_n}$ is an n -local Pauli string with each $p_k \in \{I, X, Y, Z\}$, and we used $n = 8$ for this experiment. While computing F for a Clifford circuit using DFE can incur significant overhead in the general case, here the effect of U simplifies to a Pauli operator acting on a computational basis state. Therefore the output of this circuit is uniquely described by a computational basis bitstring that can be efficiently simulated according to the Gottesman-Knill theorem [247] such that F can be determined in constant time.

We estimated each of the metrics F_{LE} , $\langle F_{\text{LE}}^{\text{rand}} \rangle$, and F_0 on a subset of qubit assignments satisfying linear connectivity on the Rainbow device, for each choice of circuit described above. Fig. 7.4 demonstrates results for diagnostic runs for the GHZ state (corresponding figures for the other circuits can be found in Appendix C.4). Each fidelity F was computed using DFE using $t = 1.5 \times 10^4$ repetitions for each choice of circuit. The corresponding values for F_{LE} for each qubit assignment were computed with variance $\text{Var}(F_{\text{LE}}) = F_{\text{LE}}(1 - F_{\text{LE}})t^{-1}$ resulting in negligible statistical error. Values for $\langle F_{\text{LE}}^{\text{rand}} \rangle$ were computed using an average over different random circuits, with F_{LE} being estimated using t repetitions for

each. $F_0(\hat{\epsilon}, \hat{\eta})$ was computed for each circuit assignment using gate counts n_1, n_2 for the corresponding circuit and taking $\hat{\epsilon}$ to be the average *RB* error per gate and $\hat{\eta}$ to be the average $\sqrt{\text{iSWAP}}$ XEB error per cycle [8, 278] (all circuits were implemented using the hardware-native $\sqrt{\text{iSWAP}}$ entangling gate). All F_0 values were calculated using RB and XEB calibration data estimated in parallel for separate qubits, and we have verified that calculating F_0 using non-parallel calibration data has no significant effect on our analysis.

The goal of using F_{LE} is to determine the qubit assignment corresponding to the highest value of F by proxy, which will only be possible if there is a strong monotonic relationship between F_{LE} and F . We chose to compute Kendall's τ_b coefficient [279] which captures the concordance between two random variables $X \sim p_X$ and $Y \sim p_Y$ according to

$$\tau_b(X, Y) = \mathbb{E}_{X, Y} \left[\text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j) \right] \quad (7.18)$$

evaluated over all pairs (i, j) with $i, j = 1 \dots N$. This coefficient has a simple interpretation in terms of joint behavior of changing X and Y . A value $\tau_b = k$ indicates that when the variable X increases (decreases), then the event that Y also increases (decreases) is a factor of k more likely than the event that Y decreases (increases). $\tau_b = 0$ indicates no monotonic relationship between variables. Table 7.1 computes τ_b with respect to the fidelity F for each diagnostic on each circuit implemented. In addition, we consider a quantity describing conditional performance of a metric as

$$P(Y_k | X_k) \equiv \frac{P(X > X_k, Y > Y_k)}{P(X > X_k)} \quad (7.19)$$

which provides the probability of a signal Y exceeding the k^{th} percentile Y_k of p_Y when an input X is drawn from p_X subject to $X > X_k$, where X_k is the k^{th} percentile of p_X . Practically, this quantity predicts the likelihood of a qubit assignment having high fidelity relative to other assignments when the assignment is chosen using only knowledge of the distribution of a diagnostic $X \in \{F_{\text{LE}}, \langle F_{\text{LE}}^{\text{rand}} \rangle, F_0\}$. Both τ_b and $P(Y_k | X_k)$ are consistently the highest between F_{LE} and F compared to other metrics considered, indicating that the Loschmidt Echo is the most reliable proxy when direct measurement of F is unavailable.

We now present the results of SA for qubit selection on a hardware device. For demonstration purposes we consider the case of *offline* qubit assignment, in which the value of F_{LE} was determined for every possible assignment and then cached in a lookup table to allow for many SA experiments to be repeated. For $n \gtrsim 4$, the number of qubit assignments grows super-exponentially and it becomes infeasible to construct a comprehensive

Circuit	GHZ	Clifford	SWAPnet
n	8	8	9
m	275	350	250
depth	17	25	17
Diagnostic X	$\tau_b(X, F) [10^{-2}]$		
F_{LE}	78.2 (1.4)	57.0 (2.9)	63.5 (2.4)
$\langle F_{LE}^{rand} \rangle$	29.4 (3.8)	52.7 (2.7)	31.8 (4.0)
F_0	9.4 (4.4)	10.4 (4.1)	-10.1 (4.5)
Diagnostic X	$P(F_{85} X_{85}) [10^{-2}]$		
F_{LE}	80.5 (5.7)	56.4 (6.9)	60.0 (9.0)
$\langle F_{LE}^{rand} \rangle$	27.8 (7.2)	53.8 (7.2)	14.3 (7.5)
F_0	22.2 (6.8)	25.6 (6.8)	8.6 (6.2)

Table 7.1: Monotonic relationship coefficients τ_b and probabilities $P(F_k|X_k)$ for the $k = 85^{th}$ percentile computed for each experiment and each diagnostic $X \in \{F_{LE}, \langle F_{LE}^{rand} \rangle, F_0\}$ evaluated on Rainbow. Standard deviations computed via bootstrapping [280] are provided in parentheses. Each experiment was run for $t = 1.5 \times 10^4$ repetitions on each of m distinct, randomly drawn assignments consisting of n qubits composing a simple path sampled from the Rainbow device using networkx [281]. For $n = 8$ and $n = 9$ there are 2984 and 4972 such paths respectively (including reversed paths), highlighting the need for optimization heuristics. In each case, F_{LE} captures a stronger monotonic relationship to F than the other metrics considered, though $\langle F_{LE}^{rand} \rangle$ is an effective method for assessing the performance of other random circuits.

lookup table, in which case SA would need to be performed *online* to optimize the qubit assignment.

To test the performance of Simulated Annealing for qubit assignment on hardware, we prepared a 4-qubit GHZ state restricted to logical line connectivity for each of the 1116 possible length-4 simple paths on the 53-qubit Weber grid layout. We evaluated both orderings of qubit assignments over each simple path, with the expectation that each ordering may produce a different fidelity. Roughly 27% of assignments were rejected on account of exceed our maximum threshold for 0.15 qubit visibility (see Appendix C.2), with SA being performed over the remaining dataset of 808 simple paths. Fig. 7.5 shows the results of SA for qubit selection using F_{LE} . SA optimization runs are divided according

to the number of unique states n_s for which F_{LE} would need to be computed on hardware during an online run. SA frequently found the qubit assignment corresponding to the optimal F_{LE}^* using n_s as small as 30, fewer than 5% of the assignments available in the dataset. Since prior knowledge of F_0 computed from diagnostic data could not be relied on to find a high scoring qubit assignment, we chose best-of- n_s random sampling as a competitive baseline optimization technique on the dataset. We found that SA consistently outperformed this scheme for all choices of n_s .

We also performed SA for qubit assignment on the 23 qubit Rainbow device for an $n = 3$ Quantum Fourier Transform (QFT) circuit. Each circuit was executed with a specific input basis state $|j\rangle$, $j \in [0, 2^n - 1]$ so that DFE could be performed using $O(1)$ experiments. Due to the much smaller state space (only 148 possible length-3 simple paths) we found that random assignment was competitive with SA (see Appendix C.4.1). This highlights the importance of considering circuit size and device layout when considering the qubit assignment problem.

7.5 Discussion

Even with the availability of detailed error rates for a quantum device, the task of assigning a logical circuit to a subset of hardware qubits that maximize the fidelity of the prepared state can be extremely difficult. In this work we have demonstrated instances of circuits for which the Loschmidt Echo provides an effective stand-in for the fidelity function in qubit assignment, and have provided theoretical justification for this relationship in the small-error limit. Conversely, we have shown that even excellent knowledge of device error rates (e.g. computed using Randomized Benchmarking) can be insufficient to predict the performance of a circuit mapped onto hardware. Furthermore, we provided analytical arguments that the advantage of F_{LE} over F_0 for determining relative performance would persist even if F_0 could be estimated to higher precision using scalable techniques (e.g. Refs. [282, 283]). We have also showed that Simulated Annealing can be used to find performant qubit assignments using significantly fewer resources than a complete exploration of possible qubit assignments would require.

Our theoretical arguments have relied on a low-error limit which limits this analysis to circuits with either few qubits or low gate depth. However our experimental results indicate that a strong connection between the Loschmidt Echo and the state fidelity function may still hold in a regime of higher errors. Future work may determine that such a relationship exists for a broader class of noise models or outside of the perturbative regime.

In this work we have evaluated the use of a Loschmidt Echo diagnostic on a specific selection of circuits employing only readout error mitigation, and so the optimal qubit assignments determined in this work may not remain optimal when employing platform-specific error mitigation. In particular, control errors which may go undetected by (see Appendix C.1.2.2) can be mitigated using Floquet Calibration [284], thereby strengthening the relationship between F_{LE} and F . Therefore such device-specific calibration procedures should be combined with qubit assignment based on F_{LE} to maximize the final assignment performance. Other error mitigation strategies may be found, for example in [285, 286].

However, as the number of qubit assignments grows exponentially with respect to the number of available qubits on the device and the number of qubits in the circuit, it becomes essential to minimize the experimental overhead associated with device calibration and computation of F_{LE} for each assignment. Future work will be necessary to determine the appropriate balance of implementing error mitigation so that F_{LE} remains faithful to F while minimizing overhead such that a significant number of qubit assignments can be probed during the optimization procedure.

While we have focused on maximizing the fidelity of a state prepared by a specific unitary U , our method can be readily adapted to analyze more general situations. For instance, a Hilbert-Schmidt test [228] which computes $\text{Tr}(U^\dagger V)$ for unitaries U, V could be used to compute $\text{Tr}(\mathcal{E}_{U^\dagger} \circ \mathcal{E}_U(|0\rangle\langle 0|))$ as a proxy for the process fidelity [287, 288], $F(U, \mathcal{E}_U) = \int \mathcal{F}(U\psi U^\dagger, \mathcal{E}_U(\psi)) d\psi$. However such a modification will introduce significant two-qubit gate overhead, for which our weak-error analysis of the Loschmidt echo technique is less applicable.

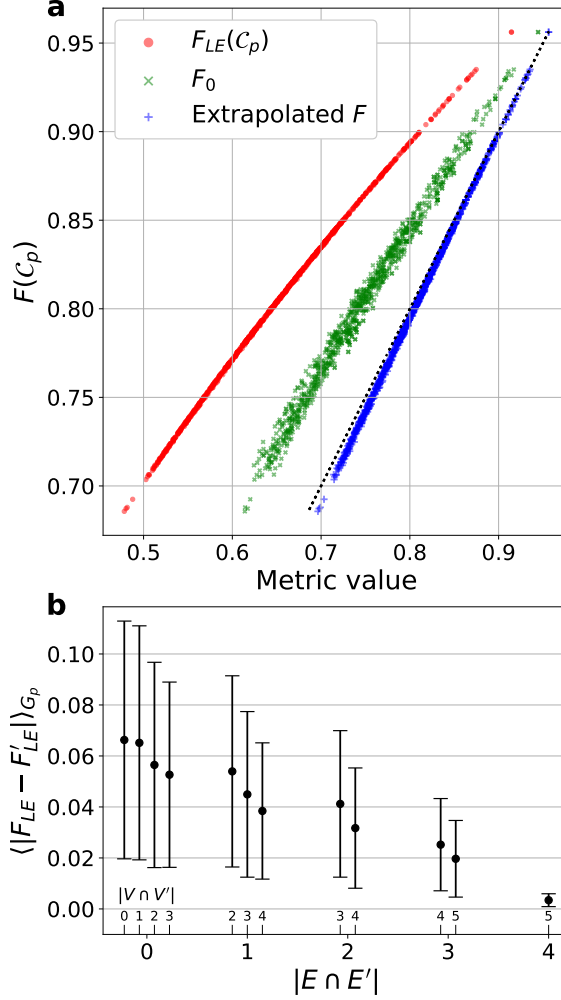


Figure 7.2: **a.** Simulated 5-qubit GHZ circuit exposed to a 2-local depolarizing noise model shows a strong monotonic relationship between F_{LE} and F compared to a weaker relationship between F_0 and F . Also shown is the extrapolation $F \approx \frac{1}{2} (F_{LE}F_0^{-1} + F_0)$ (Eq. C.33) which closely matches F in the low error limit $F \rightarrow 1$ well outside of the single-error limit (compare to dashed line $y = x$, indicating perfect prediction of $F(C_p)$). **b.** Computing the difference in the cost function for every pair of qubit assignments in G_p as a function of the number of shared qubits and edges demonstrates that the cost function is local in S . Note that even when all edges and qubits are shared ($k = 0$) any linear path and its reversal may differ in the cost function.

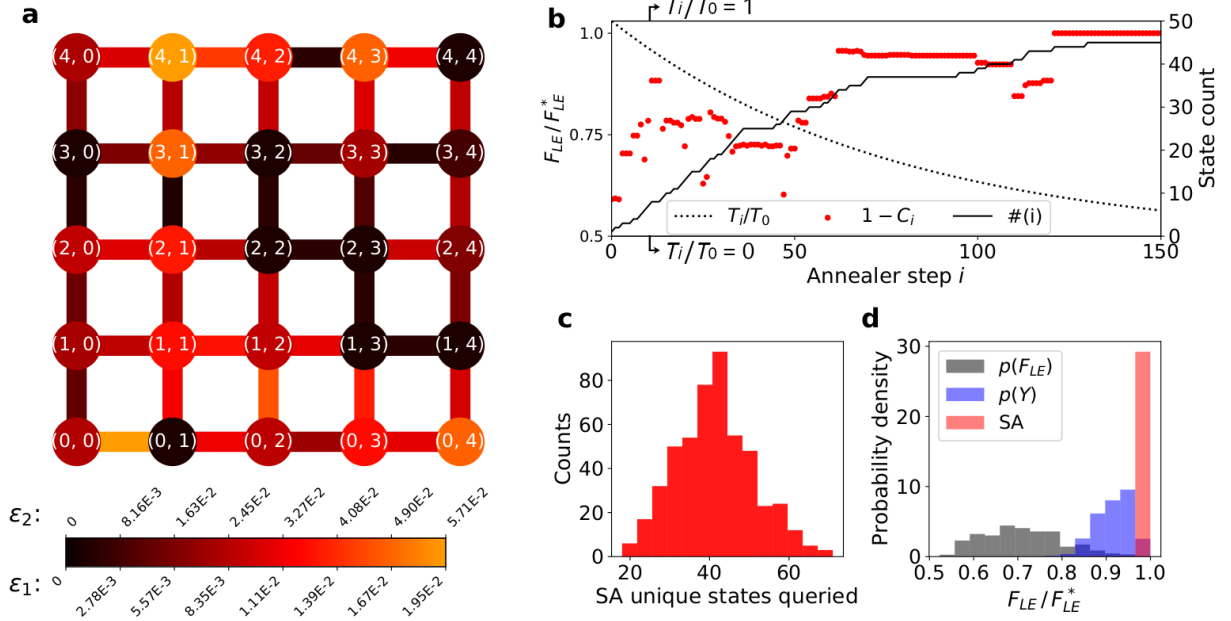


Figure 7.3: **a.** Sample simulated noise graph engineered to contain an optimal 5-qubit assignment used to validate SA qubit assignment in simulation. Each vertex and edge is color coded by the weight corresponding to the depolarization probability for channels acting after each one- and two-qubit gate respectively **b.** Sample score history for a Simulated Annealer optimization run using simulated noise model on the engineered noisemap, with exponential temperature decay $T_i = T_0 \alpha^i$ ($\alpha = 0.987$, $T_0 = 0.10$) and neighborhood size $k_{max} = 2$ for 150 steps. The score $1 - C_i$ and number of unique states $\#(i)$ queried initially change quickly at high temperatures, before stabilizing as T approaches zero (S^* is often found and maintained early). The dotted curve for T_i/T_0 is plotted on a linear, 0 (bottom of the y-axis) to 1 (top of the y-axis) scale. **c.** The distribution $p(n)$ describing the number of unique assignments sampled by 500 trials of SA is generally smaller than the total number of iterations. **d.** We compare SA to assignment based on random sampling: Given the set of scores S , for 10,000 trials of $n \sim p(n)$ we sample a subset of scores $K \subset S$ with $|K| = n$ and compute $Y = \max(K)$. The resulting distribution $p(Y)$ is consistently outperformed by SA. Also shown is $P(F_{LE})$, the distribution of scores over all qubit assignments.

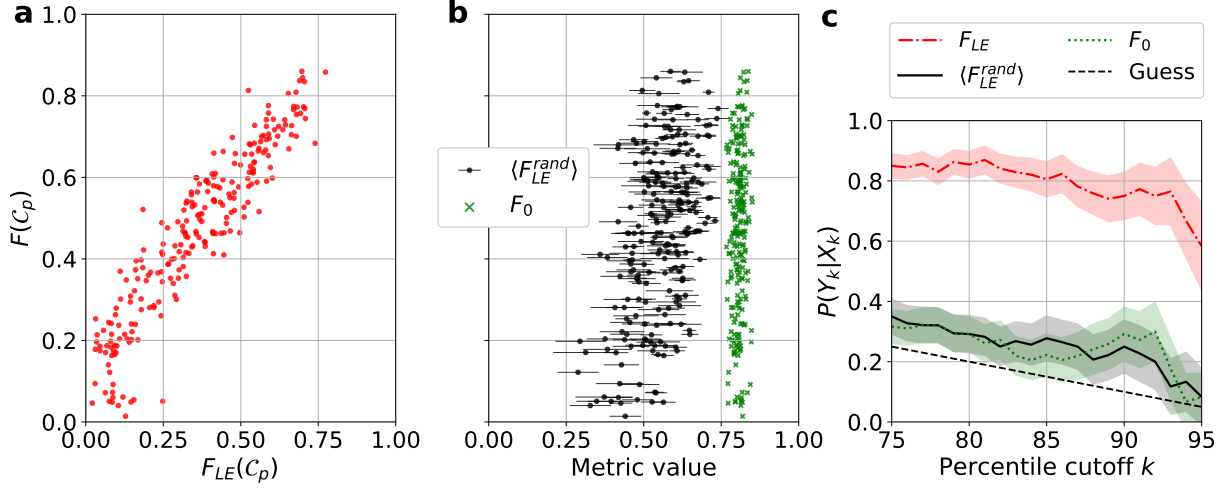


Figure 7.4: **a.** Experimental results comparing F to F_{LE} for an 8-qubit GHZ state demonstrate strong monotonic relationship ($\tau_b = 0.782$) between the two diagnostics. **b.** Experiments comparing the average Loschmidt survival over 5 random circuits $\langle F_{LE}^{rand} \rangle$ ($\tau_b = 0.294$) and the metric $F_0 := F_0(\hat{\epsilon}, \hat{\eta})$ ($\tau_b = 0.094$) computed from calibration data $(\hat{\epsilon}, \hat{\eta})$ queried from the quantum processor's most recent calibration (typically 4-8 hours prior to an experiment) show only a weak relationship between either metric and the fidelity F . Error bars for $\langle F_{LE}^{rand} \rangle$ indicate the sample standard deviation computed over 5 random circuits. **c.** The diagnostic results can be interpreted by computing the conditional probability of the fidelity F in the k^{th} percentile of fidelities given that an observed metric $X \in \{F_{LE}, \langle F_{LE}^{rand} \rangle, F_0\}$ falls in the k^{th} percentile X_k of its own distribution. Shading indicates standard deviation obtained via bootstrapping. For reference, the dashed line describes uniformly random sampling of F which gives $P(Y_k|X_k) = 1 - k/100$.

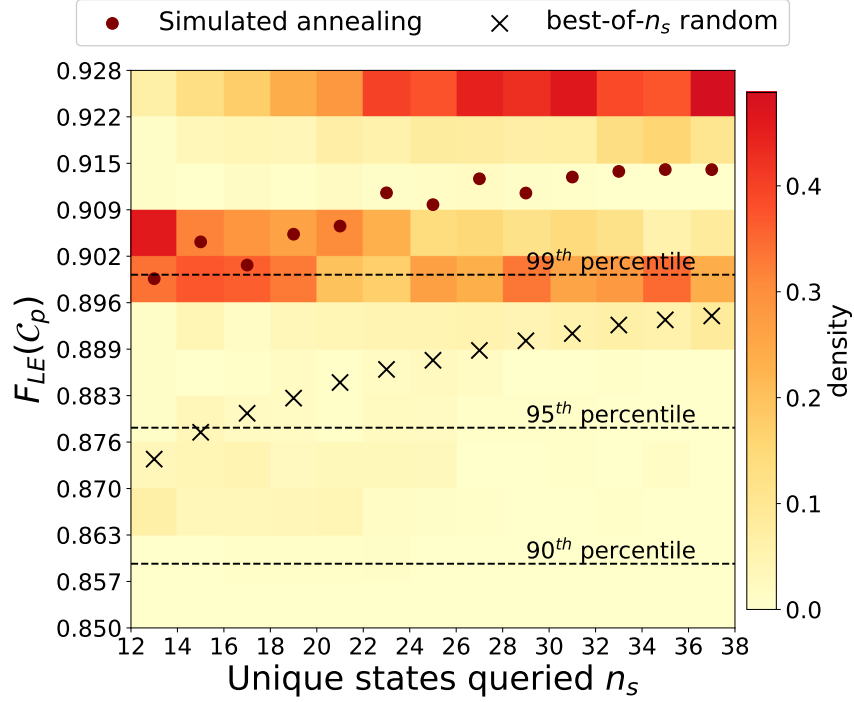


Figure 7.5: 2D histogram of best F_{LE} values found over 2500 trials of hardware SA using F_{LE} values computed for $n = 4$ GHZ over the complete state space of circuit assignments on Weber. All annealing trials used exponential temperature decay $T_i = T_0 \alpha^i$ ($\alpha = 0.988$, $T_0 = 0.07$), chosen to balance exploration of the state space with suppressing n_s . Columns are renormalized since the typical number of unique states queried n_s concentrates around 23 for this choice of SA hyperparameters. SA optimization frequently finds the best possible value F_{LE} from the state space (contained in the maximum y -bin shown) but the n_s necessary to do so varies. Given the poor predictive capabilities of device diagnostic data for GHZ state fidelity, we compare this method to a random sampling scheme that consists of drawing n_s circuit assignments without replacement (10000 trials per n_s) and keeping the assignment corresponding to the maximum F_{LE} observed. SA consistently outperforms random sampling for the $n = 4$ GHZ dataset. Averaged over all choices of n_s the final F_{LE} achieved by SA outperformed the random sampling scheme by 2.8%.

Chapter 8

Conclusion

In this thesis, we have studied several topics at the intersection of machine learning and quantum information, as well as quantum error mitigation. We began by describing spectra associated with classical kernel methods, which provide an elegant mathematical description for how learning algorithms learn from data. We were able to apply spectra in Chapter 2 to demonstrate how the kernel bandwidth technique – which limits the effective dimension of a feature space used by a quantum kernel method – affects generalization of certain learning algorithms. While an analogous technique is ubiquitous in classical machine learning (e.g. the bandwidth parameter of an RBF kernel), we cannot validate techniques in QML relying on empirical evidence alone. Currently, applications of such techniques are usually limited to noisy devices or circuits that are small enough to be classically simulable (or both), and so spectral methods provide us with the rare opportunity to justify or refute such techniques theoretically. However, the kernel bandwidth technique introduces potential complications to quantum kernel methods. Namely, by reducing the effective dimension spanned by the set of feature vectors $\{\phi(x)\}_{x \in \mathcal{X}}$, it is possible that we have removed the property that made the kernel hard to evaluate in the first place [62].

While the study of benign overfitting in classical machine learning models was originally motivated by the observation that deep neural networks with millions of parameters often both overfit data and generalize well [66–70], we probably cannot expect a similar situation for QML algorithms in the near term. Such computations will probably be small enough or noisy enough to be classically simulable, and therefore irrelevant for practical machine learning. And in the absence of fault-tolerant devices, we cannot rely on empirical studies to understand whether benign overfitting occurs in QML algorithms applied to real datasets. Thus, the work presented in Chapter 3 to study benign overfitting in quantum machine learning models represents a first step towards a theoretical characterization of

this phenomenon. Looking forward, the relevance of this line of study will ultimately rely on the observation of (or lack thereof) benign overfitting in a realistic application of QML, which itself depends on finding an interesting learning problem that is well-suited for QML.

Following the original framework of Ref. [19], we have begun to theoretically characterize the role of correlated auxiliary noise in communication and learning tasks. While Chapter 4 showed information-theoretic demonstrations given access to correlated auxiliary noise, these analyses fundamentally rely on discrete data. Moving beyond this setting into the setting of continuous-valued data found more commonly in learning problems will pose some difficulty. For instance, it is straightforward to repurpose the arguments of [155] and Sec. 5.3.1 to construct an information-theoretic error lower bound for the learning of sparse vectors given access to correlated auxiliary noise. However, this will only result in a vacuous error lower bound and will not actually demonstrate an advantage for correlated noise. On the other hand, an error upper bound of the form we have derived in Sec. 5.3.2 will require more sophisticated techniques to account for the convexity of the relative entropy. In the quantum setting, future work will consider a broader class of channels for which correlated auxiliary noise provides a provable advantage for communication and hypothesis testing tasks.

In Chapter 5, we introduced and characterized a quantum information processing task that blurs the line between learning from data and manipulating entanglement. The overlap of these topics arises naturally in a discrete picture: classical hypothesis testing is successful when a learner’s random variable and the unknown random variable are strongly *classically* correlated, and the singlet fraction of a quantum state is large when the learner’s subsystem and the reference system are strongly *quantumly* correlated. However, formalizing this connection in a continuous variable setting becomes more difficult, since many quantities in information theory (e.g. differential entropy [115]) lose their operational meaning when extended to continuous support in the obvious way [289]. Along these lines, this work has also stopped short of exploring continuous-variable extensions of quantum PAC learning, which we recall involves query access to observations in the form $\sum_{x \in \mathcal{X}} \sqrt{p(x)} |x, f_i(x)\rangle$ [21]. As with exact learning, the naive extension of the definition of this state to continuous support – some object that looks like $\int_{\mathcal{X}} dx \sqrt{p(x)} |x, f_i(x)\rangle$ – does not necessarily result in a valid state. However, as with the task discussed in Chapter 5, future work may reveal a more natural generalization of quantum PAC learning by recruiting ideas from quantum probability and continuous variable quantum mechanics.

This thesis also presented error mitigation techniques for quantum devices in the absence of fault tolerance. In Chapters 6 and 7 we have introduced methods for approximately mitigating readout error and optimizing the assignment of logical qubits to physical qubits respectively. We have used perturbative techniques to derive certain performance guaran-

tees for these methods, and while the low error rate necessary for these guarantees to hold is usually not achievable by a near-term device in the first place, we were able to provide empirical evidence that these techniques are effective in a broader range of conditions. We ought to explore how to combine these techniques with quantum error correction or fault tolerance so that these techniques (along with a vast literature dedicated to NISQ computing) might remain relevant upon the arrival of fault-tolerant quantum computation.

References

- [1] Abdulkadir Canatar et al. “Bandwidth Enables Generalization in Quantum Kernel Models”. In: *Transactions on Machine Learning Research* (June 2023) (cit. on pp. [iv](#), [5](#), [17](#), [23](#), [49](#)).
- [2] Evan Peters and Maria Schuld. “Generalization despite overfitting in quantum machine learning models”. In: *Quantum* 7 (Dec. 2023), p. 1210. DOI: [10.22331/q-2023-12-20-1210](#) (cit. on pp. [iv](#), [27](#)).
- [3] Evan Peters. *Bounds and guarantees for learning and entanglement*. 2024. arXiv: [2404.07277 \[quant-ph\]](#) (cit. on pp. [iv](#), [76](#)).
- [4] Evan Peters, Andy C. Y. Li, and Gabriel N. Perdue. “Perturbative readout-error mitigation for near-term quantum computers”. In: *Phys. Rev. A* 107 (6 June 2023), p. 062426. DOI: [10.1103/PhysRevA.107.062426](#) (cit. on pp. [iv](#), [v](#), [109](#), [131](#), [218](#), [242](#)).
- [5] Evan Peters et al. “Qubit Assignment Using Time Reversal”. In: *PRX Quantum* 3 (4 Dec. 2022), p. 040333. DOI: [10.1103/PRXQuantum.3.040333](#) (cit. on pp. [iv](#), [v](#), [125](#), [130](#), [230](#)).
- [6] Evan Peters et al. “Machine learning of high dimensional data on a noisy quantum processor”. In: *npj Quantum Information* 7.1 (2021), pp. 1–5 (cit. on pp. [v](#), [17](#), [21](#), [109](#), [131](#)).
- [7] Erik Gustafson et al. “Large scale multi-node simulations of $\mathbb{Z}_2$ gauge theory quantum circuits using Google Cloud Platform”. In: *2021 IEEE/ACM Second International Workshop on Quantum Computing Software (QCS)*. 2021, pp. 72–79 (cit. on p. [v](#)).
- [8] Sergei V. Isakov et al. *Simulations of Quantum Circuits with Approximate Noise using qsim and Cirq*. arXiv:211.02396. 2021. arXiv: [2111.02396 \[quant-ph\]](#) (cit. on pp. [v](#), [137](#)).

- [9] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002. DOI: [10.7551/mitpress/4175.001.0001](https://doi.org/10.7551/mitpress/4175.001.0001) (cit. on pp. 2, 9).
- [10] Nachman Aronszajn. “Theory of reproducing kernels”. In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404 (cit. on pp. 2, 8).
- [11] Vojtěch Havlíček et al. “Supervised learning with quantum-enhanced feature spaces”. In: *Nature* 567.7747 (Mar. 2019), pp. 209–212. DOI: <https://doi.org/10.1038/s41586-019-0980-2> (cit. on pp. 2, 9, 28, 109, 111).
- [12] Maria Schuld and Nathan Killoran. “Quantum Machine Learning in Feature Hilbert Spaces”. In: *Phys. Rev. Lett.* 122 (4 Feb. 2019), p. 040504. DOI: [10.1103/PhysRevLett.122.040504](https://doi.org/10.1103/PhysRevLett.122.040504) (cit. on pp. 2, 9, 28, 111).
- [13] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. “A rigorous and robust quantum speed-up in supervised machine learning”. In: *Nature Physics* 17.9 (July 2021), pp. 1013–1017. DOI: [10.1038/s41567-021-01287-z](https://doi.org/10.1038/s41567-021-01287-z) (cit. on pp. 2, 9, 88).
- [14] P.W. Shor. “Algorithms for quantum computation: discrete logarithms and factoring”. In: *Proceedings 35th Annual Symposium on Foundations of Computer Science*. 1994, pp. 124–134. DOI: [10.1109/SFCS.1994.365700](https://doi.org/10.1109/SFCS.1994.365700) (cit. on p. 2).
- [15] Jonas Jäger and Roman V Krems. “Universal expressiveness of variational quantum classifiers and quantum kernels for support vector machines”. In: *Nature Communications* 14.1 (2023), p. 576 (cit. on pp. 2, 9).
- [16] Scott Aaronson and Andris Ambainis. “Forrelation: A Problem That Optimally Separates Quantum from Classical Computing”. In: *SIAM Journal on Computing* 47.3 (2018), pp. 982–1038. DOI: [10.1137/15M1050902](https://doi.org/10.1137/15M1050902) (cit. on p. 2).
- [17] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. “On learning with integral operators”. In: *Journal of Machine Learning Research* 11.30 (2010), pp. 905–934 (cit. on pp. 2, 18, 25).
- [18] Jonas M. Kübler, Simon Buchholz, and Bernhard Schölkopf. “The Inductive Bias of Quantum Kernels”. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (2021), pp. 12661–12673. arXiv: [2106.03747](https://arxiv.org/abs/2106.03747) [quant-ph] (cit. on pp. 3, 19–24, 88).
- [19] Aida Ahmadzadegan et al. “Neural networks can learn to utilize correlated auxiliary noise”. In: *Scientific reports* 11.1 (2021), p. 21624 (cit. on pp. 3, 50, 51, 146).

- [20] Siddharth Karumanchi et al. “Quantum Channel Capacities with Passive Environment Assistance”. In: *IEEE Trans. Inform. Theory* 62.4 (2016), pp. 1733–1747. DOI: [10.1109/TIT.2016.2522192](https://doi.org/10.1109/TIT.2016.2522192) (cit. on pp. 3, 52, 66, 70, 72).
- [21] 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*. COLT 1995. New York, NY, USA: Association for Computing Machinery, 1995, pp. 118–127. DOI: [10.1145/225298.225312](https://doi.org/10.1145/225298.225312) (cit. on pp. 3, 88, 146).
- [22] Rocco A. Servedio and Steven J. Gortler. “Equivalences and Separations Between Quantum and Classical Learnability”. In: *SIAM Journal on Computing* 33.5 (2004), pp. 1067–1092. DOI: [10.1137/S0097539704412910](https://doi.org/10.1137/S0097539704412910) (cit. on pp. 3, 88).
- [23] Alp Atıcı and Rocco A. Servedio. “Quantum Algorithms for Learning and Testing Juntas”. In: *Quantum Information Processing* 6.5 (Sept. 2007), pp. 323–348. DOI: [10.1007/s11128-007-0061-6](https://doi.org/10.1007/s11128-007-0061-6) (cit. on pp. 3, 88).
- [24] Srinivasan Arunachalam and Ronald De Wolf. “Optimal quantum sample complexity of learning algorithms”. In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2879–2878 (cit. on pp. 3, 88).
- [25] John Preskill. “Quantum Computing in the NISQ era and beyond”. In: *Quantum* 2 (Aug. 2018), p. 79. DOI: [10.22331/q-2018-08-06-79](https://doi.org/10.22331/q-2018-08-06-79) (cit. on pp. 3, 125).
- [26] Michael R. Geller. “Rigorous measurement error correction”. In: *Quantum Sci. Technol.* 5.3 (June 2020), 03LT01. DOI: [10.1088/2058-9565/ab9591](https://doi.org/10.1088/2058-9565/ab9591) (cit. on pp. 4, 110).
- [27] J. Loschmidt. “Über den Zustand des Wärmegleichgewichts eines Systems von Körpern mit Rücksicht auf die Schwerkraft”. In: *Sitzungsberichte der Akademie der Wissenschaften, Wien* II.73 (1876), pp. 128–142 (cit. on pp. 4, 129).
- [28] Peter JM Van Laarhoven and Emile HL Aarts. “Simulated annealing”. In: *Simulated annealing: Theory and applications*. Springer, 1987 (cit. on pp. 4, 126, 132).
- [29] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2011. DOI: [10.1017/CB09780511976667](https://doi.org/10.1017/CB09780511976667) (cit. on p. 7).
- [30] Lorenzo Rosasco and Mikhail Belkin. “On Learning with Integral Operators”. In: *Journal of machine learning research (JMLR)* (2010), p. 32 (cit. on p. 8).

- [31] James Mercer. “Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations”. In: *Philosophical Transactions of the Royal Society A* 209 (1909), pp. 415–446 (cit. on p. 8).
- [32] George Kimeldorf and Grace Wahba. “Some results on Tchebycheffian spline functions”. In: *Journal of Mathematical Analysis and Applications* 33.1 (1971), pp. 82–95. DOI: [https://doi.org/10.1016/0022-247X\(71\)90184-3](https://doi.org/10.1016/0022-247X(71)90184-3) (cit. on p. 9).
- [33] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. “A generalized representer theorem”. In: *International conference on computational learning theory*. Springer, 2001, pp. 416–426 (cit. on p. 9).
- [34] Hsin-Yuan Huang et al. “Power of data in quantum machine learning”. In: *Nat. Commun.* 12.1 (2021), p. 2631 (cit. on pp. 9, 17, 20, 28, 40, 88).
- [35] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018. DOI: [10.1017/9781316848142](https://doi.org/10.1017/9781316848142) (cit. on pp. 10, 86).
- [36] Maria Schuld. “Supervised quantum machine learning models are kernel methods”. In: *arXiv:2101.11020* (2021). DOI: [10.48550/arXiv.2101.11020](https://doi.org/10.48550/arXiv.2101.11020) (cit. on pp. 10, 40).
- [37] Grace Wahba. *Spline models for observational data*. SIAM, 1990 (cit. on p. 11).
- [38] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. “Regularization networks and support vector machines”. In: *Advances in computational mathematics* 13 (2000), pp. 1–50 (cit. on p. 11).
- [39] J. Shawe-Taylor et al. “On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA”. In: *IEEE Transactions on Information Theory* 51.7 (2005), pp. 2510–2522. DOI: [10.1109/TIT.2005.850052](https://doi.org/10.1109/TIT.2005.850052) (cit. on p. 11).
- [40] Elies Gil-Fuster, Jens Eisert, and Vedran Dunjko. “On the expressivity of embedding quantum kernels”. In: *Machine Learning: Science and Technology* () (cit. on p. 12).
- [41] Sofiene Jerbi et al. “Quantum machine learning beyond kernel methods”. In: *Nat. Commun.* 14.1 (2023), p. 517. DOI: <https://doi.org/10.1038/s41467-023-36159-y> (cit. on pp. 12, 39).
- [42] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. “Effect of data encoding on the expressive power of variational quantum-machine-learning models”. In: *Phys. Rev. A* 103.3 (Mar. 2021), p. 032430. DOI: [10.1103/physreva.103.032430](https://doi.org/10.1103/physreva.103.032430) (cit. on pp. 13, 16, 29, 40, 205).
- [43] Salomon Bochner et al. *Lectures on Fourier integrals*. Vol. 42. Princeton University Press, 1959 (cit. on pp. 13, 16).

- [44] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. “On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels”. In: *Proceedings of the 33rd Conference on Learning Theory (COLT) 125* (2020), pp. 2683–2711 (cit. on pp. 13, 19, 31).
- [45] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. “Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks”. In: *Nature Communications* 12.1 (2021), pp. 1–12. DOI: [10.1038/s41467-021-23103-1](https://doi.org/10.1038/s41467-021-23103-1) (cit. on pp. 17, 19, 23).
- [46] Huaiyu Zhu et al. “Gaussian regression and optimal finite dimensional linear models”. In: (1997) (cit. on p. 17).
- [47] Persi Diaconis, Sharad Goel, and Susan Holmes. “Horseshoes in multidimensional scaling and local kernel methods”. In: *The Annals of Applied Statistics* 2.3 (Sept. 2008). DOI: [10.1214/08-aos165](https://doi.org/10.1214/08-aos165) (cit. on p. 17).
- [48] Ha Quang Minh. “Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory”. In: *Constructive Approximation* 32.2 (2010), pp. 307–338 (cit. on p. 17).
- [49] Jennifer R. Glick et al. “Covariant quantum kernels for data with group structure”. In: *arXiv:2105.03406* (2021). DOI: [10.48550/arXiv.2105.03406](https://doi.org/10.48550/arXiv.2105.03406) (cit. on p. 17).
- [50] Sau Lan Wu et al. “Application of quantum machine learning using the quantum kernel algorithm on high energy physics analysis at the LHC”. In: *Physical Review Research* 3.3 (Sept. 2021). DOI: [10.1103/physrevresearch.3.033221](https://doi.org/10.1103/physrevresearch.3.033221) (cit. on p. 17).
- [51] Thomas Hubregtsen et al. “Evaluation of Parameterized Quantum Circuits: on the relation between classification accuracy, expressibility and entangling capability”. In: *Quantum Mach. Intell.* 3 (Mar. 2021), p. 1. DOI: [10.1007/s42484-021-00038-w](https://doi.org/10.1007/s42484-021-00038-w) (cit. on pp. 17, 28).
- [52] Ali Rahimi and Benjamin Recht. “Random Features for Large-Scale Kernel Machines”. In: *Advances in Neural Information Processing Systems*. Vol. 20. 2007 (cit. on pp. 18, 32).
- [53] Benoît Collins and Ion Nechita. “Random matrix techniques in quantum information theory”. In: *Journal of Mathematical Physics* 57.1 (Jan. 2016), p. 015215. DOI: [10.1063/1.4936880](https://doi.org/10.1063/1.4936880) (cit. on p. 18).

- [54] Zbigniew Puchała and Jarosław Adam Miszczak. “Symbolic integration with respect to the Haar measure on the unitary group”. In: *Bulletin of the Polish Academy of Sciences: Technical Sciences* 65.1 (2017), pp. 21–27. DOI: [10.1515/bpasts-2017-0003](https://doi.org/10.1515/bpasts-2017-0003). eprint: [1109.4244](https://arxiv.org/abs/1109.4244) (cit. on pp. [19](#), [196](#), [197](#)).
- [55] Antonio Anna Mele. “Introduction to Haar Measure Tools in Quantum Information: A Beginner’s Tutorial”. In: *Quantum* 8 (May 2024), p. 1340. DOI: [10.22331/q-2024-05-08-1340](https://doi.org/10.22331/q-2024-05-08-1340) (cit. on p. [19](#)).
- [56] Rainer Dietrich, Manfred Oppen, and Haim Sompolinsky. “Statistical mechanics of support vector networks”. In: *Physical Review Letters* 82.14 (1999), p. 2975. DOI: [10.1103/PhysRevLett.82.2975](https://doi.org/10.1103/PhysRevLett.82.2975) (cit. on p. [19](#)).
- [57] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. *The Curse of Dimensionality for Local Kernel Machines*. Tech. rep. TR-1258. Université de Montréal, Mar. 2005 (cit. on p. [19](#)).
- [58] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. “Spectrum dependent learning curves in kernel regression and wide neural networks”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1024–1034 (cit. on p. [19](#)).
- [59] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014 (cit. on p. [21](#)).
- [60] Ruslan Shaydulin and Stefan M. Wild. “Importance of kernel bandwidth in quantum machine learning”. In: *Phys. Rev. A* 106 (4 Oct. 2022), p. 042407. DOI: [10.1103/PhysRevA.106.042407](https://doi.org/10.1103/PhysRevA.106.042407) (cit. on p. [21](#)).
- [61] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018 (cit. on p. [23](#)).
- [62] Lucas Slattery et al. “Numerical evidence against advantage with quantum fidelity kernels on classical data”. In: *Physical Review A* 107.6 (2023), p. 062417 (cit. on pp. [26](#), [145](#)).
- [63] Michael A Nielsen. *Neural networks and deep learning*. Determination Press, 2015 (cit. on p. [28](#)).
- [64] Stuart Geman, Elie Bienenstock, and René Doursat. “Neural Networks and the Bias/Variance Dilemma”. In: *Neural Comput.* 4.1 (1992), pp. 1–58. DOI: [10.1162/neco.1992.4.1.1](https://doi.org/10.1162/neco.1992.4.1.1) (cit. on p. [28](#)).
- [65] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009. DOI: [http://dx.doi.org/10.1007/978-0-387-84858-7](https://doi.org/http://dx.doi.org/10.1007/978-0-387-84858-7) (cit. on p. [28](#)).

- [66] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. “Deep learning: a statistical viewpoint”. In: *Acta Numerica* 30 (2021), pp. 87–201. DOI: [10.1017/S0962492921000027](https://doi.org/10.1017/S0962492921000027) (cit. on pp. 28, 145, 189).
- [67] Mikhail Belkin. “Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation”. In: *Acta Numerica* 30 (2021), pp. 203–248. DOI: [doi:10.1017/S0962492921000039](https://doi.org/10.1017/S0962492921000039) (cit. on pp. 28, 145).
- [68] Peter L. Bartlett et al. “Benign overfitting in linear regression”. In: *Proc. Natl. Acad. Sci.* 117.48 (2020), pp. 30063–30070. DOI: [10.1073/pnas.1907378117](https://doi.org/10.1073/pnas.1907378117) (cit. on pp. 28, 31, 37, 145, 185, 189, 211).
- [69] Mikhail Belkin et al. “Reconciling modern machine-learning practice and the classical bias-variance trade-off”. In: *Proc. Natl. Acad. Sci.* 116.32 (2019), pp. 15849–15854. DOI: [10.1073/pnas.1903070116](https://doi.org/10.1073/pnas.1903070116) (cit. on pp. 28, 145).
- [70] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. “Does data interpolation contradict statistical optimality?” In: *Proceedings of Machine Learning Research*. Vol. 89. PMLR, Apr. 2019, pp. 1611–1619 (cit. on pp. 28, 145).
- [71] Vidya Muthukumar et al. “Harmless interpolation of noisy data in regression”. In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 67–83. DOI: [10.1109/ISIT.2019.8849614](https://doi.org/10.1109/ISIT.2019.8849614) (cit. on pp. 28, 29, 31, 33, 34, 36, 185, 187–189, 193).
- [72] Vidya Muthukumar et al. “Classification vs regression in overparameterized regimes: Does the loss function matter?” In: *J. Mach. Learn. Res.* 22.222 (2021), pp. 1–69 (cit. on pp. 28, 29).
- [73] Yehuda Dar, Vidya Muthukumar, and Richard G. Baraniuk. *A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning*. 2021. DOI: <https://doi.org/10.48550/arXiv.2109.02355>. arXiv: [2109.02355](https://arxiv.org/abs/2109.02355) [stat.ML] (cit. on pp. 28, 29, 34).
- [74] Marcello Benedetti et al. “Parameterized quantum circuits as machine learning models”. In: *Quantum Sci. Technol.* 4.4 (Nov. 2019), p. 043001. DOI: [10.1088/2058-9565/ab4eb5](https://doi.org/10.1088/2058-9565/ab4eb5) (cit. on p. 28).
- [75] K. Mitarai et al. “Quantum circuit learning”. In: *Physical Review A* 98.3 (Sept. 2018). DOI: [10.1103/physreva.98.032309](https://doi.org/10.1103/physreva.98.032309) (cit. on pp. 28, 39, 111).
- [76] Maria Schuld et al. “Evaluating analytic gradients on quantum hardware”. In: *Phys. Rev. A* 99.3 (Mar. 2019), p. 032331. DOI: [10.1103/physreva.99.032331](https://doi.org/10.1103/physreva.99.032331) (cit. on p. 28).

- [77] Seth Lloyd and Christian Weedbrook. “Quantum Generative Adversarial Learning”. In: *Phys. Rev. Lett.* 121.4 (July 2018), p. 040502. DOI: [10.1103/physrevlett.121.040502](https://doi.org/10.1103/physrevlett.121.040502) (cit. on p. 28).
- [78] Pierre-Luc Dallaire-Demers and Nathan Killoran. “Quantum generative adversarial networks”. In: *Phys. Rev. A* 98.1 (July 2018), p. 012324. DOI: [10.1103/physreva.98.012324](https://doi.org/10.1103/physreva.98.012324) (cit. on p. 28).
- [79] Amira Abbas et al. “The power of quantum neural networks”. In: *Nat. Comput. Sci.* 1.6 (June 2021), pp. 403–409. DOI: [10.1038/s43588-021-00084-1](https://doi.org/10.1038/s43588-021-00084-1) (cit. on pp. 28, 49).
- [80] Logan G. Wright and Peter L. McMahon. “The Capacity of Quantum Neural Networks”. In: *2020 Conference on Lasers and Electro-Optics (CLEO)*. 2020, pp. 1–2 (cit. on p. 28).
- [81] Sukin Sim, Peter D. Johnson, and Alán Aspuru-Guzik. “Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms”. In: *Adv. Quantum Technol.* 2.12 (Oct. 2019), p. 1900070. DOI: [10.1002/qute.201900070](https://doi.org/10.1002/qute.201900070) (cit. on p. 28).
- [82] Jarrod R McClean et al. “Barren plateaus in quantum neural network training landscapes”. In: *Nat. Commun.* 9.1 (2018), p. 4812. DOI: <https://doi.org/10.1038/s41467-018-07090-4> (cit. on pp. 28, 49, 202, 203).
- [83] Marco Cerezo et al. “Cost function dependent barren plateaus in shallow parametrized quantum circuits”. In: *Nat. Commun.* 12.1 (2021), p. 1791. DOI: <https://doi.org/10.1038/s41467-021-21728-w> (cit. on pp. 28, 49).
- [84] Matthias C. Caro et al. “Encoding-dependent generalization bounds for parametrized quantum circuits”. In: *Quantum* 5 (Nov. 2021), p. 582. DOI: [10.22331/q-2021-11-17-582](https://doi.org/10.22331/q-2021-11-17-582) (cit. on pp. 28, 38).
- [85] Matthias C. Caro et al. “Generalization in quantum machine learning from few training data”. In: *Nat. Commun.* 13.1 (Aug. 2022), p. 4919. DOI: [10.1038/s41467-022-32550-3](https://doi.org/10.1038/s41467-022-32550-3) (cit. on pp. 28, 49).
- [86] Leonardo Banchi, Jason Pereira, and Stefano Pirandola. “Generalization in Quantum Machine Learning: A Quantum Information Standpoint”. In: *PRX Quantum* 2 (4 Nov. 2021), p. 040321. DOI: [10.1103/PRXQuantum.2.040321](https://doi.org/10.1103/PRXQuantum.2.040321) (cit. on pp. 28, 49, 77).
- [87] Francisco Javier Gil Vidal and Dirk Oliver Theis. “Input redundancy for parameterized quantum circuits”. In: *Front. Phys.* 8 (2020), p. 297. DOI: <https://doi.org/10.3389/fphy.2020.00297> (cit. on pp. 29, 40).

- [88] David Wierichs et al. “General parameter-shift rules for quantum gradients”. In: *Quantum* 6 (2022), p. 677. DOI: <https://doi.org/10.22331/q-2022-03-30-677> (cit. on p. 29).
- [89] Kendall E Atkinson. *An introduction to numerical analysis*. John Wiley & Sons, 2008 (cit. on p. 31).
- [90] Song Mei and Andrea Montanari. “The generalization error of random features regression: Precise asymptotics and the double descent curve”. In: *Commun. Pure Appl. Math.* 75.4 (2022), pp. 667–766. DOI: <https://doi.org/10.1002/cpa.22008> (cit. on p. 31).
- [91] Trevor Hastie et al. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *Ann. Stat.* 50.2 (2022), pp. 949–986. DOI: [10.1214/21-AOS2133](https://doi.org/10.1214/21-AOS2133) (cit. on p. 31).
- [92] Walter Rudin. “The Basic Theorems of Fourier Analysis”. In: *Fourier Analysis on Groups*. John Wiley & Sons, Ltd, 1990. DOI: <https://doi.org/10.1002/9781118165621.ch1> (cit. on p. 32).
- [93] Edward Farhi and Hartmut Neven. *Classification with Quantum Neural Networks on Near Term Processors*. 2018. DOI: <https://doi.org/10.48550/arXiv.1802.06002>. arXiv: [1802.06002](https://arxiv.org/abs/1802.06002) [quant-ph] (cit. on p. 39).
- [94] Maria Schuld et al. “Circuit-centric quantum classifiers”. In: *Phys. Rev. A* 101.3 (Mar. 2020), p. 032308. DOI: [10.1103/physreva.101.032308](https://doi.org/10.1103/physreva.101.032308) (cit. on p. 39).
- [95] Adrián Pérez-Salinas et al. “Data re-uploading for a universal quantum classifier”. In: *Quantum* 4 (Feb. 2020), p. 226. DOI: [10.22331/q-2020-02-06-226](https://doi.org/10.22331/q-2020-02-06-226) (cit. on pp. 39, 48, 217).
- [96] Casper Gyurik, Dyon Vreumingen van, and Vedran Dunjko. “Structural risk minimization for quantum linear classifiers”. In: *Quantum* 7 (Jan. 2023), p. 893. DOI: [10.22331/q-2023-01-13-893](https://doi.org/10.22331/q-2023-01-13-893) (cit. on p. 40).
- [97] S. Shin, Y. S. Teo, and H. Jeong. “Exponential data encoding for quantum supervised learning”. In: *Phys. Rev. A* 107 (1 Jan. 2023), p. 012422. DOI: [10.1103/PhysRevA.107.012422](https://doi.org/10.1103/PhysRevA.107.012422) (cit. on pp. 43, 205, 207).
- [98] Sophie Piccard. *Sur les ensembles de distances des ensembles de points d’un espace euclidien*. Memoires de l’Universite de Neuchatel. Secretariat de l’Universite, 1939 (cit. on pp. 43, 208).

- [99] Ville Bergholm et al. *PennyLane: Automatic differentiation of hybrid quantum-classical computations*. 2018. DOI: <https://doi.org/10.48550/arXiv.1811.04968>. arXiv: [1811.04968 \[quant-ph\]](#) (cit. on p. 46).
- [100] Dave Wecker et al. “Solving strongly correlated electron models on a quantum computer”. In: *Phys. Rev. A* 92 (6 Dec. 2015), p. 062318. DOI: [10.1103/PhysRevA.92.062318](#) (cit. on p. 47).
- [101] Ian D. Kivlichan et al. “Quantum Simulation of Electronic Structure with Linear Depth and Connectivity”. In: *Phys. Rev. Lett.* 120 (11 Mar. 2018), p. 110501. DOI: [10.1103/PhysRevLett.120.110501](#) (cit. on p. 47).
- [102] Martín Larocca et al. “Group-Invariant Quantum Machine Learning”. In: *PRX Quantum* 3 (3 Sept. 2022), p. 030341. DOI: [10.1103/PRXQuantum.3.030341](#) (cit. on p. 47).
- [103] Johannes Jakob Meyer et al. “Exploiting Symmetry in Variational Quantum Machine Learning”. In: *PRX Quantum* 4 (1 Mar. 2023), p. 010328. DOI: [10.1103/PRXQuantum.4.010328](#) (cit. on p. 47).
- [104] Hela Mhiri et al. *Constrained and Vanishing Expressivity of Quantum Fourier Models*. 2024. arXiv: [2403.09417 \[quant-ph\]](#) (cit. on p. 48).
- [105] Jonas Landman et al. “Classically Approximating Variational Quantum Machine Learning with Random Fourier Features”. In: *The Eleventh International Conference on Learning Representations*. 2023 (cit. on p. 48).
- [106] Franz J Schreiber, Jens Eisert, and Johannes Jakob Meyer. “Classical surrogates for quantum learning models”. In: *Physical Review Letters* 131.10 (2023), p. 100803 (cit. on p. 48).
- [107] Ryan Sweke et al. *Potential and limitations of random Fourier features for dequantizing quantum machine learning*. 2023. arXiv: [2309.11647 \[quant-ph\]](#) (cit. on p. 48).
- [108] Martin Larocca et al. “Theory of overparametrization in quantum neural networks”. In: *Nat. Comput. Sci.* 3.6 (2023), pp. 542–551. DOI: <https://doi.org/10.1038/s43588-023-00467-6> (cit. on p. 49).
- [109] Yuxuan Du et al. “Expressive power of parametrized quantum circuits”. In: *Phys. Rev. Res.* 2.3 (July 2020), p. 033125. DOI: [10.1103/physrevresearch.2.033125](#) (cit. on p. 49).

- [110] Zoë Holmes et al. “Connecting Ansatz Expressibility to Gradient Magnitudes and Barren Plateaus”. In: *PRX Quantum* 3 (1 Jan. 2022), p. 010313. DOI: [10.1103/PRXQuantum.3.010313](https://doi.org/10.1103/PRXQuantum.3.010313) (cit. on p. 49).
- [111] Samson Wang et al. “Noise-induced barren plateaus in variational quantum algorithms”. In: *Nat. Commun.* 12.1 (2021), p. 6961. DOI: <https://doi.org/10.1038/s41467-021-27045-6> (cit. on p. 49).
- [112] Hsin-Yuan Huang et al. “Quantum advantage in learning from experiments”. In: *Science* 376.6598 (2022), pp. 1182–1186. DOI: [10.1126/science.abn7293](https://doi.org/10.1126/science.abn7293) (cit. on p. 49).
- [113] Sitan Chen et al. “Exponential Separations Between Learning With and Without Quantum Memory”. In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. 2022, pp. 574–585. DOI: [10.1109/FOCS52979.2021.00063](https://doi.org/10.1109/FOCS52979.2021.00063) (cit. on p. 49).
- [114] Hsin-Yuan Huang, Richard Kueng, and John Preskill. “Information-Theoretic Bounds on Quantum Advantage in Machine Learning”. In: *Phys. Rev. Lett.* 126 (19 May 2021), p. 190505. DOI: [10.1103/PhysRevLett.126.190505](https://doi.org/10.1103/PhysRevLett.126.190505) (cit. on p. 49).
- [115] C. E. Shannon. “A mathematical theory of communication”. In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: [10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x) (cit. on pp. 51, 54, 76, 85, 146).
- [116] Koji Yamaguchi et al. “Superadditivity of channel capacity through quantum fields”. In: *Phys. Rev.D* 101.10 (2020), p. 105009 (cit. on p. 51).
- [117] C. E. Shannon. “Channels with Side Information at the Transmitter”. In: *IBM Journal of Research and Development* 2.4 (1958), pp. 289–293. DOI: [10.1147/rd.24.0289](https://doi.org/10.1147/rd.24.0289) (cit. on p. 52).
- [118] A.J. Goldsmith and P.P. Varaiya. “Capacity of fading channels with channel side information”. In: *IEEE Transactions on Information Theory* 43.6 (1997), pp. 1986–1992. DOI: [10.1109/18.641562](https://doi.org/10.1109/18.641562) (cit. on p. 52).
- [119] D. Slepian and J. Wolf. “Noiseless coding of correlated information sources”. In: *IEEE Trans. Inform. Theory* 19.4 (1973), pp. 471–480. DOI: [10.1109/TIT.1973.1055037](https://doi.org/10.1109/TIT.1973.1055037) (cit. on p. 52).
- [120] A. Wyner and J. Ziv. “A theorem on the entropy of certain binary sequences and applications–I”. In: *IEEE Trans. Inform. Theory* 19.6 (1973). Conference Name: IEEE Trans. Inform. Theory, pp. 769–772. DOI: [10.1109/TIT.1973.1055107](https://doi.org/10.1109/TIT.1973.1055107) (cit. on p. 52).

- [121] A. Wyner. “A theorem on the entropy of certain binary sequences and applications-II”. In: *IEEE Trans. Inform. Theory* 19.6 (1973), pp. 772–777. DOI: [10.1109/TIT.1973.1055108](https://doi.org/10.1109/TIT.1973.1055108) (cit. on p. 52).
- [122] J. Korner and K. Marton. “How to encode the modulo-two sum of binary sources (Corresp.)”. In: *IEEE Trans. Inform. Theory* 25.2 (1979). Conference Name: IEEE Trans. Inform. Theory, pp. 219–221. DOI: [10.1109/TIT.1979.1056022](https://doi.org/10.1109/TIT.1979.1056022) (cit. on p. 52).
- [123] U. Erez and R. Zamir. “Noise prediction for channels with side information at the transmitter”. In: *IEEE Transactions on Information Theory* 46.4 (July 2000). Conference Name: IEEE Transactions on Information Theory, pp. 1610–1617. DOI: [10.1109/18.850704](https://doi.org/10.1109/18.850704) (cit. on p. 52).
- [124] U. Erez, R. Zamir, and S. Shamai. “Additive noise channels with side information at the transmitter”. In: *21st IEEE Convention of the Electrical and Electronic Engineers in Israel. Proceedings (Cat. No.00EX377)*. 21st IEEE Convention of the Electrical and Electronic Engineers in Israel. Proceedings (Cat. No.00EX377). Apr. 2000, pp. 373–376. DOI: [10.1109/EEEI.2000.924437](https://doi.org/10.1109/EEEI.2000.924437) (cit. on p. 52).
- [125] Patrick Hayden and Christopher King. “Correcting Quantum Channels by Measuring the Environment”. In: *Quantum Info. Comput.* 5.2 (Mar. 2005), pp. 156–160. DOI: <https://doi.org/10.48550/arXiv.quant-ph/0409026> (cit. on pp. 52, 66, 69).
- [126] Siddharth Karumanchi et al. “Classical capacities of quantum channels with environment assistance”. In: *Probl Inf Transm* 52.3 (2016), pp. 214–238. DOI: [10.1134/S0032946016030029](https://doi.org/10.1134/S0032946016030029) (cit. on pp. 52, 66).
- [127] Robert Fano. *Class notes for MIT course 6.574: Transmission of information*. Feb. 1952 (cit. on pp. 54, 76).
- [128] Vladimir Kovalevsky. “The problem of character recognition from the point of view of mathematical statistics”. In: *Character Readers and Pattern Recognition* (1968), pp. 3–30 (cit. on pp. 54, 59).
- [129] M. Hellman and J. Raviv. “Probability of error, equivocation, and the Chernoff bound”. In: *IEEE Trans. Inf. Theory* 16.4 (1970), pp. 368–372. DOI: [10.1109/TIT.1970.1054466](https://doi.org/10.1109/TIT.1970.1054466) (cit. on pp. 54, 59).
- [130] M. Feder and N. Merhav. “Relations between entropy and error probability”. In: *IEEE Trans. Inf. Theory* 40.1 (Jan. 1994), pp. 259–266. DOI: [10.1109/18.272494](https://doi.org/10.1109/18.272494) (cit. on pp. 54, 59, 84, 86).

- [131] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999 (cit. on pp. 54, 56, 84).
- [132] Alastair Kay. *Quantikz*. 2019. DOI: [10.17637/RH.7000520](https://doi.org/10.17637/RH.7000520).
- [133] Raymond Laflamme et al. “Perfect Quantum Error Correcting Code”. In: *Phys. Rev. Lett.* 77 (1 July 1996), pp. 198–201. DOI: [10.1103/PhysRevLett.77.198](https://doi.org/10.1103/PhysRevLett.77.198) (cit. on p. 64).
- [134] Charles H. Bennett et al. “Mixed-state entanglement and quantum error correction”. In: *Phys. Rev. A* 54 (5 Nov. 1996), pp. 3824–3851. DOI: [10.1103/PhysRevA.54.3824](https://doi.org/10.1103/PhysRevA.54.3824) (cit. on p. 64).
- [135] Emanuel Knill, Raymond Laflamme, and Lorenza Viola. “Theory of Quantum Error Correction for General Noise”. In: *Phys. Rev. Lett.* 84.11 (Mar. 2000), pp. 2525–2528. DOI: [10.1103/physrevlett.84.2525](https://doi.org/10.1103/physrevlett.84.2525) (cit. on p. 65).
- [136] W. Forrest Stinespring. “Positive Functions on C*-Algebras”. In: *Proceedings of the American Mathematical Society* 6.2 (1955), pp. 211–216 (cit. on pp. 66, 96).
- [137] John A. Smolin, Frank Verstraete, and Andreas Winter. “Entanglement of assistance and multipartite state distillation”. In: *Phys. Rev. A* 72 (5 Nov. 2005), p. 052317. DOI: [10.1103/PhysRevA.72.052317](https://doi.org/10.1103/PhysRevA.72.052317) (cit. on p. 66).
- [138] Andreas Winter. “On environment-assisted capacities of quantum channels”. In: *Markov Proc. Rel. Fields* 13 (1–2 2007), pp. 297–314. DOI: <https://doi.org/10.48550/arXiv.quant-ph/0507045> (cit. on p. 66).
- [139] Seth Lloyd. “Capacity of the noisy quantum channel”. In: *Phys. Rev. A* 55 (3 Mar. 1997), pp. 1613–1622. DOI: [10.1103/PhysRevA.55.1613](https://doi.org/10.1103/PhysRevA.55.1613) (cit. on p. 67).
- [140] I. Devetak. “The private classical capacity and quantum capacity of a quantum channel”. In: *IEEE Trans. Inf. Theory* 51.1 (Jan. 2005), pp. 44–55. DOI: [10.1109/TIT.2004.839515](https://doi.org/10.1109/TIT.2004.839515) (cit. on p. 67).
- [141] Howard Barnum, M. A. Nielsen, and Benjamin Schumacher. “Information transmission through a noisy quantum channel”. In: *Phys. Rev. A* 57 (6 June 1998), pp. 4153–4175. DOI: [10.1103/PhysRevA.57.4153](https://doi.org/10.1103/PhysRevA.57.4153) (cit. on p. 67).
- [142] Peter Shor. “The quantum channel capacity and coherent information”. In: *lecture notes, MSRI Workshop on Quantum Computation (Quantum Information and Cryptography)* (2002) (cit. on p. 67).

- [143] fork40. *Does no-cloning imply zero capacity of antidegradable channels?* Quantum Computing Stack Exchange. <https://quantumcomputing.stackexchange.com/questions/37848/does-no-cloning-imply-zero-capacity-of-antidegradable-channels> (cit. on p. 67).
- [144] Igor Devetak and Peter W Shor. “The capacity of a quantum channel for simultaneous transmission of classical and quantum information”. In: *Commun. Math. Phys.* 256 (2005), pp. 287–303. DOI: <https://doi.org/10.1007/s00220-005-1317-6> (cit. on pp. 67, 70).
- [145] Farzad Kianvash, Marco Fanizza, and Vittorio Giovannetti. “Bounding the quantum capacity with flagged extensions”. In: *Quantum* 6 (Feb. 2022), p. 647. DOI: [10.22331/q-2022-02-09-647](https://doi.org/10.22331/q-2022-02-09-647) (cit. on p. 70).
- [146] Michael M. Wolf and David Pérez-García. “Quantum capacities of channels with small environment”. In: *Phys. Rev. A* 75 (1 Jan. 2007), p. 012303. DOI: [10.1103/PhysRevA.75.012303](https://doi.org/10.1103/PhysRevA.75.012303) (cit. on p. 70).
- [147] Antonio D’Arrigo, Giuliano Benenti, and Giuseppe Falci. “Quantum capacity of dephasing channels with memory”. In: *New Journal of Physics* 9.9 (Sept. 2007), pp. 310–310. DOI: [10.1088/1367-2630/9/9/310](https://doi.org/10.1088/1367-2630/9/9/310) (cit. on p. 70).
- [148] Mark M. Wilde. *From Classical to Quantum Shannon Theory*. Nov. 2016. DOI: [10.1017/9781316809976.001](https://doi.org/10.1017/9781316809976.001). arXiv: [1106.1445\[quant-ph\]](https://arxiv.org/abs/1106.1445) (cit. on p. 73).
- [149] I. A. Ibragimov, R. Z. Has’minskii, and Samuel Kotz. *Statistical estimation: Asymptotic theory*. eng. 1st ed. 1981. Stochastic Modelling and Applied Probability, 16. New York: Springer Science+Business Media, LLC, 1981 (cit. on pp. 76, 79).
- [150] Emmanuel J. Candes and Terence Tao. “Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?” In: *IEEE Trans. Inf. Theory* 52.12 (2006), pp. 5406–5425. DOI: [10.1109/TIT.2006.885507](https://doi.org/10.1109/TIT.2006.885507) (cit. on p. 76).
- [151] D.L. Donoho. “Compressed sensing”. In: *IEEE Trans. Inf. Theory* 52.4 (2006), pp. 1289–1306. DOI: [10.1109/TIT.2006.871582](https://doi.org/10.1109/TIT.2006.871582) (cit. on p. 76).
- [152] Emmanuel J. Candes and Terence Tao. “The Power of Convex Relaxation: Near-Optimal Matrix Completion”. In: *IEEE Trans. Inf. Theory* 56.5 (2010), pp. 2053–2080. DOI: [10.1109/TIT.2010.2044061](https://doi.org/10.1109/TIT.2010.2044061) (cit. on p. 76).
- [153] Emmanuel Candes and Benjamin Recht. “Exact matrix completion via convex optimization”. In: *Found Comput Math* 9 (2009), pp. 717–772. DOI: <https://doi.org/10.1007/s10208-009-9045-5> (cit. on p. 76).

- [154] Wei Wang, Martin J. Wainwright, and Kannan Ramchandran. “Information-theoretic limits on sparse support recovery: Dense versus sparse measurements”. In: *2008 IEEE International Symposium on Information Theory*. 2008, pp. 2197–2201. DOI: [10.1109/ISIT.2008.4595380](https://doi.org/10.1109/ISIT.2008.4595380) (cit. on p. 76).
- [155] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. “Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q -Balls”. In: *IEEE Trans. Inf. Theory* 57.10 (Oct. 2011), pp. 6976–6994. DOI: [10.1109/TIT.2011.2165799](https://doi.org/10.1109/TIT.2011.2165799) (cit. on pp. 76, 146).
- [156] Po-Ling Loh and Martin J. Wainwright. “Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression”. In: *2012 IEEE International Symposium on Information Theory Proceedings*. 2012 IEEE International Symposium on Information Theory - ISIT. Cambridge, MA, USA: IEEE, July 2012, pp. 2601–2605. DOI: [10.1109/ISIT.2012.6283989](https://doi.org/10.1109/ISIT.2012.6283989) (cit. on pp. 76, 106).
- [157] David Gross et al. “Quantum State Tomography via Compressed Sensing”. In: *Phys. Rev. Lett.* 105 (15 Oct. 2010), p. 150401. DOI: [10.1103/PhysRevLett.105.150401](https://doi.org/10.1103/PhysRevLett.105.150401) (cit. on p. 76).
- [158] Charles H. Bennett et al. “Purification of Noisy Entanglement and Faithful Teleportation via Noisy Channels”. In: *Phys. Rev. Lett.* 76 (5 Jan. 1996), pp. 722–725. DOI: [10.1103/PhysRevLett.76.722](https://doi.org/10.1103/PhysRevLett.76.722) (cit. on p. 76).
- [159] Charles H. Bennett et al. “Mixed-state entanglement and quantum error correction”. In: *Phys. Rev. A* 54 (5 Nov. 1996), pp. 3824–3851. DOI: [10.1103/PhysRevA.54.3824](https://doi.org/10.1103/PhysRevA.54.3824) (cit. on p. 76).
- [160] Frank Verstraete and Henri Verschelde. “Fidelity of mixed states of two qubits”. In: *Phys. Rev. A* 66.2 (Aug. 2002). DOI: [10.1103/physreva.66.022307](https://doi.org/10.1103/physreva.66.022307) (cit. on p. 76).
- [161] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. “General teleportation channel, singlet fraction, and quasidistillation”. In: *Phys. Rev. A* 60 (3 Sept. 1999), pp. 1888–1898. DOI: [10.1103/PhysRevA.60.1888](https://doi.org/10.1103/PhysRevA.60.1888) (cit. on p. 76).
- [162] Robert Koenig, Renato Renner, and Christian Schaffner. “The operational meaning of min- and max-entropy”. In: *IEEE Trans. Inf. Theory* 55.9 (Sept. 2009), pp. 4337–4347. DOI: [10.1109/TIT.2009.2025545](https://doi.org/10.1109/TIT.2009.2025545) (cit. on pp. 77, 86, 87).
- [163] Hsin-Yuan Huang et al. “Power of data in quantum machine learning”. In: *Nat. Commun.* 12.1 (May 2021). DOI: [10.1038/s41467-021-22539-9](https://doi.org/10.1038/s41467-021-22539-9) (cit. on p. 77).
- [164] Matthias C. Caro et al. “Generalization in quantum machine learning from few training data”. In: *Nat. Commun.* 13.1 (Aug. 2022), p. 4919. DOI: [10.1038/s41467-022-32550-3](https://doi.org/10.1038/s41467-022-32550-3) (cit. on p. 77).

- [165] Evan Peters and Maria Schuld. “Generalization despite overfitting in quantum machine learning models”. In: *Quantum* 7 (Dec. 2023), p. 1210. DOI: [10.22331/q-2023-12-20-1210](https://doi.org/10.22331/q-2023-12-20-1210) (cit. on p. 77).
- [166] Matthias C. Caro et al. “Out-of-distribution generalization for learning quantum dynamics”. In: *Nat. Commun.* 14.1 (July 2023). DOI: [10.1038/s41467-023-39381-w](https://doi.org/10.1038/s41467-023-39381-w) (cit. on p. 77).
- [167] Elies Gil-Fuster, Jens Eisert, and Carlos Bravo-Prieto. “Understanding quantum machine learning also requires rethinking generalization”. In: *Nat. Commun.* 15.1 (Mar. 2024). DOI: [10.1038/s41467-024-45882-z](https://doi.org/10.1038/s41467-024-45882-z) (cit. on p. 77).
- [168] Hsin-Yuan Huang, Richard Kueng, and John Preskill. “Information-Theoretic Bounds on Quantum Advantage in Machine Learning”. In: *Phys. Rev. Lett.* 126 (19 May 2021), p. 190505. DOI: [10.1103/PhysRevLett.126.190505](https://doi.org/10.1103/PhysRevLett.126.190505) (cit. on p. 77).
- [169] Matthias Caro et al. *Information-theoretic generalization bounds for learning from quantum data*. arXiv:2311.05529. Nov. 2023. arXiv: [2311.05529](https://arxiv.org/abs/2311.05529) [quant-ph] (cit. on p. 77).
- [170] Leslie G Valiant. “A theory of the learnable”. In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142 (cit. on p. 78).
- [171] John Duchi. *Lecture Notes for Statistics 311/Electrical Engineering 377*. Stanford, Nov. 2021 (cit. on p. 79).
- [172] Johannes Jakob Meyer et al. “Quantum metrology in the finite-sample regime”. In: *Quantum Information Processing 2024*. Jan. 2024. arXiv: [2307.06370](https://arxiv.org/abs/2307.06370) [quant-ph] (cit. on p. 81).
- [173] Benjamin Schumacher. “Quantum coding”. In: *Phys. Rev. A* 51 (4 Apr. 1995), pp. 2738–2747. DOI: [10.1103/PhysRevA.51.2738](https://doi.org/10.1103/PhysRevA.51.2738) (cit. on p. 85).
- [174] A. Jamiołkowski. “Linear transformations which preserve trace and positive semidefiniteness of operators”. In: *Reports on Mathematical Physics* 3.4 (1972), pp. 275–278. DOI: [https://doi.org/10.1016/0034-4877\(72\)90011-0](https://doi.org/10.1016/0034-4877(72)90011-0) (cit. on p. 85).
- [175] Man-Duen Choi. “Completely positive linear maps on complex matrices”. In: *Linear Algebra and its Applications* 10.3 (1975), pp. 285–290. DOI: [https://doi.org/10.1016/0024-3795\(75\)90075-0](https://doi.org/10.1016/0024-3795(75)90075-0) (cit. on p. 85).
- [176] John Watrous. *Is there any connection between the diamond norm and the distance of the associated states?* Theoretical Computer Science Stack Exchange. <https://cstheory.stackexchange.com/q/4920> (cit. on p. 85).

- [177] John Watrous. *Advanced topics in Quantum Information Theory*. Lecture Notes. Available at <https://cs.uwaterloo.ca/~watrous/QIT-notes/>. 2020 (cit. on p. 86).
- [178] Benjamin Schumacher. “Sending entanglement through noisy quantum channels”. In: *Phys. Rev. A* 54 (4 Oct. 1996), pp. 2614–2628. DOI: [10.1103/PhysRevA.54.2614](https://doi.org/10.1103/PhysRevA.54.2614) (cit. on pp. 86, 87).
- [179] K. M. R. Audenaert et al. “Discriminating States: The Quantum Chernoff Bound”. In: *Phys. Rev. Lett.* 98 (16 Apr. 2007), p. 160501. DOI: [10.1103/PhysRevLett.98.160501](https://doi.org/10.1103/PhysRevLett.98.160501) (cit. on p. 87).
- [180] K. M. R. Audenaert et al. “Asymptotic Error Rates in Quantum Hypothesis Testing”. In: *Comm. Math. Phys.* 279.1 (Apr. 2008), pp. 251–283. DOI: [10.1007/s00220-008-0417-5](https://doi.org/10.1007/s00220-008-0417-5) (cit. on p. 87).
- [181] Casper Gyurik, Chris Cade, and Vedran Dunjko. “Towards quantum advantage via topological data analysis”. In: *Quantum* 6 (Nov. 2022), p. 855. DOI: [10.22331/q-2022-11-10-855](https://doi.org/10.22331/q-2022-11-10-855) (cit. on p. 88).
- [182] Hsin-Yuan Huang et al. “Provably efficient machine learning for quantum many-body problems”. In: *Science* 377.6613 (Sept. 2022). DOI: [10.1126/science.abk3333](https://doi.org/10.1126/science.abk3333) (cit. on p. 88).
- [183] Casper Gyurik and Vedran Dunjko. *Exponential separations between classical and quantum learners*. 2023. arXiv: [2306.16028](https://arxiv.org/abs/2306.16028) [quant-ph] (cit. on p. 88).
- [184] Sumeet Khatri et al. “Quantum-assisted quantum compiling”. In: *Quantum* 3 (May 2019), p. 140. DOI: [10.22331/q-2019-05-13-140](https://doi.org/10.22331/q-2019-05-13-140) (cit. on p. 88).
- [185] Kunal Sharma et al. “Reformulation of the No-Free-Lunch Theorem for Entangled Datasets”. In: *Phys. Rev. Lett.* 128 (7 Feb. 2022), p. 070501. DOI: [10.1103/PhysRevLett.128.070501](https://doi.org/10.1103/PhysRevLett.128.070501) (cit. on p. 88).
- [186] Matthias C. Caro et al. “Out-of-distribution generalization for learning quantum dynamics”. In: *Nat. Commun.* 14.1 (July 2023). DOI: [10.1038/s41467-023-39381-w](https://doi.org/10.1038/s41467-023-39381-w) (cit. on p. 88).
- [187] Matthias C. Caro. *Learning Quantum Processes and Hamiltonians via the Pauli Transfer Matrix*. 2023. arXiv: [2212.04471](https://arxiv.org/abs/2212.04471) [quant-ph] (cit. on p. 88).
- [188] Hsin-Yuan Huang, Sitan Chen, and John Preskill. “Learning to Predict Arbitrary Quantum Processes”. In: *PRX Quantum* 4 (4 Dec. 2023), p. 040337. DOI: [10.1103/PRXQuantum.4.040337](https://doi.org/10.1103/PRXQuantum.4.040337) (cit. on p. 88).

- [189] Sofiene Jerbi et al. *The power and limitations of learning quantum dynamics incoherently*. 2023. arXiv: [2303.12834 \[quant-ph\]](#) (cit. on p. 88).
- [190] Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd. “Quantum Algorithm for Linear Systems of Equations”. In: *Phys. Rev. Lett.* 103 (15 Oct. 2009), p. 150502. DOI: [10.1103/PhysRevLett.103.150502](#) (cit. on p. 88).
- [191] Dana Angluin. “Queries and concept learning”. In: *Machine learning* 2 (1988), pp. 319–342 (cit. on p. 89).
- [192] Andris Ambainis et al. “Quantum Identification of Boolean Oracles”. In: *Proceedings of the 21st International Symposium on Theoretical Aspects of Computer Science (STACS 2004)*. Springer, Berlin, Heidelberg, 2004, pp. 105–116. DOI: [https://doi.org/10.1007/978-3-540-24749-4_10](#) (cit. on p. 89).
- [193] Robin Kothari. “An optimal quantum algorithm for the oracle identification problem”. In: *31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014)*. Vol. 25. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2014, pp. 482–493. DOI: [10.4230/LIPIcs.STACS.2014.482](#) (cit. on p. 89).
- [194] A. Yu. Khrennikov. “Two-particle wave function as an integral operator and the random field approach to quantum correlations”. In: *Theoretical and Mathematical Physics* 164.3 (Sept. 2010), pp. 1156–1162. DOI: [10.1007/s11232-010-0094-3](#) (cit. on p. 91).
- [195] Jens Eisert, Christoph Simon, and Martin B Plenio. “On the quantification of entanglement in infinite-dimensional quantum systems”. In: *Journal of Physics A: Mathematical and General* 35.17 (2002), p. 3911 (cit. on p. 92).
- [196] J. Eisert and M. B. Plenio. “Introduction to the basics of entanglement theory in continuous-variable systems”. In: *International Journal of Quantum Information* 1.04 (2003), pp. 479–506 (cit. on p. 92).
- [197] A. S. Holevo, Maksim Shirokov, and Reinhard Werner. “On the notion of entanglement in Hilbert spaces”. In: *Russian Mathematical Surveys* 60.2 (2005), pp. 359–360 (cit. on p. 92).
- [198] Gerardo Adesso and Fabrizio Illuminati. “Entanglement in continuous-variable systems: recent advances and current perspectives”. In: *Journal of Physics A: Mathematical and Theoretical* 40.28 (June 2007), p. 7821. DOI: [10.1088/1751-8113/40/28/S01](#) (cit. on p. 92).
- [199] Fabian Furrer, Johan Åberg, and Renato Renner. “Min- and Max-Entropy in Infinite Dimensions”. In: *Comm. Math. Phys.* 306.1 (Aug. 2011), pp. 165–186. DOI: [10.1007/s00220-011-1282-1](#) (cit. on p. 93).

- [200] Bartosz Regula et al. “Operational Quantification of Continuous-Variable Quantum Resources”. In: *Phys. Rev. Lett.* 126.11 (Mar. 2021). DOI: [10.1103/physrevlett.126.110403](https://doi.org/10.1103/physrevlett.126.110403) (cit. on p. 93).
- [201] Erkkä Haapasalo et al. “Operational Characterization of Infinite-Dimensional Quantum Resources”. In: *Phys. Rev. Lett.* 127.25 (Dec. 2021). DOI: [10.1103/physrevlett.127.250401](https://doi.org/10.1103/physrevlett.127.250401) (cit. on p. 93).
- [202] Hayata Yamasaki et al. *Entanglement cost for infinite-dimensional physical systems*. 2024. arXiv: [2401.09554](https://arxiv.org/abs/2401.09554) [[quant-ph](#)] (cit. on p. 93).
- [203] A. S. Holevo. “The Choi–Jamiołkowski forms of quantum Gaussian channels”. In: *Journal of Mathematical Physics* 52.4 (Apr. 2011). DOI: [10.1063/1.3581879](https://doi.org/10.1063/1.3581879) (cit. on p. 93).
- [204] A. S. Holevo, M. E. Shirokov, and R. F. Werner. *Separability and Entanglement-Breaking in Infinite Dimensions*. 2005. arXiv: [quant-ph/0504204](https://arxiv.org/abs/quant-ph/0504204) [[quant-ph](#)] (cit. on p. 93).
- [205] A. S. Holevo. *Statistical structure of quantum theory*. Lecture Notes in Physics Monographs 67. Berlin Heidelberg: Springer, 2001. DOI: <https://doi.org/10.1007/3-540-44998-1> (cit. on p. 93).
- [206] Stephane Attal. *Lectures in Quantum Noise Theory*. 2014 (cit. on p. 96).
- [207] Karl Kraus et al. *States, Effects, and Operations Fundamental Notions of Quantum Theory*. Vol. 190. Springer, 1983. DOI: [10.1007/3-540-12732-1](https://doi.org/10.1007/3-540-12732-1) (cit. on p. 96).
- [208] Laura J. Henderson, Rishi Goel, and Sally Shrapnel. *A General Form for Continuous Variable Quantum Kernels*. 2024. arXiv: [2401.05647](https://arxiv.org/abs/2401.05647) (cit. on p. 107).
- [209] David Haussler. “Decision theoretic generalizations of the PAC model for neural net and other learning applications”. In: *Information and Computation* 100.1 (1992), pp. 78–150. DOI: [https://doi.org/10.1016/0890-5401\(92\)90010-D](https://doi.org/10.1016/0890-5401(92)90010-D) (cit. on p. 107).
- [210] John Preskill. “Quantum Computing in the NISQ era and beyond”. In: *Quantum* 2 (Aug. 2018), p. 79. DOI: [10.22331/q-2018-08-06-79](https://doi.org/10.22331/q-2018-08-06-79) (cit. on p. 109).
- [211] Matthew Neeley et al. “Generation of three-qubit entangled states using superconducting phase qubits”. In: *Nature* 467.7315 (Sept. 2010), pp. 570–573. DOI: [10.1038/nature09418](https://doi.org/10.1038/nature09418) (cit. on pp. 109, 131).
- [212] D. Willsch et al. “Testing quantum fault tolerance on small systems”. In: *Phys. Rev. A* 98.5 (Nov. 2018). DOI: [10.1103/physreva.98.052348](https://doi.org/10.1103/physreva.98.052348) (cit. on p. 109).

- [213] A. Dewes et al. “Characterization of a Two-Transmon Processor with Individual Single-Shot Qubit Readout”. In: *Physical Review Letters* 108.5 (Feb. 2012). DOI: [10.1103/physrevlett.108.057002](https://doi.org/10.1103/physrevlett.108.057002) (cit. on pp. 109, 131).
- [214] Yanzhu Chen et al. “Detector tomography on IBM quantum computers and mitigation of an imperfect measurement”. In: *Phys. Rev. A* 100.5 (Nov. 2019). DOI: [10.1103/physreva.100.052315](https://doi.org/10.1103/physreva.100.052315) (cit. on p. 109).
- [215] Ming Gong et al. “Genuine 12-Qubit Entanglement on a Superconducting Quantum Processor”. In: *Phys. Rev. Lett.* 122.11 (Mar. 2019). DOI: [10.1103/physrevlett.122.110501](https://doi.org/10.1103/physrevlett.122.110501) (cit. on p. 109).
- [216] Ken X. Wei et al. “Verifying multipartite entangled Greenberger-Horne-Zeilinger states via multiple quantum coherences”. In: *Phys. Rev. A* 101.3 (Mar. 2020). DOI: [10.1103/physreva.101.032343](https://doi.org/10.1103/physreva.101.032343) (cit. on p. 109).
- [217] Miroslav Urbanek, Benjamin Nachman, and Wibe A. de Jong. “Error detection on quantum computers improving the accuracy of chemical calculations”. In: *Phys. Rev. A* 102.2 (Aug. 2020). DOI: [10.1103/physreva.102.022427](https://doi.org/10.1103/physreva.102.022427) (cit. on pp. 109, 120).
- [218] Benjamin Nachman et al. “Unfolding quantum computer readout noise”. In: *npj Quantum Inf.* 6.1 (2020), pp. 1–7. DOI: [10.1038/s41534-020-00309-7](https://doi.org/10.1038/s41534-020-00309-7) (cit. on pp. 109, 120, 131).
- [219] Rebecca Hicks, Christian W. Bauer, and Benjamin Nachman. “Readout rebalancing for near-term quantum computers”. In: *Phys. Rev. A* 103 (2 Feb. 2021). DOI: [10.1103/PhysRevA.103.022407](https://doi.org/10.1103/PhysRevA.103.022407) (cit. on pp. 109, 120, 223).
- [220] Lena Funcke et al. “Measurement error mitigation in quantum computers through classical bit-flip correction”. In: *Phys. Rev. A* 105.6 (June 2022). DOI: [10.1103/physreva.105.062404](https://doi.org/10.1103/physreva.105.062404) (cit. on p. 109).
- [221] Sergey Bravyi et al. “Mitigating measurement errors in multiqubit experiments”. In: *Phys. Rev. A* 103.4 (Apr. 2021). DOI: [10.1103/physreva.103.042605](https://doi.org/10.1103/physreva.103.042605) (cit. on pp. 109, 113, 131).
- [222] Ewout van den Berg, Zlatko K. Mineev, and Kristan Temme. “Model-free readout-error mitigation for quantum expectation values”. In: *Phys. Rev. A* 105 (3 Mar. 2022), p. 032620. DOI: [10.1103/PhysRevA.105.032620](https://doi.org/10.1103/PhysRevA.105.032620) (cit. on p. 109).
- [223] Benjamin Nachman and Michael R. Geller. *Categorizing Readout Error Correlations on Near Term Quantum Computers*. 2021. DOI: [10.48550/ARXIV.2104.04607](https://doi.org/10.48550/ARXIV.2104.04607). arXiv: [2104.04607](https://arxiv.org/abs/2104.04607) [quant-ph] (cit. on pp. 110, 120).

- [224] Paul D. Nation et al. “Scalable Mitigation of Measurement Errors on Quantum Computers”. In: *PRX Quantum* 2 (4 Nov. 2021), p. 040326. DOI: [10.1103/PRXQuantum.2.040326](https://doi.org/10.1103/PRXQuantum.2.040326) (cit. on pp. 111, 121, 131, 228).
- [225] Bo Yang, Rudy Raymond, and Shumpei Uno. “Efficient quantum readout-error mitigation for sparse measurement outcomes of near-term quantum devices”. In: *Phys. Rev. A* 106 (1 July 2022), p. 012423. DOI: [10.1103/PhysRevA.106.012423](https://doi.org/10.1103/PhysRevA.106.012423) (cit. on p. 111).
- [226] Mingxia Huo and Ying Li. “Dual-state purification for practical quantum error mitigation”. In: *Phys. Rev. A* 105 (2 Feb. 2022), p. 022427. DOI: [10.1103/PhysRevA.105.022427](https://doi.org/10.1103/PhysRevA.105.022427) (cit. on p. 111).
- [227] Evan Peters et al. “Qubit Assignment Using Time Reversal”. In: *PRX Quantum* 3 (4 Dec. 2022), p. 040333. DOI: [10.1103/PRXQuantum.3.040333](https://doi.org/10.1103/PRXQuantum.3.040333) (cit. on p. 111).
- [228] Sumeet Khatri et al. “Quantum-assisted quantum compiling”. In: *Quantum* 3 (May 2019), p. 140. DOI: [10.22331/q-2019-05-13-140](https://doi.org/10.22331/q-2019-05-13-140) (cit. on pp. 111, 140).
- [229] Michael R Geller and Mingyu Sun. “Toward efficient correction of multiqubit measurement errors: pair correlation method”. In: *Quantum Sci. Technol.* 6.2 (Feb. 2021), p. 025009. DOI: [10.1088/2058-9565/abd5c9](https://doi.org/10.1088/2058-9565/abd5c9) (cit. on p. 116).
- [230] Kun Wang, Yu-Ao Chen, and Xin Wang. *Measurement Error Mitigation via Truncated Neumann Series*. 2021. arXiv: [2103.13856](https://arxiv.org/abs/2103.13856) [quant-ph] (cit. on p. 118).
- [231] Richard J Lipton, Donald J Rose, and Robert Endre Tarjan. “Generalized nested dissection”. In: *SIAM journal on numerical analysis* 16.2 (1979), pp. 346–358 (cit. on p. 120).
- [232] Emrah Kılıç and Pantelimon Stanica. “The inverse of banded matrices”. In: *Journal of Computational and Applied Mathematics* 237.1 (2013), pp. 126–135. DOI: <https://doi.org/10.1016/j.cam.2012.07.018> (cit. on p. 120).
- [233] Will Finigan et al. “Qubit Allocation for Noisy Intermediate-Scale Quantum Computers”. In: *arXiv:1810.08291* (Oct. 2018) (cit. on pp. 125, 126, 131).
- [234] Swamit S. Tannu and Moinuddin K. Qureshi. “Not all qubits are created equal”. In: *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. Providence RI USA: ACM, Apr. 2019, pp. 987–999 (cit. on p. 125).
- [235] Joseph Emerson, Robert Alicki, and Karol Życzkowski. “Scalable noise estimation with random unitary operators”. In: *J. opt. B Quantum semiclass. opt.* 7.10 (Sept. 2005), pp. 347–352 (cit. on pp. 125, 131).

- [236] E. Knill et al. “Randomized benchmarking of quantum gates”. In: *Phys. Rev. A* 77.1 (Jan. 2008) (cit. on pp. [125](#), [131](#), [132](#)).
- [237] Sergio Boixo et al. “Characterizing quantum supremacy in near-term devices”. In: *Nature Physics* 14.6 (Apr. 2018), pp. 595–600. DOI: [10.1038/s41567-018-0124-x](#) (cit. on pp. [125](#), [131](#), [234](#)).
- [238] C. Neill et al. “A blueprint for demonstrating quantum supremacy with superconducting qubits”. In: *Science* 360.6385 (Apr. 2018), pp. 195–199. DOI: [10.1126/science.aao4309](#) (cit. on pp. [125](#), [131](#)).
- [239] Zijun Chen et al. “Exponential suppression of bit or phase errors with cyclic error correction”. In: *Nature* 595.7867 (July 2021), pp. 383–387. DOI: [10.1038/s41586-021-03588-y](#) (cit. on p. [126](#)).
- [240] M. McEwen et al. “Removing leakage-induced correlated errors in superconducting quantum error correction”. In: *Nature Communications* 12.1 (Mar. 2021), p. 1761. DOI: [10.1038/s41467-021-21982-y](#) (cit. on p. [126](#)).
- [241] Edward H. Chen et al. “Calibrated Decoders for Experimental Quantum Error Correction”. In: *Phys. Rev. Lett.* 128 (11 Mar. 2022), p. 110504. DOI: [10.1103/PhysRevLett.128.110504](#) (cit. on p. [126](#)).
- [242] Youwei Zhao et al. “Realization of an Error-Correcting Surface Code with Superconducting Qubits”. In: *Phys. Rev. Lett.* 129 (3 July 2022), p. 030501. DOI: [10.1103/PhysRevLett.129.030501](#) (cit. on p. [126](#)).
- [243] Rajeev Acharya et al. *Suppressing quantum errors by scaling a surface code logical qubit*. 2022. DOI: [10.48550/ARXIV.2207.06431](#) (cit. on p. [126](#)).
- [244] Asher Peres. “Stability of quantum motion in chaotic and regular systems”. In: *Phys. Rev. A* 30 (4 Oct. 1984), pp. 1610–1615. DOI: [10.1103/PhysRevA.30.1610](#) (cit. on pp. [126](#), [129](#)).
- [245] Timothy Proctor et al. “Measuring the Capabilities of Quantum Computers”. In: *Nat. Phys.* (2021) (cit. on pp. [126](#), [136](#)).
- [246] Timothy Proctor et al. *Scalable randomized benchmarking of quantum computers using mirror circuits*. 2021. arXiv: [2112.09853 \[quant-ph\]](#) (cit. on pp. [126](#), [136](#)).
- [247] Daniel Gottesman. “The Heisenberg representation of quantum computers”. In: *Proceedings of the XXII International Colloquium on Group Theoretical Methods in Physics*. Cambridge, MA, 1998, pp. 32–43. arXiv: [quant-ph/9807006](#) (cit. on pp. [126](#), [136](#)).

- [248] Scott Aaronson and Daniel Gottesman. “Improved simulation of stabilizer circuits”. In: *Phys. Rev. A* 70.5 (Nov. 2004) (cit. on p. 126).
- [249] Shin Nishio et al. “Extracting Success from IBM’s 20-Qubit Machines Using Error-Aware Compilation”. In: *ACM J. Emerg. Technol. Comput.* 16.3 (July 2020), pp. 1–25 (cit. on p. 126).
- [250] Dmitri Maslov, Sean M. Falconer, and Michele Mosca. “Quantum Circuit Placement”. In: *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 27.4 (Apr. 2008), pp. 752–763 (cit. on p. 126).
- [251] Andrew M. Childs, Eddie Schoute, and Cem M. Unsal. “Circuit Transformations for Quantum Architectures”. In: *14th Conference on the Theory of Quantum Computation, Communication and Cryptography*. Vol. 135. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 3:1–3:24. arXiv: [1902.09102](https://arxiv.org/abs/1902.09102) (cit. on p. 126).
- [252] Robert Wille, Lukas Burgholzer, and Alwin Zulehner. “Mapping Quantum Circuits to IBM QX Architectures Using the Minimal Number of SWAP and H Operations”. In: *Proceedings of the 56th Annual Design Automation Conference 2019*. DAC ’19. Las Vegas, NV, USA: Association for Computing Machinery, 2019 (cit. on p. 126).
- [253] Bochen Tan and Jason Cong. “Optimal layout synthesis for quantum computing”. In: *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*. IEEE. 2020, pp. 1–9. arXiv: [2007.15671](https://arxiv.org/abs/2007.15671) (cit. on p. 126).
- [254] Blake Gerard and Martin Kong. *String Abstractions for Qubit Mapping*. 2021. arXiv: [2111.03716](https://arxiv.org/abs/2111.03716) [[cs.ET](#)] (cit. on p. 126).
- [255] Xiangzhen Zhou, Sanjiang Li, and Yuan Feng. “Quantum Circuit Transformation Based on Simulated Annealing and Heuristic Search”. In: *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 39.12 (Dec. 2020), pp. 4683–4694. DOI: [10.1109/tcad.2020.2969647](https://doi.org/10.1109/tcad.2020.2969647) (cit. on p. 127).
- [256] Marcos Yukio Siraichi et al. “Qubit Allocation”. In: *Proceedings of the 2018 International Symposium on Code Generation and Optimization*. CGO 2018. Vienna, Austria: Association for Computing Machinery, 2018, pp. 113–125. DOI: [10.1145/3168822](https://doi.org/10.1145/3168822) (cit. on p. 127).
- [257] A Yu Kitaev. “Quantum computations: algorithms and error correction”. In: *Russ. Math. Surv.* 52.6 (Dec. 1997), pp. 1191–1249. DOI: [10.1070/rm1997v052n06abeh002155](https://doi.org/10.1070/rm1997v052n06abeh002155) (cit. on p. 127).
- [258] Ali Javadi-Abhari et al. *Quantum computing with Qiskit*. 2024. DOI: [10.48550/arXiv.2405.08810](https://doi.org/10.48550/arXiv.2405.08810). arXiv: [2405.08810](https://arxiv.org/abs/2405.08810) [[quant-ph](#)] (cit. on pp. 127, 134, 245, 246).

- [259] Cirq Developers. *Cirq*. Version v0.11.0. See full list of authors on Github: <https://github.com/quantumlib/Cirq/graphs/contributors>. May 2021. DOI: [10.5281/zenodo.4750446](https://doi.org/10.5281/zenodo.4750446) (cit. on pp. 127, 134).
- [260] Michael A Nielsen and Isaac Chuang. *Quantum computation and quantum information*. 2002 (cit. on p. 128).
- [261] Marcus P. da Silva, Olivier Landon-Cardinal, and David Poulin. “Practical Characterization of Quantum Devices without Tomography”. In: *Phys. Rev. Lett.* 107 (21 Nov. 2011), p. 210404. DOI: [10.1103/PhysRevLett.107.210404](https://doi.org/10.1103/PhysRevLett.107.210404) (cit. on p. 129).
- [262] Steven T. Flammia and Yi-Kai Liu. “Direct Fidelity Estimation from Few Pauli Measurements”. In: *Phys. Rev. Lett.* 106 (23 June 2011), p. 230501. DOI: [10.1103/PhysRevLett.106.230501](https://doi.org/10.1103/PhysRevLett.106.230501) (cit. on p. 129).
- [263] Hsin-Yuan Huang, Richard Kueng, and John Preskill. “Predicting many properties of a quantum system from very few measurements”. In: *Nat. Phys.* 16.10 (2020), pp. 1050–1057 (cit. on p. 129).
- [264] Arseni Wisniacki. *Loschmidt echo*. 2012. DOI: [10.4249/scholarpedia.11687](https://doi.org/10.4249/scholarpedia.11687). arXiv: [1206.6348](https://arxiv.org/abs/1206.6348) (cit. on p. 129).
- [265] Frank Arute et al. “Quantum supremacy using a programmable superconducting processor”. In: *Nature* 574.7779 (2019), pp. 505–510 (cit. on pp. 131, 132, 240).
- [266] Prakash Murali et al. “Noise-adaptive compiler mappings for noisy intermediate-scale quantum computers”. In: *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. 2019, pp. 1015–1029. arXiv: [1901.11054](https://arxiv.org/abs/1901.11054) (cit. on p. 131).
- [267] Siyuan Niu et al. “A Hardware-Aware Heuristic for the Qubit Mapping Problem in the NISQ Era”. In: *IEEE Transactions on Quantum Engineering* 1 (2020), pp. 1–14. DOI: [10.1109/tqe.2020.3026544](https://doi.org/10.1109/tqe.2020.3026544) (cit. on p. 131).
- [268] Prakash Murali et al. “Full-Stack, Real-System Quantum Computer Studies: Architectural Comparisons and Design Insights”. In: *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*. 2019, pp. 527–540 (cit. on p. 131).
- [269] Kenneth Rudinger et al. “Probing Context-Dependent Errors in Quantum Processors”. In: *Phys. Rev. X* 9 (2 June 2019), p. 021045. DOI: [10.1103/PhysRevX.9.021045](https://doi.org/10.1103/PhysRevX.9.021045) (cit. on p. 131).

- [270] Easwar Magesan, J. M. Gambetta, and Joseph Emerson. “Scalable and Robust Randomized Benchmarking of Quantum Processes”. In: *Phys. Rev. Lett.* 106.18 (May 2011) (cit. on pp. 132, 234).
- [271] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson. “Characterizing quantum gates via randomized benchmarking”. In: *Phys. Rev. A* 85.4 (Apr. 2012) (cit. on p. 132).
- [272] Yunchao Liu et al. *Benchmarking near-term quantum computers via random circuit sampling*. 2021. arXiv: 2105.05232 [quant-ph] (cit. on p. 132).
- [273] Dimitris Bertsimas, John Tsitsiklis, et al. “Simulated annealing”. In: *Statist. Sci.* 8.1 (1993), pp. 10–15 (cit. on pp. 132, 244).
- [274] Bruce Hajek. “Cooling schedules for optimal annealing”. In: *Math. Oper. Res.* 13.2 (1988), pp. 311–329 (cit. on p. 133).
- [275] Adam Holmes et al. “Impact of qubit connectivity on quantum algorithm performance”. In: *Quantum Sci. Technol.* 5.2 (Mar. 2020), p. 025009 (cit. on p. 134).
- [276] Evan Peters and Prasanth Shyamsundar. *Qubit assignment using time reversal*. <https://github.com/peterse/qubit-assignment-time-reversal>. 2022 (cit. on p. 134).
- [277] Otfried Gühne et al. “Toolbox for entanglement detection and fidelity estimation”. In: *Phys. Rev. A* 76.3 (Sept. 2007) (cit. on pp. 136, 252).
- [278] Google Quantum AI. *Calibration metrics*. <https://quantumai.google/cirq/google/calibration>. 2021 (cit. on p. 137).
- [279] Maurice G Kendall. “A new measure of rank correlation”. In: *Biometrika* 30.1/2 (1938), pp. 81–93 (cit. on p. 137).
- [280] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Boca Raton, Florida: CRC press, 1994 (cit. on p. 138).
- [281] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring Network Structure, Dynamics, and Function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA, 2008, pp. 11–15 (cit. on p. 138).
- [282] Robin Harper, Steven T. Flammia, and Joel J. Wallman. “Efficient learning of quantum noise”. In: *Nature Physics* 16.12 (Aug. 2020), pp. 1184–1188. DOI: 10.1038/s41567-020-0992-8 (cit. on p. 139).
- [283] Steven T. Flammia. *Averaged circuit eigenvalue sampling*. 2021. DOI: 10.48550/ARXIV.2108.05803 (cit. on p. 139).

- [284] C. Neill et al. “Accurately computing the electronic properties of a quantum ring”. In: *Nature* 594.7864 (June 2021), pp. 508–512. DOI: [10.1038/s41586-021-03576-2](https://doi.org/10.1038/s41586-021-03576-2) (cit. on p. 140).
- [285] Abhinav Kandala et al. “Error mitigation extends the computational reach of a noisy quantum processor”. In: *Nature* 567.7749 (Mar. 2019), pp. 491–495. DOI: [10.1038/s41586-019-1040-7](https://doi.org/10.1038/s41586-019-1040-7) (cit. on p. 140).
- [286] Chao Song et al. “Quantum computation with universal error mitigation on a superconducting quantum processor”. In: *Science Advances* 5.9 (2019), eaaw5686. DOI: [10.1126/sciadv.aaw5686](https://doi.org/10.1126/sciadv.aaw5686). eprint: <https://www.science.org/doi/pdf/10.1126/sciadv.aaw5686> (cit. on p. 140).
- [287] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. “General teleportation channel, singlet fraction, and quasidistillation”. In: *Phys. Rev. A* 60 (3 Sept. 1999), pp. 1888–1898 (cit. on p. 140).
- [288] Michael A Nielsen. “A simple formula for the average gate fidelity of a quantum dynamical operation”. In: *Phys. Lett. A* 303.4 (2002), pp. 249–252 (cit. on p. 140).
- [289] E. T. Jaynes. “Information Theory and Statistical Mechanics”. In: *Phys. Rev.* 106 (4 May 1957), pp. 620–630. DOI: [10.1103/PhysRev.106.620](https://doi.org/10.1103/PhysRev.106.620) (cit. on p. 146).
- [290] A. Tsigler and P. L. Bartlett. *Benign overfitting in ridge regression*. 2020. arXiv: [2009.14286](https://arxiv.org/abs/2009.14286) [math.ST] (cit. on pp. 185, 211).
- [291] Vladimir Koltchinskii and Karim Lounici. “Concentration inequalities and moment bounds for sample covariance operators”. In: *Bernoulli* 23.1 (2017), pp. 110–133. DOI: [10.3150/15-BEJ730](https://doi.org/10.3150/15-BEJ730) (cit. on p. 191).
- [292] Daniel A. Roberts and Beni Yoshida. “Chaos and complexity by design”. In: *J. High Energy Phys.* 2017 (Apr. 2017), p. 121. DOI: [10.1007/jhep04\(2017\)121](https://doi.org/10.1007/jhep04(2017)121) (cit. on p. 203).
- [293] Wallace C. Babcock. “Intermodulation interference in radio systems frequency of occurrence and control by channel selection”. In: *Bell Syst. tech. j.* 32.1 (1953), pp. 63–73. DOI: [10.1002/j.1538-7305.1953.tb01422.x](https://doi.org/10.1002/j.1538-7305.1953.tb01422.x) (cit. on p. 208).
- [294] M. Atkinson, N. Santoro, and J. Urrutia. “Integer Sets with Distinct Sums and Differences and Carrier Frequency Assignments for Nonlinear Repeaters”. In: *IEEE Trans. Commun.* 34.6 (1986), pp. 614–617. DOI: [10.1109/TCOM.1986.1096587](https://doi.org/10.1109/TCOM.1986.1096587) (cit. on p. 208).

- [295] J. Robinson and A. Bernstein. “A class of binary recurrent codes with limited error propagation”. In: *IEEE Trans. Inf.* 13.1 (1967), pp. 106–113. DOI: [10.1109/TIT.1967.1053951](#) (cit. on p. 208).
- [296] R. J. F. Fang and W. A. Sandrin. “Carrier frequency assignment for nonlinear repeaters”. In: *COMSAT Technical Review* 7 (May 1977), pp. 227–245 (cit. on p. 208).
- [297] Helmut Lutkepohl. *Handbook of matrices*. Vol. 2. Wiley, 1997, p. 167 (cit. on p. 219).
- [298] Asuman Ozdaglar, Devavrat Shah, and Christina Lee Yu. *Asynchronous Approximation of a Single Component of the Solution to a Linear System*. 2019. arXiv: [1411.2647 \[cs.DS\]](#) (cit. on p. 223).
- [299] Joel J. Wallman and Joseph Emerson. “Noise tailoring for scalable quantum computation via randomized compiling”. In: *Phys. Rev. A* 94.5 (Nov. 2016) (cit. on pp. 236, 237).
- [300] Yuxuan Du et al. “Learnability of Quantum Neural Networks”. In: *PRX Quantum* 2 (4 Nov. 2021), p. 040337 (cit. on p. 239).
- [301] Xiao Mi et al. “Information scrambling in quantum circuits”. In: *Science* 374.6574 (2021), pp. 1479–1483 (cit. on p. 241).
- [302] Lorenza Viola, Emanuel Knill, and Seth Lloyd. “Dynamical decoupling of open quantum systems”. In: *Phys. Rev. Lett.* 82.12 (1999), p. 2417 (cit. on p. 244).
- [303] E. L. Hahn. “Spin Echoes”. In: *Phys. Rev.* 80 (4 Nov. 1950), pp. 580–594. DOI: [10.1103/PhysRev.80.580](#) (cit. on p. 244).
- [304] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. “Error Mitigation for Short-Depth Quantum Circuits”. In: *Phys. Rev. Lett.* 119.18 (Nov. 2017) (cit. on p. 244).
- [305] Ying Li and Simon C. Benjamin. “Efficient Variational Quantum Simulator Incorporating Active Error Minimization”. In: *Phys. Rev. X* 7.2 (June 2017) (cit. on p. 244).
- [306] Adi Botea, Akihiro Kishimoto, and Radu Marinescu. “On the complexity of quantum circuit compilation”. In: *Eleventh annual symposium on combinatorial search*. Vol. 9. 2018, pp. 138–142 (cit. on p. 245).

APPENDICES

Appendix A

Supplementary material to Chapter 3

A.1 Solution for classical overparameterized model

In this section we derive the optimal solution and generalization error for the classical overparameterized weighted Fourier functions model of Eq. 3.21. We then discuss the conditions under which benign overfitting may be observed and construct examples of the phenomenon.

A.1.1 Linearization of overparameterized Fourier model

We first show that the classical overparameterized Fourier features model may be cast as a *linear* model under an appropriate orthogonal transformation. We are interested in learning a target function of the form

$$g(x) = \sum_{k \in \Omega_{n_0}} \hat{g}_k e^{i2\pi kx}, \quad (\text{A.1})$$

with the additional constraint that the Fourier coefficients satisfy $\hat{g}_k = \hat{g}_{-k}^*$ such that g is real. The spectrum of Eq. (A.1) for odd n_0 only contains integer frequencies,

$$\Omega_{n_0} = \left\{ -\frac{n_0-1}{2}, \dots, 0, \dots, \frac{n_0-1}{2} \right\}, \quad (\text{A.2})$$

and we accordingly call g a *bandlimited* function with bandlimit $n_0/2 - 1$. To learn g , we first sample n equally spaced datapoints on the interval $[0, 1]$,

$$x_j = \frac{j}{n}, \quad j \in [n], \quad (\text{A.3})$$

where $[n] = \{0, 1, 2, \dots, n-1\}$ and we assume n is odd, and we then evaluate $g(x_j)$ with additive error. This noisy sampling process yields n observations of the form $y_j = g(x_j) + \epsilon$ with $\mathbb{E}[\epsilon^2] = \sigma^2$. We will fit the observations y_j using overparameterized Fourier features models of the form

$$f(x) = \langle \phi(x), \boldsymbol{\alpha} \rangle = \sum_{k \in \Omega_d} \alpha_k \sqrt{\nu_k} e^{i2\pi kx}, \quad (\text{A.4})$$

with $\boldsymbol{\alpha} \in \mathbb{C}^d$, and we have introduced weighted Fourier features $\phi : \mathbb{R} \rightarrow \mathbb{C}^d$ defined elementwise as

$$\phi(x)_k = \sqrt{\nu_k} e^{i2\pi kx}. \quad (\text{A.5})$$

In Eq. (A.4), Ω_d describes the set of frequencies available to the model for any choice of $d \geq n \geq n_0$. We are interested in the case where f interpolates the observations y_j , i.e., $f(x_j) = y_j$ for all $j = 0, \dots, n-1$. To this end, we define a $n \times d$ feature matrix Φ whose rows are given by $\phi(x_j)^\dagger$:

$$\Phi_{jk} = [\phi(x_j)]_k^* = \sqrt{\nu_k} e^{-i2\pi jk/n}. \quad (\text{A.6})$$

The interpolation condition may then be stated in matrix form as

$$\Phi \boldsymbol{\alpha} = \mathbf{y}, \quad (\text{A.7})$$

where $(\mathbf{y})_j = y_j$ is the vector of noisy observations. Ω_d contains alias frequencies of Ω_{n_0} , and so there the choice of $\boldsymbol{\alpha}$ that satisfies Eq. (A.7) is not unique. Here we will focus on the minimum- ℓ_2 norm interpolating solution,

$$\boldsymbol{\alpha}^{opt} = \arg \min_{\Phi \boldsymbol{\alpha} = \mathbf{y}} \|\boldsymbol{\alpha}\|_2. \quad (\text{A.8})$$

A.1.1.1 Fourier transform into the linear model

We will now show that Eq. (A.8) with uniformly-spaced datapoints x_j can be solved using methods from ordinary linear regression under a suitable choice of transformation. Defining

the n -th root of identity as $\omega = e^{i2\pi/n}$, then the j -th row of the LHS of Eq. (A.7) is equivalent to

$$\sum_{k \in \Omega_d} \alpha_k \sqrt{\nu_k} \omega^{-kj} = \sum_{k \in \Omega_n} \sum_{\ell \in S(k)} \alpha_{k+n\ell} \sqrt{\nu_{k+n\ell}} \omega^{-(k+n\ell)j} \quad (\text{A.9})$$

$$= \sum_{k \in \Omega_n} \omega^{-kj} \sum_{\ell \in S(k)} \alpha_{k+n\ell} \sqrt{\nu_{k+n\ell}} \quad (\text{A.10})$$

$$:= \sum_{k \in \Omega_n} \omega^{-kj} \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle. \quad (\text{A.11})$$

where $S(k) = \{j : j \in \Omega_d, j \bmod n = k\}$ is the set of alias frequencies of k appearing in Ω_d , i.e. the set of frequencies $k + \ell n$ with $\ell \in \mathbb{Z}$ that obey

$$e^{i2\pi k x_j} = e^{i2\pi k j/n} = e^{i2\pi(k+\ell n)j/n} = e^{i2\pi(k+\ell n)x_j}, \quad (\text{A.12})$$

which follows from $e^{i2\pi \ell k} = 1$ since $k \in \mathbb{Z}$. The operator $P_{S(k)} : \mathbb{C}^d \rightarrow \mathbb{C}^d$ is the projector onto the set of standard basis vectors in $\mathbb{C}^d$ with indices in $S(k)$, and $(\mathbf{w})_k := \sqrt{\nu_k}$. The roots of unity are orthonormal since

$$\sum_{k=0}^{n-1} \omega^{k(p-q)} = n \delta_p^q \quad (\text{A.13})$$

for $p, q \in [n]$. This implies

$$\sum_{k \in \Omega_n} \omega^{k(p-q)} = \omega^{\min(\Omega_n)(p-q)} n \delta_p^q = n \delta_p^q, \quad (\text{A.14})$$

for $p, q \in \Omega_n$. Defining the discrete Fourier Transform $\hat{y}_k$ of $\mathbf{y}$ according to

$$\hat{y}_j = \frac{1}{n} \sum_{k \in \Omega_n} y_k \omega^{jk}, \quad (\text{A.15})$$

then using the identity of Eq. (A.14), we evaluate the j -th row of Eq. (A.7) as

$$\sum_{k \in \Omega_n} \omega^{-kj} \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle = y_j, \quad (\text{A.16})$$

$$\sum_{j \in \Omega_n} \omega^{pj} \sum_{k \in \Omega_n} \omega^{-kj} \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle = \sum_{j \in \Omega_n} \omega^{pj} \left(\sum_{k \in \Omega_n} \hat{y}_k \omega^{-kj} \right), \quad (\text{A.17})$$

$$\sum_{k \in \Omega_n} n \delta_k^p \langle P_{S(k)} \boldsymbol{\alpha}, P_{S(k)} \mathbf{w} \rangle = \sum_{k \in \Omega_n} n \delta_k^p \hat{y}_k, \quad (\text{A.18})$$

$$\langle P_{S(p)} \boldsymbol{\alpha}, P_{S(p)} \mathbf{w} \rangle = \hat{y}_p. \quad (\text{A.19})$$

Inspecting this final line yields a new matrix equation: Let X be an $n \times d$ matrix X with elements

$$X_{jk} = (P_{S(j)} \mathbf{w})_k = \sqrt{\nu_k} \mathbb{1}\{k \in S(j)\}, \quad (\text{A.20})$$

where the conditional operator $\mathbb{1}\{Z\}$ evaluates to 1 if the predicate Z is true, and 0 otherwise. Then we may express Eq. (A.19) for all $p \in \Omega_n$ as a matrix equation

$$X\boldsymbol{\alpha} = \hat{\mathbf{y}}, \quad (\text{A.21})$$

where $(\hat{\mathbf{y}})_j = \hat{y}_j$. We have shown that Eq. (A.7) is exactly equivalent to Eq. (A.21) for uniformly spaced inputs, and as $\boldsymbol{\alpha}$ is unchanged between these two representations this implies that the solution to Eq. (A.8) is also given by

$$\boldsymbol{\alpha}^{opt} = \arg \min_{X\boldsymbol{\alpha} = \hat{\mathbf{y}}} \|\boldsymbol{\alpha}\|_2. \quad (\text{A.22})$$

Therefore, the minimum ℓ_2 -norm solution to interpolate the input signal using weighted Fourier functions provided as samples y_j is exactly the same as the minimum ℓ_2 -norm solution for an equivalent *linear regression* problem on the matrix X with targets $\hat{\mathbf{y}}$. Furthermore, this linear regression problem is related to the original problem via Fourier transform. Let F be the (nonunitary) discrete Fourier transform defined on $\mathbb{C}^n$ elementwise as

$$F_{jk} = \omega^{jk}, \quad (\text{A.23})$$

$$F^\dagger F = n\mathbb{I}_n. \quad (\text{A.24})$$

Then for $k \in \Omega_d$, $j \in \Omega_n$,

$$\frac{1}{n}(F\Phi)_{jk} = \frac{1}{n} \sum_{\ell \in \Omega_n} \omega^{(k-j)\ell} \sqrt{\nu_k} = \frac{1}{n} (n\delta_j^{k \bmod n}) \sqrt{\nu_k} = \sqrt{\nu_k} \mathbb{1}\{k \in S(j)\} = X_{jk}, \quad (\text{A.25})$$

implying

$$\frac{1}{n}F\Phi = X. \quad (\text{A.26})$$

We may similarly recover $\frac{1}{n}F\mathbf{y} = \hat{\mathbf{y}}$ to show that the coefficients $\hat{y}_k$ are given by a discrete Fourier transform of y_j .

A.1.1.2 Error analysis of the linear model

Having shown that the discrete Fourier transform F relates the original system of trigonometric equations $\Phi\boldsymbol{\alpha} = \mathbf{y}$ to the system of linear equations (or ordinary least squares problem $X\boldsymbol{\alpha} = \hat{\mathbf{y}}$, where we treat the rows of X as observations in $\mathbb{R}^d$), we now derive the error of the problem in the Fourier representation. Standard treatment for ordinary least squares (OLS) gives the minimum ℓ_2 -norm solution to Eq. (A.22) as

$$\boldsymbol{\alpha}^{opt} = X^T(XX^T)^{-1}\hat{\mathbf{y}} \quad (\text{A.27})$$

in which case the optimal interpolating overparameterized Fourier model is

$$f^{opt}(x) = \langle \phi(x), \boldsymbol{\alpha}^{opt} \rangle \quad (\text{A.28})$$

$$= \sum_{k \in \Omega_n} \frac{\hat{y}_k}{\sum_{j \in S(k)} \nu_j} \sum_{\ell \in S(k)} \nu_\ell e^{i2\pi \ell x}. \quad (\text{A.29})$$

Once we have trained a model on noisy samples $\mathbf{y}$ of g using uniformly spaced values of x , we would like to evaluate how well the model performs for arbitrary x in $[0, 1]$. Given some function $f = \langle \phi(x), \boldsymbol{\alpha} \rangle$ the mean squared error $\mathbb{E}_x(f(x) - g(x))^2$ may be evaluated with respect to the interval $[0, 1]$. We define the *generalization error* of the model as the expected mean squared error evaluated with respect to $\mathbf{y}$:

$$L(f) := \mathbb{E}_{x, \mathbf{y}}(f(x) - g(x))^2. \quad (\text{A.30})$$

We decompose the generalization error of the model as

$$\mathbb{E}_{x, \mathbf{y}}(f(x) - g(x))^2 = \mathbb{E}_{x, \mathbf{y}}(f(x) - \mathbb{E}_{\mathbf{y}}f(x) + \mathbb{E}_{\mathbf{y}}f(x) - g(x))^2 \quad (\text{A.31})$$

$$= \mathbb{E}_{x, \mathbf{y}}(f(x) - \mathbb{E}_{\mathbf{y}}f(x))^2 + \mathbb{E}_{x, \mathbf{y}}(\mathbb{E}_{\mathbf{y}}f(x) - g(x))^2 \quad (\text{A.32})$$

$$+ 2\mathbb{E}_{x, \mathbf{y}}[(f(x) - \mathbb{E}_{\mathbf{y}}f(x))(\mathbb{E}_{\mathbf{y}}f(x) - g(x))].$$

Because of orthonormality of Fourier basis functions, the cross-terms cancel resulting in a decomposition of the generalization error of the optimal standard bias and variance terms:

$$L(f^{opt}) = \underbrace{\mathbb{E}_{x, \mathbf{y}}(f^{opt}(x) - \mathbb{E}_{\mathbf{y}}f^{opt}(x))^2}_{\text{VAR}} + \underbrace{\mathbb{E}_{x, \mathbf{y}}(\mathbb{E}_{\mathbf{y}}f^{opt}(x) - g(x))^2}_{\text{BIAS}^2}. \quad (\text{A.33})$$

We now evaluate VAR and BIAS² using the linear representation developed in the previous section. Beginning with the variance, conditional on constructing Φ from the set of

uniformly spaced points x_j we apply the discrete Fourier transformation to yield X and compute $\boldsymbol{\alpha}^{opt}$ using Eq. (A.27):

$$\text{VAR} = \mathbb{E}_{x, \mathbf{y}} (f^{opt}(x) - \mathbb{E}_{\mathbf{y}} f^{opt}(x))^2 \quad (\text{A.34})$$

$$= \mathbb{E}_{x, \mathbf{y}} |\langle \phi(x), \boldsymbol{\alpha}^{opt} - \mathbb{E}_{\mathbf{y}} \boldsymbol{\alpha}^{opt} \rangle|^2 \quad (\text{A.35})$$

$$= \mathbb{E}_{x, \mathbf{y}} |\langle \phi(x), X^T (X X^T)^{-1} (\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}} \hat{\mathbf{y}}) \rangle|^2 \quad (\text{A.36})$$

$$= \text{Tr} \left(\mathbb{E}_{\mathbf{y}} [(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}} \hat{\mathbf{y}})(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}} \hat{\mathbf{y}})^\dagger] (X X^T)^{-1} X \mathbb{E}_x [\phi(x) \phi(x)^\dagger] X^T (X X^T)^{-1} \right) \quad (\text{A.37})$$

$$= \frac{\sigma^2}{n} \text{Tr} \left((X X^T)^{-2} X \Sigma_\phi X^T \right). \quad (\text{A.38})$$

Letting $\boldsymbol{\epsilon} := (\mathbf{y} - \mathbb{E}_{\mathbf{y}} \mathbf{y})$ we have simplified the above using the following¹:

$$\mathbb{E}_{\mathbf{y}} [(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}} \hat{\mathbf{y}})(\hat{\mathbf{y}} - \mathbb{E}_{\mathbf{y}} \hat{\mathbf{y}})^\dagger]_{jk} = \frac{1}{n^2} \mathbb{E}_{\mathbf{y}} (F \boldsymbol{\epsilon})_j (F \boldsymbol{\epsilon})_k^* \quad (\text{A.39})$$

$$= \frac{1}{n^2} \mathbb{E}_{\mathbf{y}} \left(\sum_{p \in \Omega_n} \omega^{-jp} \epsilon_p \right) \left(\sum_{q \in \Omega_n} \omega^{qk} \epsilon_q \right) \quad (\text{A.40})$$

$$= \frac{1}{n^2} \sum_{p, q \in \Omega_n} \mathbb{E}_{\mathbf{y}} [\epsilon_p \epsilon_q] \omega^{-jp+kq} \quad (\text{A.41})$$

$$= \frac{\sigma^2}{n} \delta_j^k, \quad (\text{A.42})$$

where we have used the fact that the errors ϵ are independent and zero mean, $\mathbb{E}_{\mathbf{y}} [\epsilon_p \epsilon_q] = \mathbb{E}_{\mathbf{y}} [\epsilon_p^2] \delta_p^q$. We have defined the feature covariance matrix as $\Sigma_\phi = \mathbb{E}_x [\phi(x) \phi(x)^\dagger]$, which may be computed elementwise using the orthonormality of Fourier features on $[0, 1]$:

$$(\Sigma_\phi)_{jk} = \mathbb{E}_x [\phi(x)_j \phi(x)_k^*] = \sqrt{\nu_j \nu_k} \langle e^{i2\pi jx}, e^{i2\pi kx} \rangle_{L_2^{[0,1]}} = \delta_j^k \nu_j. \quad (\text{A.43})$$

¹In Eq. (A.38), $\mathbb{E}[X^T X] = \Sigma_\phi$, so that Σ_ϕ differs from the covariance matrix for the rows of X by a factor of $1/n$.

The following may be computed directly:

$$(XX^T)_{jk} = \delta_j^k \sum_{\ell \in S(j)} \nu_\ell \quad (\text{A.44})$$

$$(X\Sigma_x X^T)_{jk} = \sum_{\ell \in \Omega_d} X_{j\ell} \left(\frac{\nu_\ell}{n} \right) X_{k\ell} \quad (\text{A.45})$$

$$= \frac{1}{n} \sum_{\ell \in \Omega_d} \nu_\ell^2 \mathbb{1}\{\ell \in S(j)\} \mathbb{1}\{\ell \in S(k)\} \quad (\text{A.46})$$

$$= \frac{\delta_j^k}{n} \sum_{\ell \in S(k)} \nu_\ell^2. \quad (\text{A.47})$$

In line (A.47) we have used the identity

$$\mathbb{1}\{\ell \in S(j)\} \mathbb{1}\{\ell \in S(k)\} = \mathbb{1}\{\ell \in S(j)\} \delta_k^j, \quad (\text{A.48})$$

since $\ell \in S(k) \Rightarrow k = \ell \bmod n$ and therefore $\ell \in S(j) \Rightarrow j = \ell \bmod n = k$. We now compute the variance as

$$\text{VAR} = \frac{\sigma^2}{n} \text{Tr} \left((XX^T)^{-2} X\Sigma_\phi X^T \right) \quad (\text{A.49})$$

$$= \frac{\sigma^2}{n} \sum_{j,k \in \Omega_n} \left(\delta_k^j \sum_{\ell \in S(k)} \nu_\ell \right)^{-2} \left(\delta_j^k \sum_{\ell \in S(k)} \nu_\ell^2 \right) \quad (\text{A.50})$$

$$= \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell \right)^2}. \quad (\text{A.51})$$

To evaluate the BIAS^2 , we will first rewrite $g(x)$ in terms of its Fourier representation,

$$g(x) = \langle \phi(x), \boldsymbol{\alpha}_0 \rangle, \quad (\text{A.52})$$

$$(\boldsymbol{\alpha}_0)_k = \frac{\hat{g}_k}{\sqrt{\nu_k}} \mathbb{1}\{k \in \Omega_{n_0}\}. \quad (\text{A.53})$$

Noting that $(X\alpha_0)_k = \hat{g}_k$ implies $X\alpha_0 = \mathbb{E}_{\mathbf{y}}\hat{\mathbf{y}}$ we evaluate

$$\text{BIAS}^2 = \mathbb{E}_{x,\mathbf{y}}(\mathbb{E}_{\mathbf{y}}f^{opt}(x) - g(x))^2 \quad (\text{A.54})$$

$$= \mathbb{E}_{x,\mathbf{y}}|\langle \phi(x), \mathbb{E}_{\mathbf{y}}\alpha^{opt} - \alpha_0 \rangle|^2 \quad (\text{A.55})$$

$$= \text{Tr} \left(\mathbb{E}_x [\phi(x)\phi(x)^\dagger] (\mathbb{E}_{\mathbf{y}}\alpha^{opt} - \alpha_0)(\mathbb{E}_{\mathbf{y}}\alpha^{opt} - \alpha_0)^\dagger \right) \quad (\text{A.56})$$

$$= \text{Tr} \left(\Sigma_\phi (X^T (X X^T)^{-1} X - \mathbb{I}_d) \alpha_0 \alpha_0^\dagger (X^T (X X^T)^{-1} X - \mathbb{I}_d)^\dagger \right) \quad (\text{A.57})$$

$$= \|\Sigma_\phi^{1/2}(\mathbb{I}_d - X^T (X X^T)^{-1} X) \alpha_0\|_2^2. \quad (\text{A.58})$$

While Eq. (A.58) already completely characterizes the bias error in terms of the choice of feature weights and input data, it may be greatly simplified by taking advantage of the sparseness X . We have

$$(X^T (X X^T)^{-1} X)_{jk} = \sum_{\ell \in \Omega_n} \left(\sum_{m \in S(\ell)} \nu_m \right)^{-1} X_{\ell j} X_{\ell k} \quad (\text{A.59})$$

$$= \left(\sum_{m \in S(j \bmod n)} \nu_m \right)^{-1} \sqrt{\nu_k} \sqrt{\nu_j} \mathbb{1}\{k \in S(j \bmod n)\}, \quad (\text{A.60})$$

where in line (A.60) we have used

$$\sum_{\ell \in \Omega_n} \mathbb{1}\{k \in S(\ell)\} \mathbb{1}\{j \in S(\ell)\} = \mathbb{1}\{k \in S(j \bmod n)\} \quad (\text{A.61})$$

for $k, j \in \Omega_d$ and $\ell \in \Omega_n$, which follows from similar reasoning as Eq. (A.47). And so, writing

$$(\Sigma_\phi^{1/2}(\mathbb{I}_d - X^T(XX^T)^{-1}X)\alpha_0)_\ell \quad (\text{A.62})$$

$$= \sum_{j \in \Omega_d} \left[\delta_\ell^j \hat{g}_j \mathbb{1}\{j \in \Omega_{n_0}\} - \sum_{k \in \Omega_d} \sqrt{\nu_\ell} \delta_j^\ell \left(\sum_{m \in S(j \bmod n)} \nu_m \right)^{-1} \hat{g}_k \sqrt{\nu_j} \mathbb{1}\{k \in S(j \bmod n)\} \mathbb{1}\{k \in \Omega_{n_0}\} \right] \quad (\text{A.63})$$

$$= \left(\sum_{j \in \Omega_{n_0}} \delta_\ell^j \hat{g}_j - \sum_{k \in \Omega_{n_0}} \nu_\ell \left(\sum_{m \in S(\ell \bmod n)} \nu_m \right)^{-1} \hat{g}_\ell \mathbb{1}\{k \in S(\ell \bmod n)\} \right) \quad (\text{A.64})$$

$$= \left(\mathbb{1}\{\ell \in \Omega_{n_0}\} \hat{g}_\ell - \sum_{k \in \Omega_{n_0}} \nu_\ell \left(\sum_{m \in S(k)} \nu_m \right)^{-1} \hat{g}_k \mathbb{1}\{\ell \in S(k)\} \right). \quad (\text{A.65})$$

Letting $Q_k := \sum_{m \in S(k)} \nu_m$ the bias term of the error evaluates to

$$\text{BIAS}^2 = \sum_{\ell \in \Omega_d} |(\Sigma_x^{1/2}(\mathbb{I}_d - X^T(XX^T)^{-1}X)\alpha_0)_\ell|^2 \quad (\text{A.66})$$

$$= \sum_{\ell \in \Omega_d} \left| \mathbb{1}\{\ell \in \Omega_{n_0}\} \hat{g}_\ell - \sum_{k \in \Omega_{n_0}} \nu_\ell Q_k^{-1} \hat{g}_k \mathbb{1}\{\ell \in S(k)\} \right|^2 \quad (\text{A.67})$$

$$= \sum_{\ell \in \Omega_d} \left(\mathbb{1}\{\ell \in \Omega_{n_0}\} |\hat{g}_\ell|^2 - \sum_{k \in \Omega_{n_0}} \mathbb{1}\{\ell \in \Omega_{n_0}\} \hat{g}_\ell \hat{g}_k^* \nu_\ell Q_k^{-1} \mathbb{1}\{\ell \in S(k)\} \right. \\ \left. - \sum_{k \in \Omega_{n_0}} \mathbb{1}\{\ell \in \Omega_{n_0}\} \hat{g}_\ell^* \hat{g}_k \nu_\ell Q_k^{-1} \mathbb{1}\{\ell \in S(k)\} \right. \\ \left. + \sum_{j, k \in \Omega_{n_0}} \nu_\ell^2 Q_j^{-1} Q_k^{-1} \hat{g}_j^* \hat{g}_k \mathbb{1}\{\ell \in S(j)\} \mathbb{1}\{\ell \in S(k)\} \right) \quad (\text{A.68})$$

$$= \left(\sum_{k \in \Omega_n} |\hat{g}_k|^2 - 2 \sum_{k \in \Omega_{n_0}} \nu_k |\hat{g}_k|^2 Q_k^{-1} + \sum_{k \in \Omega_{n_0}} \sum_{\ell \in S(k)} \nu_\ell^2 Q_k^{-2} |\hat{g}_k|^2 \right) \quad (\text{A.69})$$

$$= \sum_{k \in \Omega_{n_0}} \frac{|\hat{g}_k|^2}{Q_k^2} \left((Q_k - \nu_k)^2 + \sum_{\ell \in S(k) \setminus k} \nu_\ell^2 \right) \quad (\text{A.70})$$

$$= \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \frac{\left(\sum_{\ell \in S(k) \setminus k} \nu_\ell \right)^2 + \sum_{\ell \in S(k) \setminus k} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell \right)^2}. \quad (\text{A.71})$$

With Eqs. (A.51) and (A.71) we have recovered a closed form expression for generalization error in terms of feature weights ν_k , the target function Fourier coefficients $\hat{g}_k$, and the noise variance σ^2 . This was possible in part because of the orthonormality of the Fourier features $\phi(x)_k$ and the choice to sample x_j uniformly on $[0, 1]$, resulting in diagonal Σ_ϕ and XX^T respectively. This simplicity is more advantageous for studying scenarios where benign overfitting may exist compared to prior works [68, 290]. We will now analyze choices of feature weights ν_k that may give rise to benign overfitting for overparameterized weighted Fourier Features models.

We remark that the results of this section may also be derived using the methods of Ref. [71], though we have opted here to use a language more reminiscent of linear regression

to highlight connections between the analysis of weighted Fourier features models and ordinary least squares errors.

A.1.2 Solutions for special cases

A.1.2.1 Noise-only minimum norm estimator

We now derive the cases considered in Sec. 3.3.2 of the main text. We first consider when the target function is given by $g = 0$ and we attempt to predict on a pure noise signal

$$y_j = \epsilon_j, \quad (\text{A.72})$$

with $\mathbb{E}[\epsilon^2] = \sigma^2$. We can recover an unaliased, “simple” fit to this pure noise signal by reconstructing ϵ from the DFT using Eq. (A.15):

$$\epsilon_j = \sum_{k \in \Omega_n} \hat{\epsilon}_k e^{i2\pi jk/n}. \quad (\text{A.73})$$

By setting all weights $\nu_k = 1$, an immediate choice for an interpolating f with access to frequencies in Ω_d is found by setting

$$\alpha_k = \frac{\hat{\epsilon}_{k \bmod n}}{n(m+1)}, \quad (\text{A.74})$$

where we have assumed $d = n(m+1)$. Eq. (A.74) is the minimum ℓ_2 -norm estimator and evenly distributes the weight of each Fourier coefficient $\hat{\epsilon}_k$ over m aliases frequencies in $S(k)$. The effect of higher-frequency aliases is to reduce the f to near-zero everywhere except at the interpolated training data. We can directly compute the generalization error of the interpolating estimator as

$$\mathbb{E}_{x,y} |f(x) - g(x)|^2 = \mathbb{E}_x |f(x)|^2 \quad (\text{A.75})$$

$$= \mathbb{E}_y \left| \sum_{k=0}^{d-1} \alpha_k e^{-i2\pi kx} \right|^2 \quad (\text{A.76})$$

$$= \mathbb{E}_y \sum_{k=0}^{d-1} \left| \frac{\hat{\epsilon}_{k \bmod n}}{n(m+1)} \right|^2 \quad (\text{A.77})$$

$$= \mathbb{E}_y \frac{n(m+1) \|\hat{\epsilon}\|_2^2}{n^2(m+1)^2} \quad (\text{A.78})$$

$$= \frac{\sigma^2}{n(m+1)}. \quad (\text{A.79})$$

Using $d = n(m+1)$ as the dimensionality of the feature space, we have recovered the lower bound scaling for overparameterized models derived in Ref. [71]. In line (A.76) we have used the independence of ϵ from x_k and y_k and in line (A.77) we have used orthonormality of Fourier basis functions on $[0, 1]$. line (A.79) uses Parseval's relation for the Fourier coefficients, namely:

$$\sum_{k \in \Omega_n} |\epsilon_k|^2 = n \sum_{k \in \Omega_n} |\hat{\epsilon}_k|^2, \quad (\text{A.80})$$

which implies $\mathbb{E}\|\hat{\epsilon}\|_2^2 = n^{-1}\mathbb{E}\|\epsilon\|_2^2 = \sigma^2$. Figure 3.2a of the main text shows the effect of the number of cohorts m on the behavior of f , which interpolates a pure noise signal with very little bias. As the number of cohorts increases, the function deviates very little away from the true “signal” $y = 0$, and becomes very “spiky” in the vicinity of noise

A.1.2.2 Signal-only minimum norm estimator

Now we study the opposite situation in which the pure tone is noiseless, and aliases in the spectrum of f interfere in prediction of f . In this case, we set $\sigma = 0$ and interpolate target labels

$$y_j = \hat{g}_p e^{i2\pi p x_j}, \quad (\text{A.81})$$

with $-n/2 < p < n/2$. When we set $d = n(m+1)$ and predict on y , there are exactly $|S(p)| - 1 = m$ aliases for the target function with frequency p . We again assume all feature weights are equal, $\nu_k = 1$, and by orthonormality of Fourier basis functions, only the components of f with frequency in $S(p)$ are retained:

$$f(x) = \alpha_p e^{i2\pi p x} + \sum_{k \in S(p)} \alpha_k e^{i2\pi k x}, \quad (\text{A.82})$$

which will interpolate the training points for any choice of α satisfying

$$\sum_{k \in S(p)} \alpha_k = \hat{g}_p. \quad (\text{A.83})$$

The choice of trainable weights α_k that satisfy Eq. (A.83) while minimizing ℓ_2 -norm is

$$\alpha_k = \begin{cases} \frac{\hat{g}_p}{m+1}, & k \in S(p), \\ 0, & \text{otherwise.} \end{cases} \quad (\text{A.84})$$

The problem with minimizing the ℓ_2 norm in this case is that it spreads the true signal into higher frequency aliases: The generalization error of this model is

$$\mathbb{E}_{x,y}|f(x) - g(x)|^2 = \left| \hat{g}_p \left(\frac{m}{m+1} \right) \right|^2 + m \frac{|\hat{g}_p|^2}{(m+1)^2} = \mathcal{O}(1). \quad (\text{A.85})$$

We see that this model generally fails to generalize. This poor generalization was described as “signal bleeding” by Ref. [71]: Using n samples, there is no way to distinguish the signal due to an alias of p from the signal due to the true frequency p , so the coefficients α_k become evenly distributed over aliases with very little weight allocated to the true Fourier coefficient $\hat{g}_p$ in the model f . Fig. 3.2b in the main text shows the effect of “signal bleed” for learning a pure tone in the absence of noise.

A.1.3 Conditions for Benign overfitting

The behavior of the error of the overparameterized Fourier model (Eq. (3.21) of the main text) depends on an interplay between noise variance σ^2 , signal Fourier coefficients $\hat{g}_k$, feature weights ν_k , and the size of the model d . A desirable property of the feature weights ν_k is that they should result in a model that both interpolates the sampled data while also achieving good generalization error in the limit $n \rightarrow \infty$. For our purposes we will consider cases where $\lim_{n \rightarrow \infty} L(f^{opt}) = 0$, though this condition could be relaxed to allow for more interesting or natural choices of ν_k . We now analyze the error arising from a simple weighting scheme to demonstrate benign overfitting using the overparameterized Fourier models discussed in this work.

A.1.3.1 A simple demonstration of benign overfitting

Here we demonstrate a simple example of benign overfitting when the feature weights ν_k are chosen with direct knowledge of the spectrum Ω_{n_0} of g . For some $n_0 < n < d$, fix $c \in (0, 1)$ and use the feature weights given as

$$\nu_k = \begin{cases} \frac{c}{n_0}, & k \in \Omega_{n_0}, \\ \frac{1-c}{d-n_0}, & \text{otherwise,} \end{cases} \quad (\text{A.86})$$

for all $k \in \Omega_d$. For simplicity, suppose $d = n(m+1)$ such that $|S(k)| = m+1$ for all $k \in \Omega_n$. Defining the *signal power* as

$$P := \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2, \quad (\text{A.87})$$

we can directly evaluate VAR of Eq. (A.51) and BIAS² of Eq. (A.71):

$$\text{VAR} \rightarrow \frac{\sigma^2}{n} \left(n_0 \frac{\left(\frac{c}{n_0}\right)^2 + m \left(\frac{1-c}{d-n_0}\right)^2}{\left(\frac{c}{n_0} + m \frac{1-c}{d-n_0}\right)^2} + \frac{n - n_0}{m + 1} \right). \quad (\text{A.88})$$

$$\text{BIAS}^2 \rightarrow \frac{P n_0^2 (m + 1) m}{\left(\frac{c}{1-c}(d - n_0) + m n_0\right)^2}. \quad (\text{A.89})$$

Fixing n_0 , we can bound generalization error in the asymptotic limit $n \rightarrow \infty$ as:

$$\text{VAR} = \mathcal{O} \left(\frac{1}{n} + \frac{n}{d} \right) \quad (\text{A.90})$$

$$\text{BIAS}^2 = \mathcal{O} \left(\frac{1}{n^2} \right). \quad (\text{A.91})$$

Therefore, by setting $d = \omega(n)$ the model perfectly interpolates the training data and also achieves vanishing generalization error in the limit of large number of data. A similar example was considered in Ref. [71], though a rigorous error analysis (and relationship to benign overfitting) was not considered there. This benign overfitting behavior is entirely due to the feature weights of Eq. (A.86): As $d, n \rightarrow \infty$ with $d = \omega(n)$, the feature weights are concentrated on Ω_{n_0} (suppressing BIAS) while becoming increasingly small and evenly distributed over all aliases of Ω_{n_0} (suppressing VAR).

A.1.3.2 More general conditions for benign overfitting

We have derived closed-form solutions for the BIAS² and VAR terms that determine the total generalization error of an interpolating model f and in the previous section provided a concrete example of a model that achieves benign overfitting in this setting. We now discuss conditions under which a more general choice of model can exhibit benign overfitting.

We begin by showing that the variance of Eq. (A.51) splits naturally into an error due to a (simple) prediction component and a (spiky, complex) interpolating component. Following Refs. [66, 68], we will split the variance into components corresponding to eigenspaces of Σ_ϕ with large and small eigenvalues respectively. Let $S^{\leq p}$ denote the set indices for the largest p eigenvalues of Σ_ϕ (i.e., the largest p values of ν_k), and $S^{> p} = [d] \setminus S^{\leq p}$ be its complement. Define $P^{\leq p} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ as the projector onto the subspace of $\mathbb{R}^d$ spanned by basis vectors labelled by indices in $S^{\leq p}$ (and $P^{> p} = \mathbb{I}_d - P^{\leq p}$ is defined analogously).

Then letting $(\boldsymbol{\nu})_k = \nu_k$ be the vector of feature weights and assuming $p \leq n$, we may rewrite the variance of Eq. (A.51) as

$$\text{VAR} = \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \quad (\text{A.92})$$

$$= \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\|P_{S(k)} \boldsymbol{\nu}\|_2^2}{\|P_{S(k)} \boldsymbol{\nu}\|_1^2} \quad (\text{A.93})$$

$$= \frac{\sigma^2}{n} \sum_{k \in \Omega_n} \frac{\|P^{\leq p} P_{S(k)} \boldsymbol{\nu} + P^{> p} P_{S(k)} \boldsymbol{\nu}\|_2^2}{\|P_{S(k)} \boldsymbol{\nu}\|_1^2} \quad (\text{A.94})$$

$$\leq \frac{\sigma^2}{n} \left(\sum_{k \in \Omega_n} \frac{\|P^{\leq p} P_{S(k)} \boldsymbol{\nu}\|_2^2}{\|P^{\leq p} P_{S(k)} \boldsymbol{\nu}\|_1^2} + \sum_{k \in \Omega_n} \frac{\|P^{> p} P_{S(k)} \boldsymbol{\nu}\|_2^2}{\|P^{> p} P_{S(k)} \boldsymbol{\nu}\|_1^2} \right) \quad (\text{A.95})$$

$$\leq \sigma^2 \left[\frac{p}{n} + \frac{1}{n} \sum_{k \in \Omega_n} \frac{1}{R_p^{(k)}(\Sigma_\phi)} \right], \quad (\text{A.96})$$

where we have used $P^{\leq p} P_{S(k)} \hat{e}_j = \delta_j^k \mathbb{1}\{j \leq p\}$ for $p \leq n$ and we have introduced an effective rank for the alias cohort of k ,

$$R_p^{(k)}(\Sigma_\phi) = \frac{\left(\sum_{\ell \in S(k) \cap S^{> p}} \nu_\ell\right)^2}{\sum_{\ell \in S(k) \cap S^{> p}} \nu_\ell^2}. \quad (\text{A.97})$$

Since $p = n_0$ is a relevant choice for our problem setup, we define $R^{(k)} := R_{n_0}^{(k)}(\Sigma_\phi)$ and focus on the bound

$$\text{VAR} \leq \sigma^2 \left[\frac{n_0}{n} + \frac{1}{n} \sum_{k \in \Omega_n} \frac{1}{R^{(k)}} \right], \quad (\text{A.98})$$

The first term of the decomposition of Eq. (A.98) corresponds to the variance of a (non-interpolating) model with access to n_0 Fourier modes, while the second term corresponds to excess error introduced by high frequency components of f . Given a sequence of experiments with increasing n (while g and σ^2 remain fixed), we would like to understand the choices of feature weights ν_k for which VAR vanishes as $n \rightarrow \infty$. Given that $|\{k \in \Omega_n\}| = n$, a sufficient condition for a sequence of feature weight distributions to be benign is that $R^{(k)} = \omega(1)$ for all k while a necessary condition is that there is no k for which $R^{(k)} = \mathcal{O}(1/n)$. These conditions are not difficult to satisfy: Intuitively they require

only that Σ_ϕ changes with increasing n in such a way that the values of ν_k in $\Omega_n \setminus \Omega_{n_0}$ continue “flatten out” as n increases. This is precisely the behavior engineered in the example of Sec. A.1.3.1.

We now proceed to bound the bias term of Eq. (A.58). Observe that $P^\perp := (\mathbb{I}_d - X^T(XX^T)^{-1}X)$ is a projector onto the subspace of $\mathbb{R}^d$ orthogonal to the rows of X , therefore satisfying $\|P^\perp\| \leq 1$ and $P^\perp X^T X = 0$. Then the BIAS^2 is bounded as

$$\text{BIAS}^2 = \|\Sigma_\phi^{1/2} P^\perp \alpha_0\|_2^2 \quad (\text{A.99})$$

$$= \|(\Sigma_\phi - X^T X)^{1/2} P^\perp \alpha_0\|_2^2 \quad (\text{A.100})$$

$$\leq \|\Sigma_\phi - X^T X\|_\infty \|\alpha_0\|_2^2. \quad (\text{A.101})$$

The term $X^T X$ can be interpreted as a finite-sample estimator for Σ_ϕ in the sense that $X^T X = n^{-1} \Phi^T \Phi = \Sigma_\phi$ for $n = d$. However, we cannot apply standard results on the convergence of sample covariance matrices (e.g. Ref. [291]) since the uniform spacing requirement for training data x violates the standard assumption of i.i.d. input data. To proceed, we will make a number of simplifying assumptions about the feature weights. First, to control for the possibility that large feature weights ν_k concentrate within a specific $S(k)$ we will assume that feature weights corresponding to any set of alias frequencies of Ω_n are close to their average. Letting $d = (m+1)n$, we define

$$\eta_\ell = \frac{1}{n} \sum_{k \in \Omega_n} \nu_{k+n\ell}, \quad (\text{A.102})$$

for $\ell = 0, 1, \dots, m$. We will impose that $|\nu_{k+n\ell} - \eta_\ell| \leq t$ for all $k \in \Omega_n$, $\ell = 1, \dots, m$, for some positive number t . We further assume normalization, $\sum_{k \in \Omega_d} \nu_k = 1 = n \sum_{\ell=0}^m \eta_\ell$. Under these assumptions we can bound the first term of (A.101) as

$$\|\Sigma_\phi - X^T X\|_\infty \leq \max_j \nu_j^{1/2} \left(\sum_{k \in S(j \bmod n) \setminus j} \nu_k^{1/2} \right) \quad (\text{A.103})$$

$$\leq (m-1)^{1/2} \max_j \nu_j^{1/2} \left(\sum_{k \in S(j \bmod n) \setminus j} \nu_k \right)^{1/2} \quad (\text{A.104})$$

$$\leq m^{1/2} \max_j \nu_j^{1/2} \left(\sum_{\ell=1}^m (\eta_\ell + t) \right)^{1/2} \quad (\text{A.105})$$

$$\leq \left(\frac{m}{n} + m^2 t \right)^{1/2}, \quad (\text{A.106})$$

where line (A.103) uses Gershgorin circles. Meanwhile, defining $\zeta := \left(\sum_{k \in \Omega_{n_0}} \nu_k\right)$, by Cauchy-Schwarz we have that

$$\|\alpha_0\|^2 \leq P\zeta^{-2}, \quad (\text{A.107})$$

where P is the signal power of Eq. (A.87). A necessary condition for producing benign overfitting in overparameterized Fourier models is that ζ remains relatively large as $n, d \rightarrow \infty$. If this is accomplished, then a small enough t guarantees that all feature weights associated with frequencies $k \in \Omega_d \setminus \Omega_{n_0}$ will be uniformly suppressed. For instance, if ζ is lower bounded as a constant while $t = 0$ then combining Eqs. (A.106) and (A.107) yields a bound of

$$\text{BIAS}^2 = \mathcal{O}\left(\frac{d^{1/2}}{n}\right). \quad (\text{A.108})$$

Although the analysis of Sec. A.1.3.1 yields a significantly tighter bound, this demonstrates that the mechanisms behind that simple demonstration of benign overfitting are somewhat generic. In particular, normalization and lower bounded support of the feature weights on Ω_{n_0} is almost sufficient to control the bias term of the generalization error.

A.2 Solution for quantum overparameterized model

We now derive the solution to the minimum-norm interpolating quantum model,

$$M^{opt} = \arg \min_{\substack{M=M^\dagger \\ f(x_j)=y_j}} \|M\|_F, \quad (\text{A.109})$$

$$f(x) = \sum_{k \in \Omega} e^{i2\pi kx} \sum_{\ell, m \in R(k)} \gamma_\ell M_{m\ell} \gamma_m^*. \quad (\text{A.110})$$

We will use the following notation and definitions:

$$a_k = -\left\lfloor \frac{k + \frac{|\Omega| - n}{2}}{n} \right\rfloor, \quad (\text{A.111})$$

$$b_k = \left\lfloor \frac{-k + \frac{|\Omega| + n}{2} - 1}{n} \right\rfloor, \quad (\text{A.112})$$

$$S_k = \{k - a(k)n, \dots, k, k + n, \dots, k + b(k)n\} \quad (\text{A.113})$$

$$= \{k + pn, \quad p = -a(k), \dots, b(k)\}, \quad (\text{A.114})$$

$$R(S_k) := \bigcup_{j \in S_k} R(j). \quad (\text{A.115})$$

Here, a_k and b_k characterize the number positive and negative frequency aliases of k appearing in Ω (i.e., $k + n\ell \in \Omega$ for all $a_k \leq \ell \leq b_k$) assuming that $|\Omega|$ is odd, a requirement for any quantum model. Let $L(d)$ be the space of linear operators acting on $d \times d$ matrices. Define the linear operator $P_k : L(d) \rightarrow L(d)$ for $k \in [n]$ as

$$P_k(X) = \sum_{\ell, m \in R(k)} \langle \ell | X | m \rangle | \ell \rangle \langle m | \quad (\text{A.116})$$

for any $X \in L(d)$. Importantly, P_k is not necessarily Hermitian preserving. Denoting $\Gamma := |\Gamma\rangle\langle\Gamma|$ for brevity, we may rewrite the Fourier coefficients of $f(x)$ as

$$\langle P_k(M), P_k(\Gamma) \rangle = \text{Tr} (P_k(M)^\dagger P_k(\Gamma)) \quad (\text{A.117})$$

$$= \text{Tr} \left(\sum_{\ell, m \in R(k)} M_{\ell m}^* |m\rangle \langle \ell| \sum_{i, j \in R(k)} \gamma_i \gamma_j^* |i\rangle \langle j| \right) \quad (\text{A.118})$$

$$= \sum_{\ell, m \in R(k)} \gamma_\ell^* M_{m\ell} \gamma_m, \quad (\text{A.119})$$

where in the last line we have used hermiticity of M . Applying $f(x_j) = y_j \forall j \in [n]$ and substituting into Eq. (A.110) we find the interpolation condition

$$\hat{y}_p = \sum_{q=a_k}^{b_k} \langle P_{p+nq}(M), P_{p+nq}(\Gamma) \rangle \quad (\text{A.120})$$

$$= \langle P_{S_p}(M), P_{S_p}(\Gamma) \rangle, \quad \forall p \in [n], \quad (\text{A.121})$$

where $P_{S_p} := \sum_{q=a_k}^{b_k} P_{p+nq}$. The equality follows from

$$\langle P_j(X), P_k(Y) \rangle = \langle P_j(X), P_j(Y) \rangle \delta_j^k \quad (\text{A.122})$$

due to the fact that $R(j)$ and $R(k)$ are disjoint sets for any $j \neq k$. Following the technique of Ref. [71] we apply the Cauchy-Schwarz inequality to find

$$\|P_{S_p}(M)\|_F \geq \frac{|\langle P_{S_p}(M), P_{S_p}(\Gamma) \rangle|}{\|P_{S_p}(\Gamma)\|_F}, \quad (\text{A.123})$$

with equality if and only if $P_{S_p}(M)$ is proportional to $P_{S_p}(\Gamma)$. Saturating this lower bound by setting $M_{\ell m} = c \gamma_\ell \gamma_m^*$ and solving for the proportionality c constant using Eq. (A.123), we find

$$\hat{y}_p = \langle c P_{S_p}(\Gamma), P_{S_p}(\Gamma) \rangle = \sum_{\ell, m \in R(S_p)} c^* |\gamma_\ell|^2 |\gamma_m|^2. \quad (\text{A.124})$$

This indicates an additional requirement for interpolation that $\gamma_\ell, \gamma_m > 0$ for some pair $(\ell, m) \in R(k)$ whenever $\tilde{y}_k \neq 0$ and so for simplicity we will require that $\gamma_\ell > 0$ for all $\ell = 1, \dots, d$. Within each set of indices $R(S_k)$, the elements of the optimal observable are defined piecewise with respect to that partition R :

$$(\ell, m) \in R(S_k) \Rightarrow M_{\ell m} = \hat{y}_k^* \frac{\gamma_\ell \gamma_m^*}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2}. \quad (\text{A.125})$$

Minimization of $\|M\|_F$ subject to the interpolation constraint is equivalent to minimization of $\|P_{p+nq}(M)\|_F$ for all $q \in [a_k, b_k]$, $k \in \Omega$, and so solving the constrained optimization over all distinct subspaces in $\bigcup_{k=0}^{n-1} R(S_k) = \{0, 1\}^n \times \{0, 1\}^n$ we recover the optimal observable

$$M_{opt} = \arg \min_{f(x_j)=y_j \forall j} \|M\|_F = \sum_{k \in \Omega_n} \hat{y}_k^* \sum_{\ell, m \in R(S_k)} \frac{\gamma_\ell \gamma_m^*}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2} |\ell\rangle \langle m|. \quad (\text{A.126})$$

We now verify that this matrix is Hermitian and therefore a valid observable. We will use the following:

$$\hat{y}_k = \hat{y}_{-k}^*, \quad (\text{A.127})$$

$$R(-k) = \{(m, \ell) \text{ for all } (\ell, m) \in R(k)\} := R(k)^T. \quad (\text{A.128})$$

Eq. (A.127) follows from our assumption that $y_j \in \mathbb{R} \forall j \in [n]$. And so

$$M_{\ell m} = \hat{y}_k^* \frac{\gamma_\ell \gamma_m^*}{\sum_{i,j \in R(S_k)} |\gamma_i|^2 |\gamma_j|^2} \quad \forall (\ell, m) \in R(S_k) \quad (\text{A.129})$$

$$= \hat{y}_{-k} \frac{\gamma_\ell \gamma_m^*}{\sum_{j,i \in R(S_{-k})^T} |\gamma_i|^2 |\gamma_j|^2} \quad \forall (\ell, m) \in R(S_{-k})^T \quad (\text{A.130})$$

$$= \left(\hat{y}_{-k}^* \frac{\gamma_m \gamma_\ell^*}{\sum_{i,j \in R(S_{-k})} |\gamma_i|^2 |\gamma_j|^2} \right)^* \quad \forall (m, \ell) \in R(S_{-k}) \quad (\text{A.131})$$

$$= M_{m\ell}^*. \quad (\text{A.132})$$

In line (A.130) we have used Eq. (A.127), while in line (A.132) we have observed that Eq. (A.125) holds only with respect to any fixed partition; $M_{\ell m}$ and $M_{m\ell}$ must be computed with respect to distinct partitions. The optimal model may now be rewritten in terms of base frequencies as

$$f^{opt}(x) = \sum_{k \in \Omega_n} \hat{y}_k \sum_{\ell=a_k}^{b_k} e^{i2\pi(k+n\ell)x} \frac{\sum_{i,j \in R(k+n\ell)} |\gamma_i|^2 |\gamma_j|^2}{\sum_{r=a_k}^{b_k} \sum_{c,d \in R(k+nr)} |\gamma_c|^2 |\gamma_d|^2}. \quad (\text{A.133})$$

Recall that the optimal classical model derived in Eq. (A.28) is given by

$$f(x) = \sum_{k \in \Omega_n} \hat{y}_k \sum_{\ell=a_k}^{b_k} e^{i2\pi(k+\ell n)x} \frac{\nu_{k+\ell n}}{\sum_{r=a_k}^{b_k} \nu_{k+nr}}. \quad (\text{A.134})$$

Then despite Eq. (A.110) not having a clear decomposition into scalar feature weights and trainable weights, we can identify the feature weights of the optimized quantum model as

$$\nu_k^{opt} := \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2, \quad (\text{A.135})$$

which recovers the same form of the optimal classical model of Eq. (A.28). This means that the behavior of the feature weights of the (optimal) general quantum model are strongly controlled by the degeneracy sets $R(k)$. The generalization error of the quantum model is also described by Eq. (3.21) of the main text under the identification of Eq. (A.135), and therefore exhibits a tradeoff that is predominantly controlled by the degeneracies $R(k)$ of the data-encoding Hamiltonian.

We can now substitute $\gamma_\ell = 1/\sqrt{d}$ to recover the optimal observable for the simplified model derived in Sec 3.4.1 of the main text using other means, namely

$$M_{m\ell}^{opt} = \hat{y}_k \frac{d}{\sum_{r=a_k}^{b_k} |R(k+nr)|}, \quad (\text{A.136})$$

$$f(x) = \sum_{k \in \Omega_n} \hat{y}_k \sum_{q=a_k}^{b_k} e^{i2\pi(k+nq)x} \frac{|R(k+nq)|}{\sum_{r=a_k}^{b_k} |R(k+nr)|}, \quad (\text{A.137})$$

$$\nu_k = \frac{|R(k)|}{d^2}. \quad (\text{A.138})$$

A.2.1 Computing feature weights of typical quantum models

In Sec 3.4.1 we introduced a simple model with a uniform amplitude state $|\Gamma\rangle$ as input and demonstrated that the feature weights of this simple quantum model are completely determined by the sets $R(k)$ induced by the encoding strategy. We now wish to extend the intuition that the behavior of the optimized general quantum model is strongly influenced by the distribution of the degeneracies $|R(k)|$. We do so by evaluating the optimal quantum models with respect to an “average” state preparation unitary U . We can compute

the average value of $|\gamma_i|^2|\gamma_j|^2$ for U sampled uniformly from the Haar distribution using standard results from the literature [54]²:

$$\mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} |\gamma_i|^2 |\gamma_j|^2 = \mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} U_{i0} U_{i0}^* U_{j0} U_{j0}^* = \frac{\delta_i^j + 1}{d(d+1)}. \quad (\text{A.139})$$

Since $(i, i) \in R(0)$ we can then compute the feature weights of the optimal model according to³

$$\mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} \nu_k^{opt} = \mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2 = \frac{|R(k)|}{d(d+1)} + \delta_0^k \frac{1}{d+1}. \quad (\text{A.140})$$

From Eq. (A.140) we see that the feature weights of a quantum model optimized with respect to random U are completely determined by the degeneracies $R(k)$. This expected value is useful but does not fully characterize the behavior of an encoding strategy. To demonstrate that this average behavior is meaningful, we would further like to verify that the feature weights corresponding to a random U concentrate around the mean of Eq. (A.140). We characterize this by computing the variance

$$\text{Var}(\nu_k) = \mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} (\nu_k^2) - \left(\mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} \nu_k \right)^2, \quad (\text{A.141})$$

where we have dropped the superscript on ν_k^{opt} for brevity. This computation requires significantly more counting arguments dependent on the structure of $R(k)$. When $k \neq 0$,

²Note that since ν_k^{opt} is invariant with respect to $\gamma_i \rightarrow e^{i\phi} \gamma_i$ a spherical measure would suffice here.

³It is implied that we compute the optimal ν with respect to each distinct U sampled independently and uniformly with respect to the Haar measure – without optimizing M with respect to each U we would find the trivial result

$$\mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} f(x) = \mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} \text{Tr}(M \rho(x)) = d^{-1} \text{Tr } M$$

we identify cases for which $i \neq j$ whenever $(i, j) \in R(k)$:

$$\nu_k^2 = \sum_{i,j \in R(k)} \sum_{\ell, m \in R(k)} |\gamma_i|^2 |\gamma_j|^2 |\gamma_\ell|^2 |\gamma_m|^2 \quad (\text{A.142})$$

$$= \sum_{i,j \in R(k)} \left(|\gamma_i|^4 |\gamma_j|^4 + \sum_{\substack{\ell, m \in R(k) \\ \ell \neq j, m = i}} |\gamma_i|^4 |\gamma_j|^2 |\gamma_\ell|^2 \right. \\ \left. + \sum_{\substack{m, \ell \in R(k) \\ \ell = j, m \neq i}} |\gamma_i|^2 |\gamma_j|^4 |\gamma_m|^2 + \sum_{\substack{m, \ell \in R(k) \\ \ell \neq j, m \neq i}} |\gamma_i|^2 |\gamma_j|^2 |\gamma_\ell|^2 |\gamma_m|^2 \right). \quad (\text{A.143})$$

The expected values for of these terms are evaluated using the observation that the vector $(|\gamma_1|^2, \dots, |\gamma_s|^2)$ is distributed uniformly on the d -simplex leading to simple expressions for the following expected values [54]:

$$\mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} |\gamma_p|^4 |\gamma_q|^4 = 4/D, \quad (\text{A.144})$$

$$\mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} |\gamma_p|^4 |\gamma_q|^2 |\gamma_r|^2 = 2/D, \quad (\text{A.145})$$

$$\mathbb{E}_{U \sim \mathcal{U}(\mathbb{C}^d)} |\gamma_p|^2 |\gamma_q|^2 |\gamma_r|^2 |\gamma_s|^2 = 1/D, \quad (\text{A.146})$$

where $p, q, r, s = 1, \dots, d$ and $p \neq q \neq r \neq s$, and $D := (d+3)(d+2)(d+1)d$. In computing the expected value of Eq. (A.143), by linearity the terms of each sum will become a constant and only the number of items in each sum will be relevant. The first sum of Eq. (A.143) contains $|R(k)|$ many terms and the total number of terms is $|R(k)|^2$, and so we only need compute the number of elements in the two middle sums of Eq. (A.143) (which contain an equal number of terms due to the symmetry $R(k) = R(-k)^T$). These computations may be carried out by brute-force combinatorics and are summarized in Table A.1 for a few of the models studied in this work.

For the Binary encoding strategy we compute

$$\mathbb{E}(\nu_k^2) = \begin{cases} \frac{5d-7k+(d-k)^2}{D}, & 0 < 2k \leq d \\ \frac{3(d-k)+(d-k)^2}{D}, & 2k > d, \end{cases} \quad (\text{A.147})$$

Encoding strategy	$ \{(i, j, \ell, m) : m \neq i, \ell = j, (i, j), (\ell, m) \in R(k)\} $	$\text{Var}(\nu_k^{opt})$
Hamming	$\sum_{p=0}^{n_q-2k} \binom{n_q}{p} \binom{n_q}{p+ k } \binom{n_q}{p+2 k }$	-
Binary	$\max(2^{n_q} - 2 k , 0)$	$O(2^{-3n_q})$
Golomb	0	$O(d^{-4})$

Table A.1: The size of the subset of $R(k) \times R(k)$ with a single repeated index computed for different encoding strategies and the corresponding scaling of the variance of feature weights (all entries assume $k \neq 0$). The Hamming encoding strategy computation works as follows: Each bitstring i with weight p (of which there are n -choose- p) is paired with n -choose- $(p+k)$ many bitstrings j with weight $p+k$. Then, taking $\ell = j$ there will be n -choose- $(p+2k)$ many bitstrings m with weight $(p+k)+k = p+2k$. Summing over all such p 's where $p+2k \leq n_q$ yields the desired result, however this computation does not admit a clear closed-form expression and so we have omitted the corresponding scaling of $\text{Var}(\nu_k^{opt})$ for the Binary encoding strategy.

and so

$$\text{Var}(\nu_k) = \left\{ \begin{array}{ll} \frac{(5d-7k)d(d+1) - (d-k)^2(4d+6)}{Dd(d+1)} & \text{if } 0 < 2k \leq d \\ \frac{3(d-k)d(d+1) - (d-k)^2(4d+6)}{Dd(d+1)} & \text{if } 2k > d \end{array} \right\} \leq O(d^{-3}). \quad (\text{A.148})$$

Importantly, $\mathbb{E}[\nu_k] \geq O(d^{-2})$ for the Binary encoding strategy. Taking $d = 2^{n_q}$ corresponding to n_q qubits, then while the mean decays exponentially in the number of qubits n_q , the variance decays exponentially faster. For the Golomb encoding strategy, the calculation is comparatively straightforward, yielding for $k \neq 0$

$$\text{Var}(\nu_k) = \frac{3d^2 - d - 6}{Dd(d+1)} = O(d^{-4}), \quad (\text{A.149})$$

with the variance again decaying significantly faster in d than the mean. Figure A.1 shows the average and variance of ν_k^{opt} for the general quantum model with U sampled uniformly with respect to the Haar measure. We find tight concentration of ν_k^{opt} around its average in each of these cases

Why is it that the feature weights ν_k^{opt} of the quantum model given by Eq. (A.135) only appear after deriving the optimal observable? One explanation is that while the classical Fourier Features model utilizes random Fourier features such that components of $\phi(x)$ are

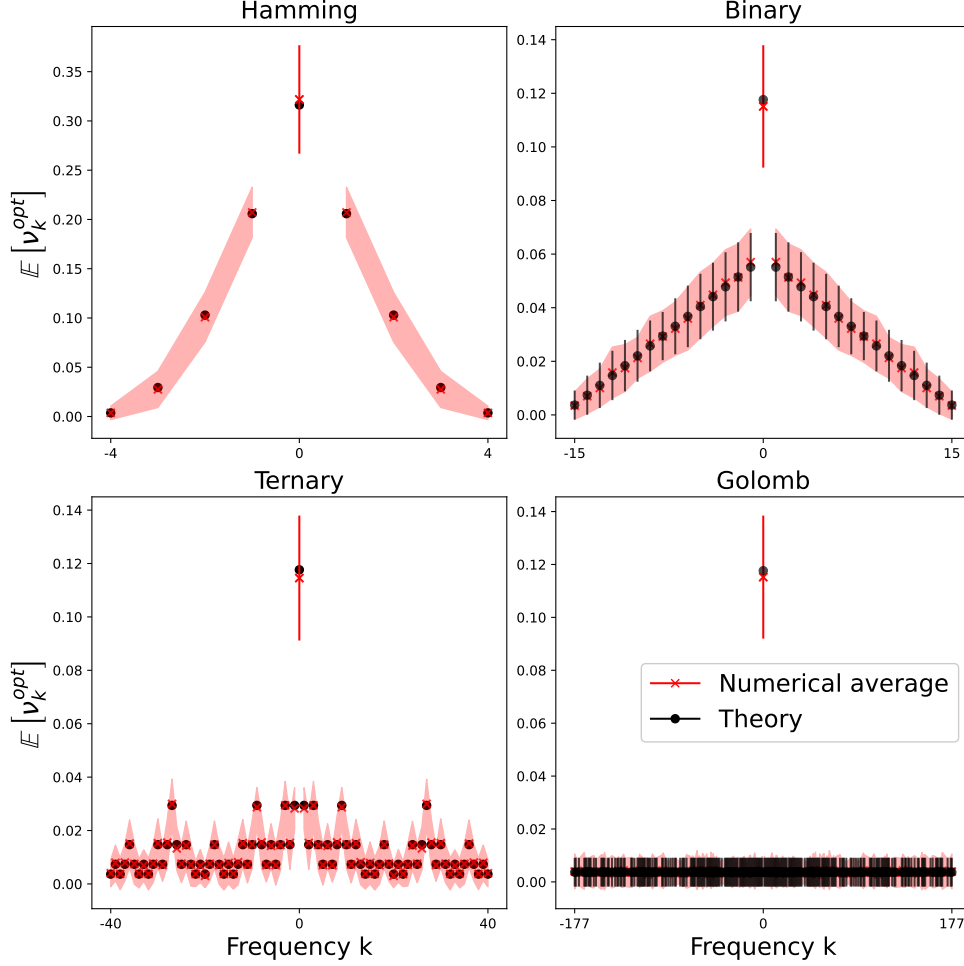


Figure A.1: Plotting $\mathbb{E}[\nu_k^{opt}]$ and $\text{Var}[\nu_k^{opt}]^{1/2}$ demonstrates tight concentration of the distribution of ν_k^{opt} around its mean with respect to Haar-random U in the general quantum model. For the Binary and Golomb encoding strategies we plot Eqs. (A.148) and (A.149) respectively. For all models we also include averages and variances computed by sampling states uniformly from the Haar measure numerically.

mutually orthogonal on $[0, 1]$, the quantum model does not. Consider the operator

$$\Sigma := \int_{\mathcal{X}} \text{Vec}(\rho(x)) \text{Vec}(\rho(x))^\dagger p(x) dx = \int_{\mathcal{X}} \rho(x) \otimes \rho^T(x) p(x) dx, \quad (\text{A.150})$$

where $p : \mathcal{X} \rightarrow \mathbb{R}$ is the probability density function describing the distribution of data in $\mathcal{X}$ and $\text{Vec} : \mathbb{C}^{d \times d} \rightarrow \mathbb{C}^{d^2}$ is the *vectorization map* that acts by stacking transposed rows of a square matrix into a column vector. Σ is analogous to a classical covariance matrix, and here determines the correlation between components of $\rho(x)$ (with the second equality holding when $\rho(x)$ is a pure state for all x). From Eq. (3.9) describing the classical Fourier features it is straightforward to compute $\mathbb{E}_{x \sim [0,1]}[\phi \phi^\dagger] = \mathbb{I}_d$, demonstrating that the classical Fourier features are indeed orthonormal. However, under the identification of the feature vector $\phi(x) \rightarrow \text{Vec}(\rho(x))$ in the quantum case, the same is not true for the quantum model:

$$\Sigma = \int_0^1 \left(\sum_{k \in \Omega} e^{i2\pi kx} \sum_{i,j \in R(k)} \gamma_i \gamma_j^* |ij\rangle \right) \left(\sum_{k' \in \Omega} e^{i2\pi k'x} \sum_{\ell, m \in R(k')} \gamma_\ell^* \gamma_m \langle \ell m| \right) dx \quad (\text{A.151})$$

$$= \sum_k \sum_{\substack{i,j \in R(k) \\ \ell, m \in R(k)}} \gamma_i \gamma_j^* \gamma_\ell^* \gamma_m |ij\rangle \langle \ell m|. \quad (\text{A.152})$$

Thus Σ is not diagonal in general, as many components of $S(x)|\Gamma\rangle$ each contribute to a single frequency k . Nor will it be possible in general to construct a quantum feature map via unitary operations that does act as an orthogonal Fourier features vector. As Σ is positive semidefinite by construction, there exists a spectral decomposition $\Sigma = WDW^\dagger$ with diagonal D and $d^2 \times d^2$ unitary W . However the linear operator $\Phi \in L(d)$ acting on d -dimensional states according to $\text{Vec}(\Phi(\rho)) = W^\dagger \text{Vec}(\rho(x))$ will not be unitary in general, and thus it may not be possible to prepare ρ in such a way that the elements of $\text{Vec}(\rho(x))$ consist of orthonormal Fourier features.

A.2.2 The effect of state preparation on feature weights in the general quantum model

We now discuss how the state preparation unitary U may affect the feature weights ν_k^{opt} in a quantum model, and what choices one can make to construct a U that gives rise to a specific distribution of feature weights ν_k^{opt} .

As an example, we consider the general quantum model using the Hamming encoding strategy and an input state $|\Gamma\rangle$. Computing the feature weights ν_k^{opt} for the Hamming

encoding strategy depends on the amplitudes γ_i and γ_j for which the weights of the indices i, j differ by k . We have

$$\nu_k^{opt} = \sum_{i,j \in R(k)} |\gamma_i|^2 |\gamma_j|^2 = \sum_{\substack{i,j: \\ w(j)-w(i)=k}} |\gamma_i|^2 |\gamma_j|^2. \quad (\text{A.153})$$

We will now show how the feature weight ν_k^{opt} may be computed with respect to a rebalanced input state which distributes each amplitude γ_i among all computational basis states with index having weight $w(i)$. We define this rebalanced state on n qubits as

$$|\Gamma'\rangle = \sum_{i \in \{0,1\}^n} \gamma'_i |i\rangle = \sum_{i=0}^n \sum_{j \in W(i)} \phi_i |j\rangle, \quad (\text{A.154})$$

$$\phi_i = \frac{1}{\sqrt{|W(i)|}} \left(\sum_{j \in W(i)} |\gamma_j|^2 \right)^{1/2}. \quad (\text{A.155})$$

where $W(i) = \{j : j \in \{0,1\}^n, w(j) = i\}$ is the set of weight- i indices and $|W(i)| = \binom{n}{i}$. Observing that the amplitudes of $|\Gamma'\rangle$ satisfy $|\gamma'_i|^2 = \phi_{w(i)}^2$, we can derive for $k \geq 0$

$$\nu_k^{opt} = \sum_{\substack{i,j: \\ w(j)-w(i)=k}} |\gamma_i|^2 |\gamma_j|^2 \quad (\text{A.156})$$

$$= \sum_{\ell=0}^{n-k} \left(\sum_{i:w(i)=\ell} |\gamma_i|^2 \right) \left(\sum_{j:w(j)=\ell+k} |\gamma_j|^2 \right) \quad (\text{A.157})$$

$$= \sum_{\ell=0}^{n-k} \left(|W(\ell)| \frac{\sum_{i:w(i)=\ell} |\gamma_i|^2}{|W(\ell)|} \right) \left(|W(\ell+k)| \frac{\sum_{j:w(j)=\ell+k} |\gamma_j|^2}{|W(\ell+k)|} \right) \quad (\text{A.158})$$

$$= \sum_{\ell=0}^{n-k} |W(\ell)| \phi_\ell^2 |W(\ell+k)| \phi_{\ell+k}^2 \quad (\text{A.159})$$

$$= \sum_{\ell=0}^{n-k} \left(\sum_{i:w(i)=\ell} \phi_{w(i)}^2 \right) \left(\sum_{j:w(j)=\ell+k} \phi_{w(j)}^2 \right) \quad (\text{A.160})$$

$$= \sum_{\substack{i,j: \\ w(j)-w(i)=k}} |\gamma'_i|^2 |\gamma'_j|^2. \quad (\text{A.161})$$

Therefore, the feature weights computed from $|\Gamma\rangle = U|0\rangle$ and the rebalanced state $|\Gamma'\rangle$ are identical. We can emphasize the significance of this observation by rewriting $|\Gamma'\rangle$ of Eq. (A.154) as

$$|\Gamma'\rangle = \sum_{i=0}^n \phi_i |\Phi_i\rangle, \quad (\text{A.162})$$

where $|\Phi_i\rangle = \sum_{j \in W(i)} |j\rangle$ is an (unnormalized) superposition of bitstrings with weight i . The state $|\Gamma'\rangle$ of Eq. (A.162) has only $n+1$ real parameters ϕ_i , and is invariant under operations that are restricted to act within the subspace spanned by components of $|\Phi\rangle$. This invariance greatly reduces the class of unitaries U that affect the distribution of feature weights ν_k^{opt} and enables some degree of tuning for these parameters.

A.2.2.1 Vanishing gradients in preparing feature weights

We briefly remark on the possibility of training feature weights ν_k^{opt} of the general quantum model variationally. For example, one could consider defining a desired distribution of feature weights as $\phi_k \in \mathbb{R}^{|\Omega|}$ and then attempting to tune parameters of U in order to minimize a cost function such as⁴

$$L = \sum_{k \in \Omega} |\nu_k^{opt} - \phi_k|^2. \quad (\text{A.163})$$

We will now demonstrate that for a sufficiently expressive class of state preparation unitaries U , such an optimization problem will be difficult on average. For simplicity, we follow the original formulation of barren plateaus presented in Ref. [82]: Let U be defined with respect to a set of parameters $\theta \in \mathbb{R}^L$ as

$$U = \prod_{\ell=1}^L U(\theta_\ell) W_\ell \quad (\text{A.164})$$

with $U(\theta_\ell) = \exp(i\theta_\ell V_\ell)$, V_ℓ being a d -dimensional Hermitian operator, and W_ℓ being a d -dimensional unitary operator. Pick a parameter θ_m and define $U_+ = \prod_{\ell=m}^L U(\theta_\ell) W_\ell$ and $U_- = \prod_{\ell=1}^m U(\theta_\ell) W_\ell$, and observe that

$$\partial_{\theta_m} |\gamma_j|^2 = i \langle 0 | U_-^\dagger [V_m, U_+^\dagger |j\rangle \langle j| U_+] U_- | 0 \rangle. \quad (\text{A.165})$$

⁴As ν_k represents a probability distribution, other distance measures such as relative entropy or earth-mover's distance may be considered more appropriate. However using alternative cost functions will not affect the main arguments here.

We can compute the derivative of ν_k with respect to θ_m using the chain rule:

$$\frac{\partial \nu_k}{\partial \theta_m} = \sum_{j=1}^d \frac{\partial \nu_k}{\partial |\gamma_j|^2} \frac{\partial |\gamma_j|^2}{\partial \theta_m} \quad (\text{A.166})$$

$$= \sum_{j=1}^d \left(\sum_{a,b \in R(k)} (|\gamma_a|^2 \delta_b^j + \delta_a^j |\gamma_b|^2) \right) i \langle 0|U_-^\dagger [V_m, U_+^\dagger |j\rangle \langle j|U_+] U_- |0\rangle \quad (\text{A.167})$$

$$= i \sum_{a,b \in R(k)} [\text{Tr}(\rho_- H_a) \text{Tr}(\rho_- [V_m, H_b]) + \text{Tr}(\rho_- [V_m, H_a]) \text{Tr}(\rho_- H_b)] \quad (\text{A.168})$$

where $\rho_- = U_- |0\rangle \langle 0| U_-^\dagger$ and $H_x = U_+^\dagger |x\rangle \langle x| U_+$, and we used the equality

$$|\gamma_a|^2 = \langle 0|U_-^\dagger U_+^\dagger |a\rangle \langle a|U_+ U_- |0\rangle = \text{Tr}(\rho_- H_a). \quad (\text{A.169})$$

We now show that each term in this sum vanishes by following Ref. [82] in letting U_- be sampled from a distribution that forms a unitary 2-design:

$$\mathbb{E}_{U_- \sim U(d)} \text{Tr}(\rho_- H_a) \text{Tr}(\rho_- [V_m, H_b]) \quad (\text{A.170})$$

$$= \text{Tr}(\mathbb{E}_{U_- \sim U(d)} [\rho_- \otimes \rho_-] H_a \otimes [V_m, H_b]) \quad (\text{A.171})$$

$$= \frac{1}{d(d+1)} \text{Tr} \left(\left(\mathbb{I}_d + \sum_{i,j=1}^d |ij\rangle \langle ji| \right) H_a \otimes [V_m, H_b] \right) \quad (\text{A.172})$$

$$= \frac{\text{Tr}([V_m, H_b] H_a)}{d(d+1)} \quad (\text{A.173})$$

$$= \frac{V_m U_+^\dagger |b\rangle \langle b| U_+ U_+^\dagger |a\rangle \langle a| U_+ - U_+^\dagger |b\rangle \langle b| U_+ V_m U_+^\dagger |a\rangle \langle a| U_+}{d(d+1)} \quad (\text{A.174})$$

$$= \frac{\delta_a^b \langle a|U_+ V_m U_+^\dagger |b\rangle - \delta_a^b \langle b|U_+ V_m U_+^\dagger |a\rangle}{d(d+1)} \quad (\text{A.175})$$

$$= 0, \quad (\text{A.176})$$

where in line (A.172) we have used a common expression for $\mathbb{E}_{U \sim U(d)} [\rho \otimes \rho]$ in terms of the projector onto the symmetric subspace (e.g. [292]). Substituting this result into Eq. (A.168) we find

$$\mathbb{E}_{U \sim U(d)} \left(\frac{\partial \nu_k}{\partial \theta_m} \right) = 0 \quad (\text{A.177})$$

By extension, the gradient of a loss function of the form of Eq. (A.163) will vanish for expressive enough state-preparation unitaries U , suggesting that solving for a choice of U to induce a specific distribution of feature weights ν_k^{opt} will be infeasible in practice.

A.3 Determining the degeneracy of quantum models

Here we will develop a theoretical framework for manipulating the degeneracy (and therefore feature weights) of quantum models. We will begin with choosing the data-encoding operator $S(z) = \exp(-iHz)$ to be separable, generated from a Hamiltonian of the form

$$H = \mathbf{r} \cdot \mathbf{Z} = \sum_{j=0}^{n-1} r_j Z^{(j)}, \quad (\text{A.178})$$

and $Z^{(j)}$ denotes the operator that applies a Pauli Z operator to the j^{th} register and acts trivially elsewhere, and $\mathbf{r} \in \mathbb{R}^n$. One can show that the diagonal elements of H are given by

$$\lambda_j = H_{jj} = 2(\mathbf{r} \cdot \mathbf{j}) - \|\mathbf{r}\|_1, \quad (\text{A.179})$$

where here and elsewhere we use interchangeably a bitstring denoted as $j = j_0 j_1 \dots j_{n-1}$ with decimal value $j = \sum_k j_k 2^k \in [0, 2^n - 1]$, and $\mathbf{j} \in \{0, 1\}^n$ as its corresponding vector representation. The Fourier spectrum and degeneracy of the encoding strategy are then computed as

$$\Omega = \{\lambda_j - \lambda_i : i, j = 0, \dots, 2^n - 1\} \quad (\text{A.180})$$

$$= \{2(\mathbf{r} \cdot (\mathbf{j} - \mathbf{i})) : \mathbf{i}, \mathbf{j} \in \{0, 1\}^n\}, \quad (\text{A.181})$$

$$R(\omega) = \{(i, j) : 2\mathbf{r} \cdot (\mathbf{j} - \mathbf{i}) = k\}. \quad (\text{A.182})$$

Note that the subtraction in the definition of Ω does not preserve $\mathbb{Z}_2$, i.e. $(\mathbf{j} - \mathbf{i}) \in \{-1, 0, 1\}^n$. From Eq. (A.180) we see that the largest possible size for $|\Omega|$ is 3^n , the number of unique choices for $(\mathbf{j} - \mathbf{i})$. Any encoding strategy of Eq. (A.178) therefore introduces a *combinatorial degeneracy*, since the set $\{(\mathbf{j} - \mathbf{i}) : \mathbf{i}, \mathbf{j} \in \{0, 1\}^n\}$ is the image of a surjective map on $\{0, 1\}^n \times \{0, 1\}^n$, with each $(\mathbf{j} - \mathbf{i})$ occurring with multiplicity

$$\#(\mathbf{j} - \mathbf{i}) = 2^{n - \|\mathbf{j} - \mathbf{i}\|_1}. \quad (\text{A.183})$$

To prove this, construct the set $\{(\mathbf{j} - \mathbf{i}) : \mathbf{i}, \mathbf{j} \in \{0, 1\}^n\}$ by reflecting the hypercube $\{0, 1\}^n$ over 2^n distinct axes, and count the number of vertices shared by among reflected images – this gives the desired degeneracy. Many choices of $\mathbf{r}$ will reduce $|\Omega|$ by reducing the size of the image of the map $(\mathbf{i}, \mathbf{j}) \rightarrow \mathbf{r} \cdot (\mathbf{j} - \mathbf{i})$. Conversely, the spectrum Ω will saturate $|\Omega| = 3^n$ only under particular conditions on $\mathbf{r}$. For example with $n = 2$, $|\Omega| = 9$ is achieved only if $\mathbf{r}$ satisfies all of the following conditions

$$r_0 \neq 0, \quad r_0 \neq \pm \frac{1}{2}r_1, \quad r_0 \neq \pm r_1, \quad r_0 \neq \pm 2r_1, \quad (\text{A.184})$$

i.e., the inverse of the dot product $\mathbf{r} \cdot \mathbf{d}$ with respect to an input $\mathbf{d}$ exists if the elements of $\mathbf{r}$ satisfy $r_j \neq r_i \pmod{2^p}$ for some integer p ; a more general, recursive statement of this property is provided in Ref. [97]. Requirements of this form may be proven using the *frame analysis operator* $T_n \in \mathbb{Z}^{3^n \times n}$, the rows of which contain each element of $\{-1, 0, 1\}^n$. And so to demonstrate Eq. (A.184) above, if we let $\mathbf{k} \in \mathbb{R}^{3^n}$ be a vector containing (repeated) frequencies in Ω , we observe that $T_2 \mathbf{r} = \mathbf{k} \in \mathbb{R}^9$ will have unique entries only under specific conditions on $\mathbf{r}$. Ω being the result of a frame reconstruction of $\mathbf{r}$ also provides a direct way to tune the elements of Ω : given a length 3^n vector $\mathbf{k}$ containing the (not necessarily unique) desired elements of Ω , one can then compute H by solving $\mathbf{r} = T_n^\dagger \mathbf{k}$.

A.3.1 Hamming encoding strategy

We instantiate the Hamiltonian of Eq. (A.178) for

$$\mathbf{r} = \frac{1}{2}(1, 1, \dots, 1) \in \mathbb{R}^n.$$

This introduces many degeneracies into the spectrum; using $T_n \mathbf{r} = \mathbf{k}$ we immediately recover the result of Ref. [42] that the unique elements of $\mathbf{k}$ are given by

$$\Omega = \{-n, -(n-1), \dots, 0, \dots, n-1, n\}, \quad (\text{A.185})$$

with $|\Omega| = 2n + 1$. To recover the degeneracy of each frequency we compute

$$\lambda_i - \lambda_j = \frac{1}{2} \sum_{t=1}^n (-1)^{i_t} - \frac{1}{2} \sum_{t=1}^n (-1)^{j_t} = \frac{1}{2} \sum_{t=1}^n (1 - 2i_t) - \frac{1}{2} \sum_{t=1}^n (1 - 2j_t) = w(j) - w(i), \quad (\text{A.186})$$

where $w : \{0, 1\}^n \rightarrow [n]$ is the the weight of a bitstring $b = b_{n-1} \dots b_1 b_0 \in \{0, 1\}^n$ defined as

$$w(b) = \sum_{k=1}^n b_k, \quad (\text{A.187})$$

and when b is written as an integer it should be interpreted according to its binary value. It will be useful to consider subsets of bitstrings having a constant weight. For each $k \in [n]$ (results for $k < 0$ follow by symmetry), we now define the k -weight set $W(k)$ as

$$W(k) = \{j \in \{0, 1\}^n : w(j) = k\}. \quad (\text{A.188})$$

Then the degeneracy set is exactly the set of bitstring indices differing by constant weight:

$$R(k) = \{(i, j) \in \{0, 1\}^n \times \{0, 1\}^n : w(j) - w(i) = k\} \quad (\text{A.189})$$

$$= \bigcup_{p=k}^n W(p) \times W(p - k). \quad (\text{A.190})$$

where $\times$ denotes the Cartesian product (this motivates the name “Hamming encoding strategy”, as the frequencies and degeneracy of this model arise from index pairs having fixed Hamming weight). From counting arguments, the degeneracy of this model is then

$$|R(k)| = \sum_{j=1}^{n-k} \binom{n}{j} \binom{n}{k+j} = \binom{2n}{n-k}. \quad (\text{A.191})$$

A.3.2 Binary encoding strategy

We instantiate the Hamiltonian of Eq. (A.178) for

$$\mathbf{r} = \frac{1}{2}(2^{n-1}, \dots, 4, 2, 1) \in \mathbb{R}^n,$$

such that $\mathbf{r}$ generates the decimal value of binary vector. We know from relations like those given in Eq. (A.184) that $|\Omega| \ll 3^n$, and indeed using Eq. (A.179) we find that $\lambda_j = (j - (2^n - 1))/2$ resulting in a frequency spectrum of

$$\Omega = \{-2^n + 1, -2^n + 2, \dots, 0, 1, \dots, 2^n - 1\}, \quad (\text{A.192})$$

such that $|\Omega| = 2^{n+1} - 1$, and the degenerate index set of the kernel for frequency k is equivalent (up to permutations) to the indices corresponding to nonzero elements of an elementary Toeplitz matrix with 1’s on the k -th diagonal:

$$R(k) = \{(i, j) : i = j - k, j \in [2^n]\}. \quad (\text{A.193})$$

A.3.3 Ternary encoding strategy

We instantiate the Hamiltonian of Eq. (A.178) for

$$\mathbf{r} = \frac{1}{2}(3^{n-1}, \dots, 9, 3, 1) \in \mathbb{R}^n,$$

resulting in the spectrum

$$\Omega = \left\{ -\frac{1}{2}(3^n - 1), -\frac{1}{2}(3^n - 1) + 1, \dots, 0, \dots, \frac{1}{2}(3^n - 1) \right\}. \quad (\text{A.194})$$

This spectrum (also studied in Ref. [97]) is interesting to study because it is the unique choice giving rise to a Fourier spectrum with frequencies spaced by 1 saturating the $|\Omega| = 3^n$ limit for spectra generated by separable Hamiltonians. Given $k = 2\mathbf{r} \cdot \mathbf{d}$, then $\mathbf{d} = \mathbf{i} - \mathbf{j}$ may be uniquely recovered from k . Suppose k has a ternary representation

$$k = k_n \dots k_1 k_0 = \sum_j k_j 3^j, \quad (\text{A.195})$$

with $n = \lceil \log_3 k \rceil$ and $k_j \in \{0, 1, 2\}$. Then defining the operation $t : \mathbb{Z} \rightarrow \{0, 1, 2\}^n$ which converts decimals to their vector-form ternary representation according to $t(k) = (k_0, k_1, \dots, k_{n-1})$, then the degeneracy of this encoding strategy follows directly from the combinatorial degeneracy given in Eq. (A.183):

$$|R(k)| = 2^{n - \|T(k)\|_1}, \quad (\text{A.196})$$

where the shifted ternary operation is defined for $k \in [0, 3^n - 1]$

$$T(k) = t(k + \|\mathbf{r}\|_1) - \mathbf{1}, \quad (\text{A.197})$$

and $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{Z}^n$ is the array of all ones. Intuitively, this computation converts a signed ternary expression $k = \sum_j k_j 3^j$ (with $\mathbf{k} \in \{-1, 0, 1\}$) to an unsigned ternary expression provided in Eq. (A.195).

A.3.4 Nonseparable models

Given a data-encoding Hamiltonian $H = \text{diag}(\Lambda)$ with $\Lambda \in \mathbb{R}^d$, a key feature of the associated spectrum is that it is preserved under permutations of Λ . This means that a large class of models where $S(x)$ is generated by a separable $H = H_0 \otimes H_1 \otimes H_{n-1}$ are capable of producing the same spectrum as models where $S(x)$ is generated by a nonseparable Hamiltonian. For instance, letting $\sim$ denote equivalence up to permutation and additive shift, it holds that $I \otimes Z \sim Z \otimes Z$, so that a data encoding unitaries

$$S_1(x) = \exp(i2\pi r(I \otimes Z)x), \quad (\text{A.198})$$

$$S_2(x) = \exp(i2\pi r(Z \otimes Z)x), \quad (\text{A.199})$$

have identical spectrum and degeneracy properties (differing only in their corresponding degeneracy index sets R). Understanding the classes of nonseparable data-encoding Hamiltonians that are equivalent to a separable encoding strategy up to permutation and translation may be an interesting problem for future work.

We therefore limit our discussion of nonseparable models to a particular choice of Hamiltonian which is provably inequivalent to any separable Hamiltonian, as it saturates the largest possible $|\Omega|$ (and smallest values of $|R(k)|$) available to any single-layer quantum model. This can be achieved by setting the diagonal of the data-encoding Hamiltonian as the elements of a Golomb ruler [98, 293]. The resulting spectrum is nondegenerate for all nonzero frequencies, though it is straightforward to prove that one cannot achieve uniform spacing in the spectrum (i.e., a *Perfect Golomb ruler*) for $d \geq 5$. As a result, the corresponding spectrum of this model generally exhibits gaps between frequencies. Further exploration of the connections to concepts from radio engineering [294, 295] and classical coding theory [296] may enrich investigations into the spectral properties of quantum models.

A.4 Demonstration of benign overfitting in the general quantum model

In this section we demonstrate an example of benign overfitting in the general quantum model by explicitly constructing a sequence of state preparation unitaries U for a particular choice of data-encoding unitary $S(d)$. We will consider a d -dimensional version of the Binary encoding strategy constructed from a data-encoding Hamiltonian $H = \text{diag}(0, 1, 2, \dots, d-1)$ that gives rise to

$$\Omega = \left\{ \frac{-d}{2} + 1, \dots, 0, \dots, \frac{d}{2} - 1 \right\}, \quad (\text{A.200})$$

$$|R(k)| = d - |k|, \quad (\text{A.201})$$

for $k \in \Omega$, and we will require d to be even for simplicity. Suppose the target function spectrum has size $n_0 < n < d$ for some integer n and satisfies $(n_0 + 1) \bmod 4 = 0$. Define the constants

$$c_1 = \frac{d}{2} - \frac{n_0 + 1}{4}, \quad (\text{A.202})$$

$$c_2 = \frac{d}{2} + \frac{n_0 + 1}{4}. \quad (\text{A.203})$$

We then choose a constant $a \in [0, 2/(n_0 + 1)]$ and prepare $|\Gamma\rangle$ with elements given by

$$\gamma_j = \begin{cases} \sqrt{a}, & j \in [c_1, c_2), \\ \sqrt{b}, & j \in [0, c_1) \cup [c_2, d), \end{cases} \quad (\text{A.204})$$

where normalization requires that

$$b = \frac{1 - \frac{n_0+1}{2}a}{d - \frac{n_0+1}{2}}. \quad (\text{A.205})$$

For this encoding strategy, the optimal feature weights of the quantum model corresponding to positive frequencies $k \in \Omega_+$ are given by

$$\nu_k = \sum_{j=0}^{d-k-1} |\gamma_j|^2 |\gamma_{j+k}|^2. \quad (\text{A.206})$$

Counting arguments and algebraic simplification lead to

$$\nu_k = \begin{cases} (d - B_0 - 2k)b^2 + 2kab + (B_0 - k)a^2, & 0 \leq k < c_2 - c_1, \\ (d - 2B_0 - k)b^2 + 2B_0ab, & c_2 - c_1 \leq k < c_1, \\ (k - B_0)b^2 + (d - B_0 - 2k)ab, & c_1 \leq k < c_2, \\ (d - k)b^2, & c_2 \leq k < d, \end{cases} \quad (\text{A.207})$$

where $B_0 = (n_0 + 1)/2$ and $\nu_{k+1} \leq \nu_k$ for all $k > 0$. Noting that $b = \mathcal{O}(d^{-1})$ and a is constant, we find that ν_k has three distinct scaling regimes: $\nu_k = \Omega(1)$ for $k \in \Omega_{n_0}$, $\nu_k = \Theta(d^{-2})$ when $k \in [c_2, d)$, and $\nu_k = \mathcal{O}(d^{-1})$ otherwise. We can use this behavior to bound the generalization error of the quantum model in a manner analogous to Sec. A.1.3.1. Defining $m = |S(k)| - 1$, we bound the bias according to

$$\text{BIAS}^2 = \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \frac{\left(\sum_{\ell \in S(k) \setminus k} \nu_\ell\right)^2 + \sum_{\ell \in S(k) \setminus k} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \quad (\text{A.208})$$

$$\leq \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2 \frac{m(m+1)(db^2 + 2B_0ab)^2}{(a^2 + mb^2)^2} \quad (\text{A.209})$$

$$\leq [m(m+1)b^2] \left(\frac{db + 2B_0a}{a^2 + mb^2}\right)^2 P \quad (\text{A.210})$$

$$= \mathcal{O}\left(\frac{1}{n^2}\right), \quad (\text{A.211})$$

where we have defined $P := \sum_{k \in \Omega_{n_0}} |\hat{g}_k|^2$. In line (A.209) we have used the monotonicity of ν_k to bound the components of $S(k) \setminus k$ in the numerator according to the upper bound for the interval $k \in [c_2 - c_1, c_1)$ and applied a lower bound for elements of $S(k) \setminus k$ in the denominator using the interval $k \in [c_2, d)$. The scaling of Eq. (A.211) follows from $m = \mathcal{O}(d/n)$, which implies that $m(m+1)b^2 = \mathcal{O}(n^{-2})$, while the remaining terms in line (A.210) scale as no greater than a constant. To bound the variance, we observe

$$\sum_{k \in \Omega_{n_0}} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \quad (\text{A.212})$$

$$\leq \sum_{k \in \Omega_{n_0}} \frac{[a^2(B_0 - k) + 2kab + (d - B_0 - 2k)b^2]^2 + [(d - 2B_0 - k)b^2 + 2B_0ab]^2}{(a^2 + mb^2)^2} \quad (\text{A.213})$$

$$\leq n_0 \frac{(a^2 B_0 + n_0 ab + db^2)^2 + m(db^2 + 2B_0 ab)^2}{(a^2 + mb^2)^2} \quad (\text{A.214})$$

$$= \mathcal{O}(1), \quad (\text{A.215})$$

which follows identically to the arguments used to reach Eq. (A.211) and the observation that $2k \leq n_0$ for $k \in \Omega_{n_0}$. Furthermore, defining $m' = \lfloor (c_1 - n_0)/n \rfloor$ as the lower bound for the number of aliases of any $k \in \{\ell \in \Omega_n \setminus \Omega_{n_0} : c_1 \leq \ell < c_2\}$ we have

$$\sum_{k \in \Omega_n \setminus \Omega_{n_0}} \frac{\sum_{\ell \in S(k)} \nu_\ell^2}{\left(\sum_{\ell \in S(k)} \nu_\ell\right)^2} \leq \frac{(m+1)(db^2 + 2B_0 ab)^2}{(m'[(d - 2B_0 - k)b^2 + 2B_0 ab] + (m+1 - m')b^2)^2} \quad (\text{A.216})$$

$$\leq \frac{n - n_0}{(m+1)} \frac{(db + 2B_0 a)^2}{\left(\frac{m'}{m+1}[(d - 2B_0 - c_1)b + 2B_0 a] + \left(1 - \frac{m'}{m+1}\right)b\right)^2} \quad (\text{A.217})$$

$$= \mathcal{O}\left(\frac{n^2}{d}\right), \quad (\text{A.218})$$

where in line (A.217) we have split the denominator into frequencies $n_0/2 \leq k < c_1$ and $c_1 \leq k < d$, bounding ν_k for the latter as b^2 . Line (A.218) follows from observing that $m' = \Theta(d/n) \Rightarrow m'/m = \Theta(1)$, and thus the entire second term of line (A.217) is bounded as $\mathcal{O}(1)$. Combining Eqs. (A.215) and (A.218) with Eq. (A.51) for the variance of the Fourier features model gives

$$\text{VAR} = \mathcal{O}\left(\frac{1}{n} + \frac{n}{d}\right). \quad (\text{A.219})$$

It follows that the minimum- $\|\cdot\|_F$ interpolating quantum model with state preparation unitary U satisfying Eq. (A.204) (up to permutations) and data encoded using the Binary encoding strategy will result in benign overfitting as long as the dimensionality of the model scales as $d = \omega(n)$ (i.e., the number of qubits satisfies $n_q = \omega(\log n)$), in which case Eqs. (A.211) and (A.219) characterizing the generalization error of the model both vanish in the large- n limit.

A convenience of this demonstration of benign overfitting in a quantum model was that the state $|\Gamma\rangle$ of Eq. (A.204) incorporates knowledge of the target function (namely the band-limited spectrum Ω_{n_0}). This could be considered a limitation of the approach, as it imposes an inductive bias on the resulting interpolating model f^{opt} , in contrast to other examples of benign overfitting that are more agnostic to the underlying distribution [68, 290]. Future work could reveal choices of $|\Gamma\rangle$ that are more data-independent (no explicit dependence on n_0) but give rise to feature weights ν_k^{opt} with the same desirable properties as Eq. (A.207). As described in Appendix A.1.3.2, the desirable properties include a more weight on all ν_k with $k \in \Omega_{n_0}$ and a long, thin tail of feature weights for all other $k \in \Omega_d \setminus \Omega_{n_0}$.

A.5 Derivation of weight behavior in deep quantum kernels

Here we derive Eq. 3.43. Assuming scalar data $x \in \mathcal{X} \subset \mathbb{R}$ without loss of generality, consider the quantum state embedding

$$\rho(x) = \mathcal{U}^{(L+1)} \circ \left(\bigcirc_{i=1}^L (\mathcal{D}^{(i)}(x) \circ \mathcal{U}^{(i)}) \right) \circ |0\rangle\langle 0|, \quad (\text{A.220})$$

where we have used the shorthand notation $\mathcal{U} \circ \rho = U \rho U^\dagger$ and $\mathcal{D}^{(i)}(x) \circ \rho = D^{(i)}(x) \rho D^{(i)}(x)^\dagger$. Here, $D^{(i)}(x) \in \text{U}(\mathbb{C}^d)$ can be considered a diagonal data-encoding unitary and $U \in \text{U}(\mathbb{C}^d)$ is an arbitrary “trainable” unitary. Let $\mathcal{H} = \mathbb{C}^d$ and recall the following equality for $A, B \in \mathcal{L}(\mathcal{H})$:

$$\text{Tr}(AB) = \text{Tr}((A \otimes B)S), \quad (\text{A.221})$$

where $S := \sum_{i,j=1}^d |ij\rangle\langle ji| \in \mathcal{L}(\mathcal{H} \otimes \mathcal{H})$ is the swap operator. Using Eq. A.221 and $\rho = \rho^\dagger$ we can compute the kernel function

$$k(x, x') = \text{Tr}((\rho(x) \otimes \rho(x'))S) \quad (\text{A.222})$$

$$= \text{Tr} \left[S \left(\mathcal{U}^{(L+1)} \right)^{\otimes 2} \circ \left(\bigcirc_{i=1}^L (\mathcal{D}^{(i)}(x) \otimes \mathcal{D}^{(i)}(x')) \circ (\mathcal{U}^{(i)})^{\otimes 2} \right) \circ |00\rangle\langle 00| \right]. \quad (\text{A.223})$$

To compute this expression's average value with respect to the Haar measure, define the twirling operator $\Phi \in \mathbb{C}(\mathcal{H}_d \otimes \mathcal{H}_d)$:

$$\Phi(A) = \int U^{\otimes 2} A (U^{\otimes 2})^\dagger \mu(dU) \quad (\text{A.224})$$

$$= \frac{1}{d(d^2 - 1)} [(d \text{Tr}(SA) - \text{Tr} A)S + (d \text{Tr} A - \text{Tr}(SA))\mathbb{I}]. \quad (\text{A.225})$$

Then, averaging Eq. A.223 uniformly at random w/r to the Haar measure is equivalent to replacing each $(\mathcal{U}^{(i)} \otimes \mathcal{U}^{(i)})$ with Φ :

$$\mathbb{E}_{U \sim \text{Haar}} k(x, x') = \text{Tr} (S [\Phi \circ (\bigcirc_{i=1}^L (\mathcal{D}^{(i)}(x) \otimes \mathcal{D}^{(i)}(x')) \circ \Phi) \circ |00\rangle\langle 00|]) \quad (\text{A.226})$$

$$= \text{Tr} (S [(\bigcirc_{i=1}^L (\mathcal{D}^{(i)}(x) \otimes \mathcal{D}^{(i)}(x')) \circ \Phi) \circ |00\rangle\langle 00|]) . \quad (\text{A.227})$$

Where in the final line I have applied $\Phi^\dagger(S) = S$ and $\Phi(\mathbb{I}) = \mathbb{I}$. Now set $d = 2$ and define D to derive a closed-form solution recursively. Here we'll consider Z-rotations, $D^{(k)}(x) = \exp(i\alpha_k x Z/2)$, so that

$$D^{(k)}(x) \otimes D^{(k)}(x') = \text{diag} \left(e^{i\alpha_k(x+x')/2}, e^{i\alpha_k(x-x')/2}, e^{-i\alpha_k(x-x')/2}, e^{-i\alpha_k(x+x')/2} \right). \quad (\text{A.228})$$

Defining $\delta = (x - x')$ we have,

$$(\mathcal{D}^{(k)}(x) \otimes \mathcal{D}^{(k)}(x')) \circ S = [D^{(k)}(x) \otimes D^{(k)}(x')] S [D^{(k)}(x) \otimes D^{(k)}(x')]^\dagger \quad (\text{A.229})$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & e^{i\alpha_k \delta} & 0 \\ 0 & e^{-i\alpha_k \delta} & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (\text{A.230})$$

$$:= J^{(k)}, \quad (\text{A.231})$$

and clearly $(\mathcal{D}^{(k)}(x) \otimes \mathcal{D}^{(k)}(x')) \circ \mathbb{I} = \mathbb{I}$. Now prepare for recursion with the following calculations (setting $d = 2$):

$$\Phi(J^{(k)}) = \frac{1}{d(d^2 - 1)} [(d \text{Tr} (S J^{(k)}) - \text{Tr} J^{(k)})S + (d \text{Tr} J^{(k)} - \text{Tr} (S J^{(k)}))\mathbb{I}] \quad (\text{A.232})$$

$$= \frac{1}{6} [((2(2 + 2 \cos(\alpha_k \delta)) - 2)S + (2(2) - (2 + 2 \cos(\alpha_k \delta))\mathbb{I})] \quad (\text{A.233})$$

$$= \frac{1}{3} [(1 + 2 \cos(\alpha_k \delta))S + (1 - \cos(\alpha_k \delta))\mathbb{I}], \quad (\text{A.234})$$

$$\Phi(|00\rangle\langle 00|) = \frac{1}{6} (\mathbb{I} + S). \quad (\text{A.235})$$

Combing Eqs. A.229 and A.232 we can compute the Haar-average of the kernel (Eq. A.227) iteratively. Rearranging the terms, we arrive at the following Lemma:

Lemma 23. Let $\sigma^{(k)} \in D(\mathcal{H} \otimes \mathcal{H})$ be the state of the system after the k -th iteration of “twirl-then-rotate”,

$$\sigma^{(k)} = (\mathcal{D}^{(k)}(x) \otimes \mathcal{D}^{(k)}(x')) \circ \Phi \circ \sigma^{(k-1)}, \quad (\text{A.236})$$

and define

$$w_j := \prod_{\ell=1}^j (1 + 2 \cos(\alpha_\ell \delta)). \quad (\text{A.237})$$

Then,

$$\sigma^{(L)} = \frac{1}{2} \left[\frac{1}{3} \left(\sum_{j=1}^{L-1} \frac{1}{3^j} (1 - \cos(\alpha_j \delta)) w_{j-1} \right) + \frac{1}{3^L} w_{L-1} J^{(L)} \right]. \quad (\text{A.238})$$

Proof. Formally one should prove this inductively, but here I’ll just compute the first few terms:

$$\sigma^{(0)} = |00\rangle\langle 00|, \quad (\text{A.239})$$

$$\sigma^{(1)} := \frac{1}{6} (\mathbb{I} + J^{(1)}) \quad (\text{using Eq A.235}), \quad (\text{A.240})$$

$$\begin{aligned} \sigma^{(2)} &= \frac{1}{6} (\mathcal{D}^{(2)}(x) \otimes \mathcal{D}^{(2)}(x')) \circ \Phi (\mathbb{I} + J^{(1)}) \\ &= \frac{1}{6} (\mathbb{I} + (\mathcal{D}^{(2)}(x) \otimes \mathcal{D}^{(2)}(x')) \circ \Phi (J^{(1)})) \\ &= \frac{1}{6} \left(\mathbb{I} + (\mathcal{D}^{(2)}(x) \otimes \mathcal{D}^{(2)}(x')) \circ \frac{1}{3} [(1 + 2 \cos(\alpha_1 \delta)) S + (1 - \cos(\alpha_1 \delta)) \mathbb{I}] \right) \\ &= \frac{1}{6} \left(\left(1 + \frac{1}{3} (1 - \cos(\alpha_1 \delta)) \right) \mathbb{I} + \frac{1}{3} (1 + 2 \cos(\alpha_1 \delta)) J^{(2)} \right), \end{aligned} \quad (\text{A.241})$$

$$\begin{aligned}
\sigma^{(3)} &= \frac{1}{6} \left(\left(1 + \frac{1}{3}(1 - \cos(\alpha_1 \delta)) \right) \mathbb{I} + \frac{1}{3}(1 + 2 \cos(\alpha_1 \delta))(\mathcal{D}^{(3)}(x) \otimes \mathcal{D}^{(3)}(x')) \circ \Phi(J^{(2)}) \right) \\
&= \frac{1}{6} \left(\left(1 + \frac{1}{3}(1 - \cos(\alpha_1 \delta)) \right) \mathbb{I} \right. \\
&\quad \left. + \frac{1}{3}(1 + 2 \cos(\alpha_1 \delta)) \frac{1}{3} [(1 + 2 \cos(\alpha_2 \delta))J^{(3)} + (1 - \cos(\alpha_2 \delta))\mathbb{I}] \right) \\
&= \frac{1}{6} \left(\left(1 + \frac{1}{3}(1 - \cos(\alpha_1 \delta)) + \frac{1}{3^2}(1 + 2 \cos(\alpha_1 \delta))(1 - \cos(\alpha_2 \delta)) \right) \mathbb{I} \right. \\
&\quad \left. + \frac{1}{3^2}(1 + 2 \cos(\alpha_1 \delta))(1 + 2 \cos(\alpha_2 \delta))J^{(3)} \right). \tag{A.242}
\end{aligned}$$

□

Now we apply Lemma 23 to arrive at

$$\mathbb{E}_{U \sim \text{Haar}} k(x, x') = \text{Tr}(\sigma^{(L)} S) \tag{A.243}$$

$$= \frac{1}{2} \left[\frac{1}{3} \left(\sum_{j=1}^{L-1} \frac{1}{3^j} (1 - \cos(\alpha_j \delta)) w_{j-1} \right) \text{Tr} S + \frac{1}{3^L} w_{L-1} \text{Tr}(J^{(L)} S) \right] \tag{A.244}$$

$$= \frac{1}{3} \left(\sum_{j=1}^{L-1} \frac{1}{3^j} (1 - \cos(\alpha_j \delta)) w_{j-1} \right) + \frac{1}{2} \cdot \frac{1}{3^L} w_{L-1} (2 + 2 \cos(\alpha_L \delta)) \tag{A.245}$$

$$= \frac{1}{3} \left(\sum_{j=1}^{L-1} \frac{1}{3^j} (1 - \cos(\alpha_j \delta)) w_{j-1} + \frac{1}{2 \cdot 3^{L-1}} w_{L-1} \right) + \frac{1}{2 \cdot 3^L} w_L, \tag{A.246}$$

where we've used $\text{Tr} S = 2$ and in the final line we split a term into two parts in preparation for the following calculation. For brevity, denote $\cos(\alpha_j \delta) := z_j$ and therefore $w_j := \prod_{\ell=1}^j (1 + 2z_\ell)$. Define $g : \mathbb{N} \rightarrow \mathbb{R}$ according to

$$g(k) := \sum_{j=1}^k \frac{1}{3^j} (1 - z_j) w_{j-1} + \frac{1}{2 \cdot 3^k} w_k. \tag{A.247}$$

We now show that this has a trivial value, $g(k) = \frac{1}{2}$ for all $k \in \mathbb{N}$. Begin by observing

$$(1 - z_j) w_{j-1} + \frac{1}{2} w_j = (1 - z_j) w_{j-1} + \frac{1}{2} w_{j-1} (1 + 2z_j) \tag{A.248}$$

$$= 3 \cdot \frac{1}{2} w_{j-1} \tag{A.249}$$

Thus,

$$g(k) = \sum_{j=1}^k \frac{1}{3^j} (1 - z_j) w_{j-1} + \frac{1}{2 \cdot 3^k} w_k \quad (\text{A.250})$$

$$= \sum_{j=1}^{k-1} \frac{1}{3^j} (1 - z_j) w_{j-1} + \frac{1}{3^k} (1 - z_k) w_{k-1} + \frac{1}{2 \cdot 3^k} w_k \quad (\text{A.251})$$

$$= \sum_{j=1}^{k-1} \frac{1}{3^j} (1 - z_j) w_{j-1} + \frac{1}{3^k} \left((1 - z_k) w_{k-1} + \frac{1}{2} w_k \right) \quad (\text{A.252})$$

$$= \sum_{j=1}^{k-1} \frac{1}{3^j} (1 - z_j) w_{j-1} + \frac{1}{3^k} \left(3 \cdot \frac{1}{2} w_{k-1} \right), \quad \text{From Eq. A.249} \quad (\text{A.253})$$

$$= \sum_{j=1}^{k-1} \frac{1}{3^j} (1 - z_j) w_{j-1} + \frac{1}{2 \cdot 3^{k-1}} w_{k-1} \quad (\text{A.254})$$

$$= g(k-1). \quad (\text{A.255})$$

Since $g(k) = g(k-1)$, then the function $g(k)$ must be flat and equal to $g(1)$ everywhere. So we evaluate,

$$g(1) = \frac{1}{3} (1 - z_1) + \frac{1}{6} (1 + 2z_1) = \frac{1}{2}. \quad (\text{A.256})$$

Now observe that

$$Z_L := \prod_{\ell=1}^L (1 + e^{i\alpha_\ell \delta} + e^{-i\alpha_\ell \delta}) \quad (\text{A.257})$$

$$= \prod_{\ell=1}^L \left(\sum_{\mathbf{j} \in \{-1, 0, 1\}} e^{ij\alpha_j \delta} \right) \quad (\text{A.258})$$

$$= \sum_{\mathbf{j} \in \{-1, 0, 1\}^L} e^{i(\mathbf{j} \cdot \boldsymbol{\alpha}) \delta}, \quad (\text{A.259})$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_L)$. This now allows us to simplify our desired quantity. From

Eq. A.246 and substituting $g(k) = \frac{1}{2}$,

$$\mathbb{E}_{U \sim \text{Haar}} k(x, x') := \frac{1}{3} \underbrace{\left(\sum_{j=1}^{L-1} \frac{1}{3^j} (1 - \cos(\alpha_j \delta)) w_{j-1} + \frac{1}{2 \cdot 3^{L-1}} w_{L-1} \right)}_{g(L-1)=1/2} + \frac{1}{2 \cdot 3^L} w_L \quad (\text{A.260})$$

$$= \frac{1}{6} + \frac{1}{2 \cdot 3^L} \sum_{\mathbf{j} \in \{-1, 0, 1\}^L} e^{i(\mathbf{j} \cdot \boldsymbol{\alpha}) \delta}. \quad (\text{A.261})$$

For the final step we will rewrite this representation in terms of quantities appearing in a $L = 1$, many-qubit kernel, i.e. the setting discussed in Chapter 3 and Appendix A.3. Fix positive integers L and ω and define

$$Q(\omega) = \{\mathbf{k} : \boldsymbol{\alpha} \cdot \mathbf{k} = \omega, \mathbf{k} \in \{-1, 0, 1\}^L\} \quad (\text{A.262})$$

$$\Omega = \{\omega : \boldsymbol{\alpha} \cdot \mathbf{k} = \omega, \mathbf{k} \in \{-1, 0, 1\}^L\} \quad (\text{A.263})$$

Then we can arrive at,

$$\mathbb{E}_{U \sim \text{Haar}} k(x, x') = \frac{1}{6} + \frac{1}{2 \cdot 3^L} \sum_{\omega \in \Omega} |Q(\omega)| e^{i\omega \delta}, \quad (\text{A.264})$$

and in particular, writing $k(x, x') = \sum_{\omega} \hat{k}_{\omega} e^{i\omega \delta}$ we have

$$\mathbb{E}_{U \sim \text{Haar}} k_{\omega} = \frac{3^{L-1} \delta_0^{\omega} + |Q(\omega)|}{2 \cdot 3^L}. \quad (\text{A.265})$$

It is interesting to compare this result to the case where $L = 1$ but the number of qubits n is allowed to grow arbitrarily. Rewriting Eq. 3.36 to use the notation above, we have

$$\mathbb{E}_{U \sim \text{Haar}} k_{\omega}^{L=1} = \frac{2^{-n} |R(\omega)| + \delta_0^k}{(2^n + 1)}. \quad (\text{A.266})$$

where we fix n and defined the set

$$R(\omega) = \{(\mathbf{i}, \mathbf{j}) : \boldsymbol{\alpha} \cdot (\mathbf{j} - \mathbf{i}) = k, \mathbf{i}, \mathbf{j} \in \{0, 1\}^n\} \quad (\text{A.267})$$

is defined as in Eq. A.182. How does this relate to the $n = 1$, arbitrarily deep data-reuploading kernel solution of Eq. A.265? We can relate the two by relating $Q(\omega)$ and $R(\omega)$. From Appendix A.3, the set $\{(\mathbf{j} - \mathbf{i}) : \mathbf{i}, \mathbf{j} \in \{0, 1\}^n\}$ can be thought of as the image of a map on $\{0, 1\}^n \times \{0, 1\}^n$, with each $\mathbf{k} = (\mathbf{j} - \mathbf{i})$ occurring with multiplicity

$\#(\mathbf{k}) = 2^{n-\|\mathbf{k}\|_1} := D_{\mathbf{k}}$. In other words, for each $\mathbf{k} \in \{-1, 0, 1\}^L$ there are $D_{\mathbf{k}}$ pairs of $(\mathbf{i}, \mathbf{j})$ such that $\mathbf{j} - \mathbf{i} = \mathbf{k}$. This further means that, if we set $n = L$ between the two models, then there is a natural association between every element $\mathbf{k} \in Q(\omega)$ and $D_{\mathbf{k}}$ many elements in $R(\omega)$, i.e.

$$|R(\omega)| = \sum_{\mathbf{k} \in Q(\omega)} D_{\mathbf{k}} = 2^n \sum_{\mathbf{k} \in Q(\omega)} 2^{-\|\mathbf{k}\|_1} \quad (\text{A.268})$$

And so

$$\mathbb{E}_{U \sim \text{Haar}} k_{\omega}^{L=1} = \frac{\sum_{\mathbf{k} \in Q(\omega)} 2^{-\|\mathbf{k}\|_1} + \delta_0^k}{(2^n + 1)}. \quad (\text{A.269})$$

This provides the start of a characterization of the relationship between very deep single-qubit quantum kernels (i.e. the “data-reuploading” setting [95]) and shallow one-qubit quantum kernels. In particular, both kernels are translation invariant (on average with respect to the trainable unitary layers) and are therefore easy to simulate classically by an algorithm that has access to a description of the unitaries appearing in the circuit.

Appendix B

Supplementary material to Chapter 6

This appendix reproduces the appendix of Ref. [4]. I developed the theory in Secs. B.1-B.2. Andy C. Y. Li mainly developed the theory in Sec. B.3, and Andy also performed and analyzed the simulations and hardware experiments for Secs. B.4-B.5.1, while I was responsible for Sec. B.5.2.

B.1 Proof of Proposition 19

The construction for R given in Eq. 6.12 is a tensor product of identical single-qubit response matrices, each of which prescribes a fixed probability for a relaxation event $p(0|1) = q$ and disallows excitation ($p(1|0) = 0$). The outline of the proof is that disallowing excitations results in an R with a block structure such that projection operations commute with the matrix product for strictly upper triangular submatrices of R acting over indices with weight less than w . The tensor structure then allows direct computation of r , and therefore also of r_T .

Q is upper triangular, and therefore R is also upper triangular. Given the tensor structure of R , we have that for indices $i, j \in \{0, 1\}^n$ in the upper triangular set the elements of R are given element-wise by

$$R_{ij} = \begin{cases} q^{w(j)-w(i)}(1-q)^{w(i)}, & \text{for } w(i) < w(j) \\ (1-q)^{w(i)}, & \text{for } i = j \\ 0, & \text{else} \end{cases} \quad (\text{B.1})$$

where the term $q^{w(j)-w(i)}(1-q)^{w(i)}$ represents the probability of $|w(j) - w(i)|$ simultaneous relaxations times the probability of the remaining $w(j) - |w(j) - w(i)| = w(i)$ bits *not* relaxing. We split $\mathbf{R}$ into a diagonal component $\mathbf{R}_0$ and a strictly upper triangular component

$$\mathbf{R}_u = \sum_{k=0}^n \sum_{\ell=k+1}^n |k\rangle\langle\ell| \otimes B_{k\ell}, \quad (\text{B.2})$$

which represents a block matrix partition of $\mathbf{R}$ into $\{B_{k\ell}\}$, each of which contains all elements $(\mathbf{R})_{ij}$ with $w(i) = k$ and $w(j) = \ell$ and has dimensions

$$\dim(B_{k\ell}) = \binom{n}{k} \times \binom{n}{\ell}. \quad (\text{B.3})$$

$\mathbf{R}_u$ is an $(n+1) \times (n+1)$ block matrix, and is strictly upper triangular with respect to this block structure. For any strictly upper triangular $m \times m$ matrix A , we have that $A^m = 0$ [297] and by similar reasoning $\mathbf{R}_u^{n+1} = 0$. Therefore, the Neumann series expansion for $\mathbf{R}^{-1}$ converges in $n+1$ terms and we obtain

$$\mathbf{R}^{-1} = \left(\sum_{k=0}^n (-\mathbf{R}_0^{-1} \mathbf{R}_u)^k \right) \mathbf{R}_0^{-1}. \quad (\text{B.4})$$

In computing $(\mathbf{R}_u)^k$ every column space over basis vectors of weight w depends only on contributions of column spaces over basis vectors of weight less than w (i.e. the columns to the left of the weight- w subspace). Therefore, we further partition $\mathbf{R}_u$ into column spaces $\{\sum_{k+\ell \geq w} B_{k\ell}\}_{w=1}^n$ with basis vectors less than or equal to w and ignore the complementary column space for computing the truncated part of $\mathbf{R}^{-1}$. That is, defining the projector $\pi \equiv P_{0,w}$ and letting $\pi_{\perp} = I - \pi$ be the projector onto the complementary subspace of π we simplify our representation of $\mathbf{R}_u$ as

$$\mathbf{R}_u \equiv \begin{pmatrix} \pi \\ \pi_{\perp} \end{pmatrix} \mathbf{R}_u \begin{pmatrix} \pi & \pi_{\perp} \end{pmatrix} \rightarrow \begin{pmatrix} \pi \mathbf{R}_u \pi & * \\ 0 & * \end{pmatrix} \quad (\text{B.5})$$

The sum in Eq. B.4 may then be computed ignoring the column space of weights greater

than w :

$$(-R_0^{-1}R_u)^k R_0^{-1} \quad (B.6)$$

$$= \begin{pmatrix} (-\pi R_0^{-1}R_u\pi)^k & * \\ 0 & * \end{pmatrix} \begin{pmatrix} \pi R_0^{-1}\pi & * \\ 0 & * \end{pmatrix} \quad (B.7)$$

$$= \begin{pmatrix} (-\pi R_0^{-1}R_u\pi)^k \pi R_0^{-1}\pi & * \\ 0 & * \end{pmatrix} \quad (B.8)$$

since $\pi^2 = \pi$ by definition. Therefore,

$$\pi(R^{-1})\pi = \pi \left(\sum_{k=0}^n (-R_0^{-1}R_u)^k R_0^{-1} \right) \pi \quad (B.9)$$

$$= \sum_{k=0}^w (-\pi R_0^{-1}R_u\pi)^k (\pi R_0^{-1}\pi) \quad (B.10)$$

$$= \sum_{k=0}^w (-\pi R_0^{-1}\pi) (\pi R_u\pi)^k (\pi R_0^{-1}\pi) \quad (B.11)$$

$$\equiv (\pi R\pi)^{-1} \quad (B.12)$$

where the series now terminates at w due to the nilpotency of $\pi R_u\pi$ in Eq. B.7. Line B.11 is simply the series expansion for the inverse of $\pi R\pi = \pi R_0\pi + \pi R_u\pi$, which concludes the proof.

B.2 Proof of Theorem 20 and Corollary 21

The size of the projected subspace that includes all strings of weight less than or equal to w is the binomial sum

$$t(w) = \sum_{k=0}^w \binom{n}{k} \quad (B.13)$$

We can compute the error in the projected readout error correction method directly:

$$\begin{aligned}
|r_T \cdot p'_T - r \cdot p'| &= |((\pi R \pi)^{-1} p'_T)_0 - (R^{-1} p')_0| \\
&= |(\pi (R^{-1}) \pi p'_T)_0 - (R^{-1} p')_0| \\
&= \left| \sum_{j=0}^{t(w)} (R^{-1})_{0j} p'_j - \sum_{j=0}^{2^n-1} (R^{-1})_{0j} p'_j \right| \\
&= \left| \sum_{j=t(w)+1}^{2^n-1} (R^{-1})_{0j} p'_j \right| \\
&= \left| \sum_{k=w+1}^n \left(\frac{q}{1-q} \right)^k \sum_{\ell \in B(w)} p'_\ell \right| \tag{B.14}
\end{aligned}$$

$$\leq \left(\frac{q}{1-q} \right)^{w+1} \tag{B.15}$$

$$\leq (2q)^{w+1} \tag{B.16}$$

where $B(w)$ denotes the set of bitstrings of weight w , and $\pi \equiv P_{0,w}$ (as defined in Appendix B.1). In line B.14 the additional factor of $(1-q)^k$ is found by explicitly computing the first row of R^{-1} : For a binary string $j = j_1 j_2 \dots j_n$ the j -th entry of r is:

$$|r_j| = |(Q^{-1})_{0j_1} (Q^{-1})_{0j_2} \dots (Q^{-1})_{0j_n}| \tag{B.17}$$

$$= \left(\frac{q}{(1-q)} \right)^{w(j)} \tag{B.18}$$

Then with the requirement that $q < 0.5$ we have $(1-q)^{-k} < 2^{-k}$ which completes the proof. To prove Corollary 21 we again explicitly compute r_j taking advantage of the tensor structure of R^{-1} :

$$|r_j| = |(Q_1^{-1})_{0j_1} (Q_2^{-1})_{0j_2} \dots (Q_k^{-1})_{0j_n}| \tag{B.19}$$

$$= \prod_{j_k} \left(\frac{q_k}{(1-q_k)} \right)^{j_k} \tag{B.20}$$

$$\leq \left(\max_k \frac{q_k}{(1-q_k)} \right)^{w(j)} \tag{B.21}$$

which upper bounds the magnitude of any element in the subspace of R^{-1} excluded by truncation of all bitstrings j with $w(j) > w$. Proof of Corollary 21 proceeds identically as

with Theorem 20. Note that the bound given is quite loose, and so the exact expression given in Eq. B.20 may be freely substituted if q_{max} is expected to be significantly larger than a typical q_k .

B.3 Reduction of Algorithm 1 to single bitstrings

Algorithm 1 can be further simplified if we are only interested in a specific element $\ell \in \{0, 1\}^n$ from the distribution. To illustrate the reduction, we construct a set $S_{\ell,w}$ consisting of all $M_{\ell,w} \equiv |S_{\ell,w}|$ basis states with at most a distance w away from ℓ :

$$S_{\ell,w} = \{ m \mid s(m, \ell) \leq w \}. \quad (\text{B.22})$$

We can define a projection operator $P_{\ell,w}$ given by

$$P_{\ell,w} \hat{e}_n = \begin{cases} \hat{e}_n, & \text{if } n \in S_{\ell,w}, \\ 0, & \text{otherwise.} \end{cases} \quad (\text{B.23})$$

It follows from Eq. 6.24 that the prior probability p_ℓ for measuring a computational basis state $|\ell\rangle$ is given by

$$\begin{aligned} p_\ell &= \sum_m \left(\delta_{m,\ell} + \sum_{k=1}^w \left[\left(- \sum_{j=1}^w q^j R_0^{-1} R_j \right)^k \right]_{\ell,m} \right) \\ &\quad \times (R_0)_{m,m}^{-1} p'_m + \mathcal{O}(q^{w+1}). \end{aligned} \quad (\text{B.24})$$

It is straightforward to show that any matrix elements $[(-\sum_{j=1}^w q^j R_0^{-1} R_j)^k]_{\ell,m}$ with $s(m, \ell) > w$ are of order beyond q^w . We can thus use the projection operator $P_{\ell,w}$ to write

$$\begin{aligned} &\left[\left(- \sum_{j=1}^w q^j R_0^{-1} R_j \right)^k \right]_{\ell,m} \\ &= \left[\left(- \sum_{j=1}^w q^j P_{\ell,w} R_0^{-1} P_{\ell,w} R_j P_{\ell,w} \right)^k \right]_{\ell,m} + \mathcal{O}(q^{w+1}) \end{aligned}$$

Defining the truncated operators to be $R_j^{(\ell,w)} = P_{\ell,w} R_j P_{\ell,w}$, we get the approximated probability to be

$$\begin{aligned} \tilde{p}_\ell = \sum_m \left[\delta_{m,\ell} + \sum_{k=1}^w \left(- \sum_{j=1}^w q^j \left(R_0^{(\ell,w)} \right)^{-1} R_j^{(\ell,w)} \right)^k \right]_{\ell,m} \\ \times \left(R_0^{(\ell,w)} \right)^{-1}_{m,m} p'_m. \end{aligned} \quad (\text{B.25})$$

The algorithm to determine $\tilde{p}_\ell$ is similar to that for the full distribution but with R_j being truncated to dimensions $M_{\ell,w} \times M_{\ell,w}$. The time complexity is thus given by $\mathcal{O}(M_{\ell,w}^3)$ which is consistent with the all-zeros bitstring case discussed in Sec. 6.4.

Similarly to the case in Sec. 6.5, the strategy of decomposing R into a set of sparse components $\{R_j\}$ can be combined with alternatives to standard matrix inversion or series approximations to matrix inversion for recovering a specific bitstring ℓ . For instance, the technique of [298] for ϵ -close approximation for specific elements of the solution to $Ax = b$ could be applied to recover p_ℓ with exponential speedup over recovering the entire distribution p provided additional conditions on R . However, the procedure to correct the readout probability for a specific bitstring $\ell \neq 0^n$ will generally require more resources than recovering the all-zeros bitstring. The subspace $S_{\ell,w}$ can be significantly larger than the subspace $S_{0,w}$, given that the majority of elements in $\{0,1\}^n$ have weight close to $\frac{n}{2}$. For example, to implement the algorithm of Sec. 6.4 to recover p_ℓ , R must be projected onto a subspace $\{x : w(\ell) - w_{\min} \leq w(x) \leq w(\ell) + w_{\max}\}$ consisting of strings with weight in $[w_{\min}, w_{\max}]$. This differs from the $\ell = 0^n$ case since the population p'_0 cannot be increased due to the excitation of other bitstrings $x \neq 0$.

From the perspective of readout error mitigation, computing p_0 (or $p(1^n)$, if necessary) is ideal, as it is the string with the fewest neighbors separated by a low-weight error event. Consequently, if one desires to compute the probability of a fixed bitstring $\ell = \ell_1 \ell_2 \dots \ell_n$ at the output of a quantum circuit U , from the perspective of mitigating readout error it is preferable to perform readout rebalancing [219] by appending a single layer of gates to construct $U' = (\sigma_x^{\ell_1} \otimes \sigma_x^{\ell_2} \otimes \dots \otimes \sigma_x^{\ell_n})U$ where σ_x is the pauli-X gate. Then the corrected probability $\tilde{p}_0$ sampled from the output of U' is a maximally efficient approximation to p_ℓ sampled from U .

B.4 Modified perturbative technique

As mentioned in the main text, our technique relies on the assumption that the Neumann series for $\mathbf{R}^{-1}$ converges, namely

$$\left\| \sum_{j=1}^w \mathbf{R}_0^{-1} \mathcal{R}_j \right\| < 1 \quad (\text{B.26})$$

In general, increasing the number of qubits n will uniformly increase the left-hand side of Eq. B.26, as the magnitude of the elements of each matrix $\mathbf{R}_j$ is constant with respect to n but the size of the matrix grows exponentially in n . Unless the characteristic error rate q is reduced simultaneously as n is increased (for example by bounding the product nq), the perturbative approximation of Algorithm 1 will no longer hold. Fig. B.1 shows the effect of increasing q on the performance of Algorithm 1. The failure point of each experiment occurs when the error in $\tilde{p}$ is comparable to the error in p' , which we observed to typically coincide with reaching a characteristic rate q_{max} for which Eq. B.26 was no longer satisfied.

Our technique may be modified slightly so that it still performs well even when the requirement of Eq. B.26 is no longer satisfied, since $\sum_{j=1}^w \mathcal{R}_j$ may still be invertible even when the Neumann series for its inverse does not converge. Fig. B.2 shows the performance of this modified algorithm, and we highlight the fact that this performance improves steadily with increasing w . This improvement comes at the cost of a larger classical computation overhead, but this scenario may still be preferable over completely characterizing $\mathbf{R}$ with an exponentially large diagnostic experiment.

B.5 Additional numerics

B.5.1 Response matrices measured on IBM QPUs

Our numerical experiments in the main text used response matrices generated with a tensor-structure assumption. In reality, the response matrix does not have a tensor structure in general, though the tensor structure could be a good approximation in many cases. We now implement our technique using response matrices measured experimentally on IBM QPUs and demonstrate that the efficacy of the technique for realistic readout errors on NISQ devices.

We measured the response matrices on three different 27-qubit IBM QPUs, namely 'ibm_cairo', 'ibm_hanoi' and 'ibmq_toronto'. The response matrices were estimated by

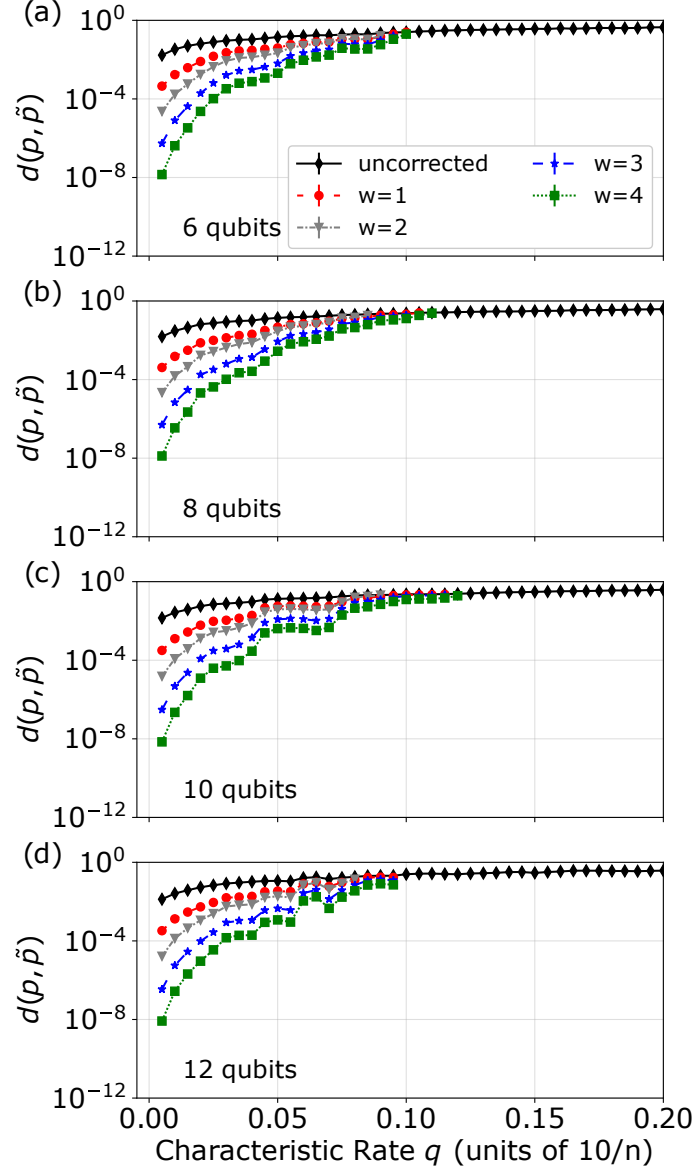


Figure B.1: Failure points for variable number of qubits occur around the same threshold q_{max} in units of $10\frac{q}{n}$. We empirically observe that this failure threshold scales inversely with n , signifying that our technique loses effectiveness when the characteristic error rate q cannot be suppressed as additional qubits are added to the system.

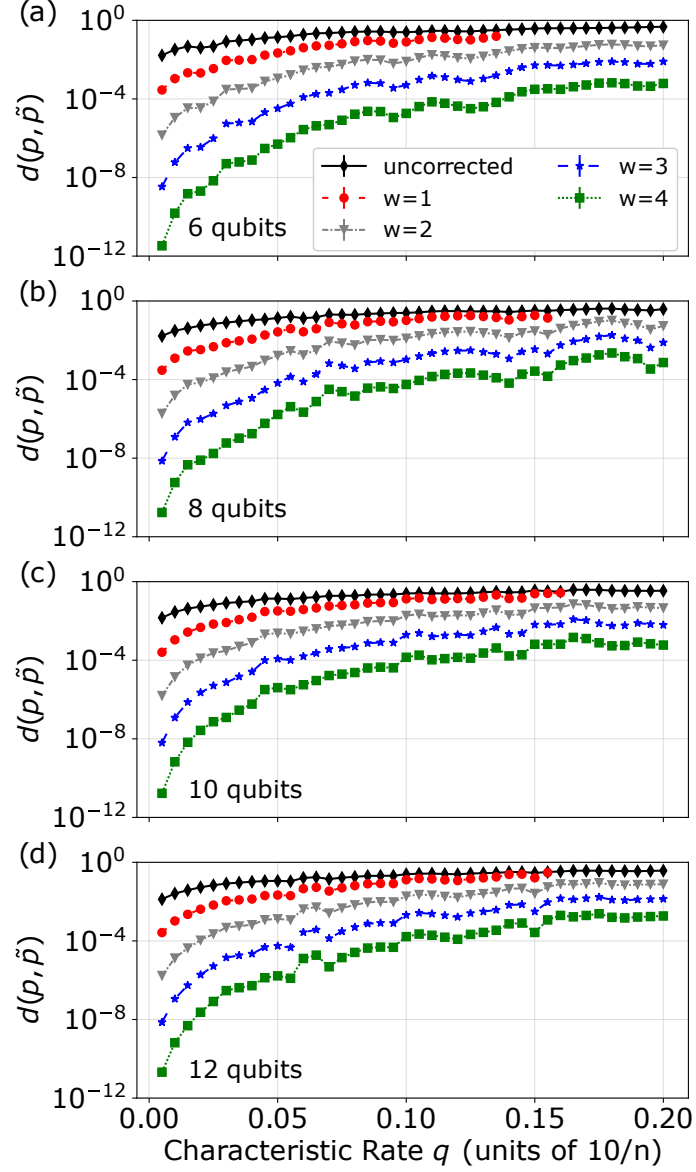


Figure B.2: Our technique can be modified to overcome the limitations shown in Fig. B.1 by exactly inverting $\sum_{j=1}^w \mathcal{R}_j$. By doing so, the failure threshold q_{max} approaches 0.5, indicating that the technique will work for arbitrary $\mathbf{R}$ constructed according to the model we have employed.

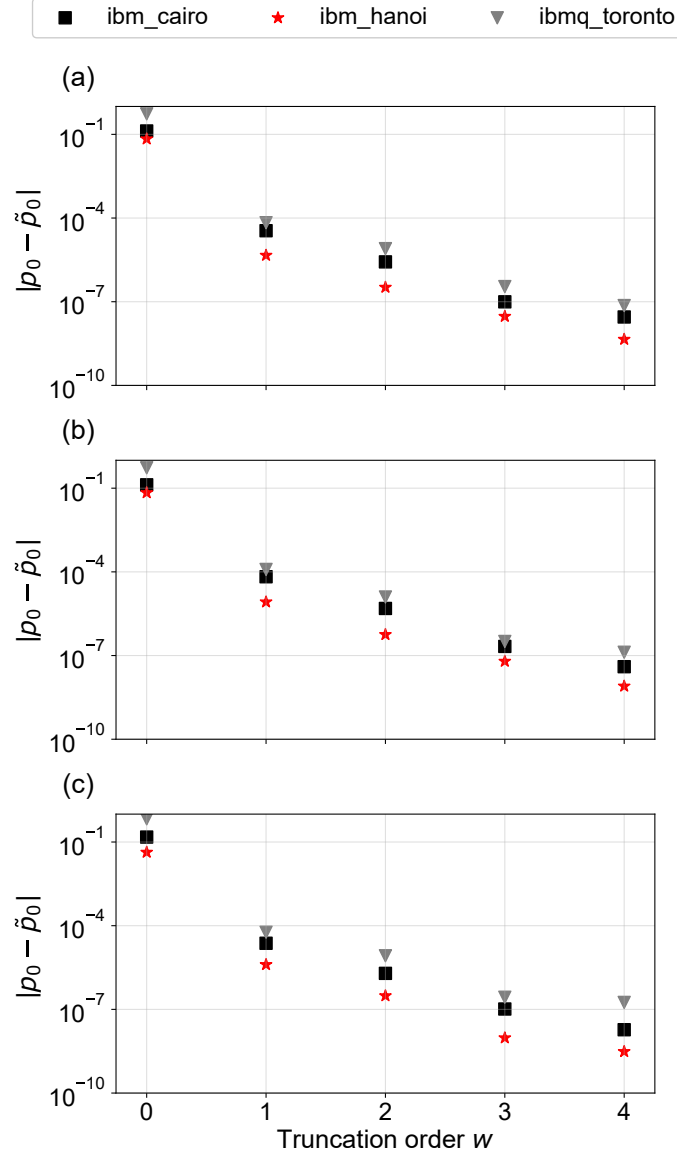


Figure B.3: Performance of all-zeros readout error mitigation with response matrices experimentally determined on IBM QPUs. We demonstrate the performance using three prior distributions, namely, (a) Gaussian distribution, (b) truncated Gaussian distribution and (c) uniform distribution. The details of the distributions are explained in the caption of Fig. 6.1.

preparing computational basis states $|j_1 \dots j_n\rangle$ using a sequence of X gates, i.e., $X_1^{j_1} \dots X_n^{j_n}$, and then determining the output distributions via parallel qubit readout. Each of these measurements requires 2^n different circuit executions with each execution to be repeated several times to generate a distribution. To avoid exponential resource overhead, we estimated R for 12 out of 27 qubits using 10000 shots per matrix element. In particular, we picked qubits 1, 2, 3, 5, 8, 11, 14, 16, 19, 22, 25 and 26 for our experiment. Note that all three QPUs have the same connectivity map.

Fig. B.3 shows the performance of our technique for recovering the all-zeros bitstrings using the response matrices measured on IBM hardware. Similar to the result using a tensor-structure assumption shown in Fig. 6.1, the error is exponentially suppressed in the truncation order w .

B.5.2 Sampling error and comparison to M3

In this section, we provide additional numerical experiments comparing the performance of our method to existing methods. We provide preliminary evidence that our technique for estimating the all-zeros bitstring provides comparable accuracy as the ‘M3’ technique of Ref. [224] for instances tested on 8 qubits. M3 performs readout error correction by operating in a subspace corresponding to bitstrings that were sampled in an experiment with finite repetitions. Thus, M3 corrects an empirical distribution $h' \in \mathbb{R}^{2^n}$ sampled according to the observed bitstring probability vector p' and takes as input a response matrix R sampled from calibration circuits on hardware. To introduce similar sampling error into our technique, we prepared independent qubit bit flip probabilities $\epsilon_k, \eta_k \sim \text{Uniform}(0, q)$ and then prepared estimates $\tilde{\epsilon}_k := B(N, \epsilon_k)/N$, $\tilde{\eta}_k := B(N, \eta_k)/N$ (where B is the binomial distribution) to simulate a series of independent calibration experiments using N circuit repetitions for each of qubits $k = 1, \dots, n$. We then used a sampled response matrix

$$\tilde{R} = \bigotimes_{k=1}^n \begin{pmatrix} 1 - \tilde{\eta}_k & \tilde{\epsilon}_k \\ \tilde{\epsilon}_k & 1 - \tilde{\epsilon}_k \end{pmatrix} \quad (\text{B.27})$$

in place of R for all parts of our algorithm. Similarly, we substituted an empirical estimate h' with components $h'_i = B(M, p'_i)/M$ for the probability vector over observed bitstrings p' to simulate sampling a circuit run for M repetitions. In Fig. B.4, we numerically simulate recovering the all-zeros bitstring for the eight-qubit case using our technique and the M3 technique. The simulation shows that the two techniques give results with a similar accuracy.

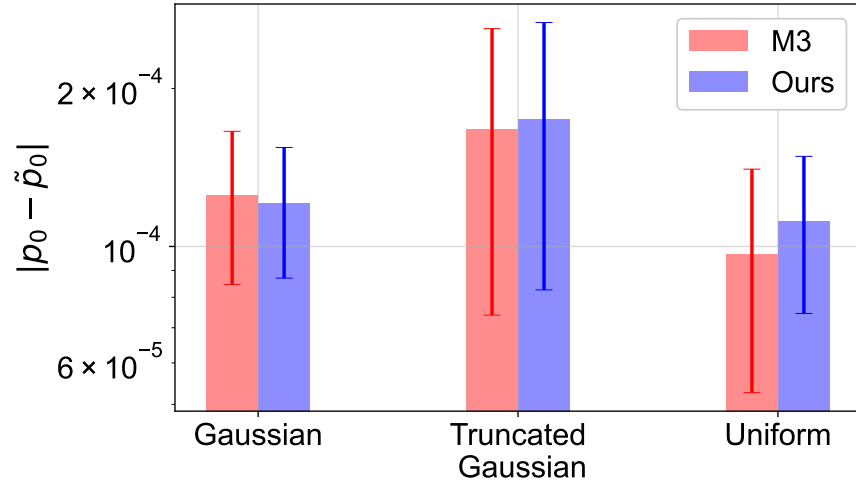


Figure B.4: The accuracy of our technique for recovering the all-zeros bitstring is statistically indistinguishable to that of M3 for the distributions considered in Fig. 6.1 using $n = 8$ qubits, with $M = N = 10^6$ circuit repetitions used for both the calibration experiment and for sampling p' (see main text). The response matrix $\mathbf{R}$ has the same characteristic error rate q as in previous numerics. Error bars denote standard deviation.

Appendix C

Supplementary material to Chapter 7

This chapter reproduces the appendix of Ref. [5]. I was the primary contributor for all material appearing here.

C.1 Certifying optimal qubit assignments

C.1.1 Local depolarizing noise

A significant claim of this work is that under some assumptions on error rates and circuit sizes, performing qubit assignment with respect to the Loschmidt Echo metric F_{LE} will provide reasonable outcomes compared to performing qubit assignment with respect to the fidelity F of a state prepared by applying a noisy implementation of U to an initial state $|0\rangle$. As described in the main text, we use the definitions

$$F_{\text{LE}}(\mathcal{C}_p) = \text{Tr}(|0\rangle\langle 0| \mathcal{E}_{U^\dagger} \circ \mathcal{E}_U(|0\rangle\langle 0|)) \quad (\text{C.1})$$

$$F(\mathcal{C}_p) = \langle 0|U^\dagger \mathcal{E}_U(|0\rangle\langle 0|)U|0\rangle \quad (\text{C.2})$$

where $\mathcal{E}_V$ is a noisy implementation of the unitary V which implicitly depends on the parameters of the circuit $\mathcal{C}_p$. We wish to compare the optimal qubit set S_{LE}^* that maximizes F_{LE} with respect to a hardware connectivity graph (weighted by error rates) to the qubit set S^* that maximizes F :

$$S_{LE}^* = \{S' : S' \subset S, S' = \arg \max_{\mathcal{C}_p \in S} F_{LE}(\mathcal{C}_p')\} \quad (\text{C.3})$$

$$S^* = \{S' : S' \subset S, S' = \arg \max_{\mathcal{C}_p \in S} F(\mathcal{C}_p')\} \quad (\text{C.4})$$

where S describes the complete state space of hardware circuits that realize the operation U in the absence of noise, $S = \{\mathcal{C}_p : U(\mathcal{C}_p) = U\}$. For notational convenience we will typically write $F = F(\mathcal{C}_p)$, $F_{LE} = F_{LE}(\mathcal{C}_p)$. Each metric will be evaluated with respect to a unitary $U = \prod_{i=1}^m U_i$ acting on an input state prepared to $|0\rangle$. We assume without loss of generality that each U_i is up to two-local. We will consider a discrete noise model that can be described by insertion of CPTP maps within an existing circuit, with each map being parameterized by error rates that depend only on the choice of qubit or qubit pair involved in the execution of a gate. We emphasize that an assumption about the device noise is necessary for any comparison between F_{LE} and F since there exists a family of simple noise models for which these two quantities are completely independent, as demonstrated in Appendix C.1.2.2.

We therefore choose a noise model that inserts error channels for each single- and two-qubit gate occurring in a circuit. Fix a device connectivity graph as $G_p = (E_p, V_p)$ and a circuit $\mathcal{C}_p$ acting over n qubits chosen from $N = |V_p|$ qubits with entangling operations subject to the edges provided in E_p . The unitary U implemented by $\mathcal{C}_p$ can be described by a sequence of single qubit gates acting on individual nodes p and two-qubit gates acting on edges $\{p, q\} \in E_p$. The circuit contains a total of n_p gates applied to each node i (where $n_p = 0$ if $\mathcal{C}_p$ does not apply any single qubit gates to i) and n_{ij} gates applied to each pair of qubits $\{p, q\}$.

In the noisy implementation $\mathcal{E}_U$ each single-qubit gate acting on qubit $p \in V_p$ is followed by some fixed, CPTP single-qubit map $\mathcal{N}_p$ also acting on p , and each two-qubit gate on $\{p, q\}$ is followed by a two-qubit channel $\mathcal{M}_{pq}$ acting on $\{p, q\}$. For convenience in the derivation, we will define a single-qubit error rate ϵ_p such that the Kraus operator representation of $\mathcal{N}_p$ is given as

$$\mathcal{N}_p(\rho) = \sum_{A \in K_p} A \rho A^\dagger \quad (\text{C.5})$$

$$K_p = \{\sqrt{1 - \epsilon_p} \mathbb{I}\} \cup \{E_1, \dots, E_{r_p}\} \quad (\text{C.6})$$

$$\sum_{k=1}^{r_p} E_k^\dagger E_k = \mathbb{I} \epsilon_p \quad (\text{C.7})$$

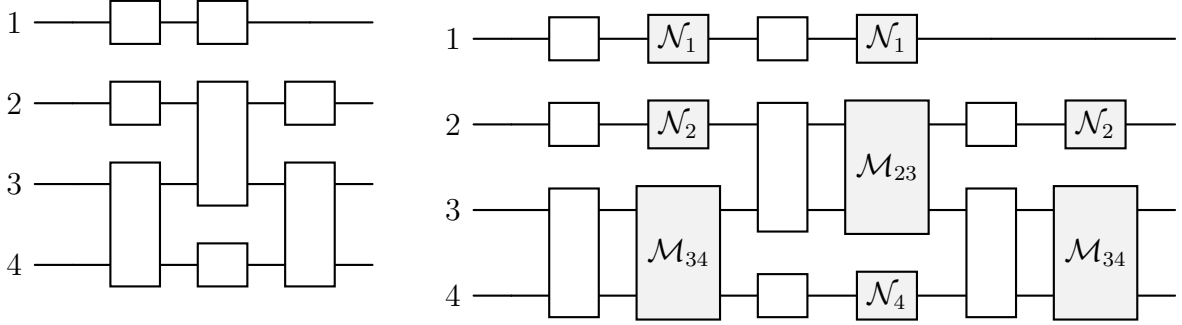


Figure C.1: The two-local noise model here appends channels to each gate depending on which vertex or edge the preceding unitary was applied to. Shaded boxes indicate completely-positive trace preserving operators that are not unitary in general (so that the diagram on the right represents maps acting n -qubit density matrices). If we consider $\mathcal{N}$ to be a channel that applies a non-identity operator with probability ϵ then the state $|\psi_{(i,j)}\rangle$ describes the output of the noisy operator in the case where a non-identity operator was drawn from the Kraus representation of $\mathcal{N}$. While a more general model could consider $\mathcal{N}_p$ and $\mathcal{M}_{pq}$ to be dependent on the gate prior to their insertion point, we do not consider this level of detail in this work.

so that ϵ_p is the probability that $\mathcal{N}_p$ applies a nontrivial error operation. We define η_{pq} to similarly be the probability that $\mathcal{M}_{pq}$ acts with a nontrivial error operation on edge $\{p, q\}$. In this sense, ϵ_p and η_{pq} describe weights in the hardware connectivity graph G_p . Fig. C.1 illustrates an example construction of this noise model.

From $\mathcal{N}$ we define another map

$$\mathcal{N}_p^{sub}(\rho) := \sum_{k=1}^{r_p} \epsilon_p^{-1} E_k \rho E_k \quad (\text{C.8})$$

which contains the (nontrivial) errors that occur when $\mathcal{N}$ is applied to some state ρ . Assuming that the operation U_k is i -th operation applied to qubit p we can then define the (renormalized) mixture over the nontrivial Kraus operators for the i^{th} channel $\mathcal{N}_p$ acting on qubit p during execution of U

$$\rho_{(p,i)} := \mathcal{U}_m \circ \dots \circ \mathcal{U}_{k+1} \circ \mathcal{N}_p^{sub} \circ \mathcal{U}_k \circ \dots \circ \mathcal{U}_1(|0\rangle\langle 0|) \quad (\text{C.9})$$

where we have used the shorthand $\mathcal{U}(\rho) = U\rho U^\dagger$. The state $\rho_{(pq,i)}$ can be defined analogously as the result of inserting $\mathcal{M}_{pq}$ after the i -th operation applied to edge $\{p, q\}$, yielding the mixture over the nontrivial Kraus operators for the i^{th} channel $\mathcal{M}_{pq}$ acting on edge $\{p, q\}$, and $|\psi\rangle = U|0\rangle$. This permits a compact analytical expression for the effect of $\mathcal{E}_U$ and $\mathcal{E}_{U^\dagger}$ for this form of noise model. The state prepared by $\mathcal{E}_U(|0\rangle\langle 0|)$ is given by

$$\begin{aligned} \mathcal{E}_U(|0\rangle\langle 0|) = F_0 & \left(|\psi\rangle\langle\psi| + \sum_{p \in V_p} \sum_{i=1}^{n_p} \frac{\epsilon_p}{1 - \epsilon_p} \rho_{(p,i)} \right. \\ & \left. + \sum_{\{p,q\} \in E_p} \sum_{i=1}^{n_{pq}} \frac{\eta_{pq}}{1 - \eta_{pq}} \rho_{(pq,i)} + \dots \right) \end{aligned} \quad (\text{C.10})$$

where we have omitted terms corresponding to two or more errors and substituted the expression

$$F_0 = \prod_{p \in V_p} (1 - \epsilon_p)^{n_p} \prod_{\{p,q\} \in E_p} (1 - \eta_{pq})^{n_{pq}} \quad (\text{C.11})$$

originally presented in Eq. 7.9 of the main text. The appearance of F_0 here motivates its interpretation as a zero-error contribution to the fidelity F . The remainder of this derivation will consider contributions only up to the single error limit (which will generally still involve high-order terms in ϵ and η). The fidelity of U can be computed directly as

$$\begin{aligned} F = F_0 & \left(1 + \sum_{p \in V_p} \sum_{i=1}^{n_p} \frac{\epsilon_p}{1 - \epsilon_p} \langle \psi | \rho_{(p,i)} | \psi \rangle \right. \\ & \left. + \sum_{\{p,q\} \in E_p} \sum_{i=1}^{n_{pq}} \frac{\eta_{pq}}{1 - \eta_{pq}} \langle \psi | \rho_{(pq,i)} | \psi \rangle + \dots \right) \end{aligned} \quad (\text{C.12})$$

$$= F_0 + F_1 + \dots \quad (\text{C.13})$$

Defining the quantity that describes drawing a nontrivial error for the i^{th} instance of $\mathcal{N}_p$ acting on the reverse circuit:

$$\mathcal{E}_{U^\dagger}^{(p,i)}(|\psi\rangle\langle\psi|) := \mathcal{U}_1^\dagger \circ \dots \circ \mathcal{N}_p^{\text{sub}} \circ \mathcal{U}_k^\dagger \circ \mathcal{U}_{k+1} \circ \dots \circ \mathcal{U}_m^\dagger(|\psi\rangle\langle\psi|) \quad (\text{C.14})$$

and defining $\mathcal{E}_{U^\dagger}^{(pq,i)}$ analogously for $\mathcal{M}_{pq}$, we can similarly compute F_{LE} by applying $\mathcal{E}_{U^\dagger}$ to Eq. C.10 to arrive at

$$F_{\text{LE}} = F_0^2 \left(1 + \sum_{p \in V_p} \sum_{i=1}^{n_p} \frac{\epsilon_p}{1 - \epsilon_p} \left(\langle \psi | \rho_{(p,i)} | \psi \rangle + \langle 0 | \mathcal{E}_{U^\dagger}^{(p,i)} (|\psi\rangle\langle\psi|) | 0 \rangle \right) \right. \quad (\text{C.15})$$

$$\left. + \sum_{\{p,q\} \in E_p} \sum_{i=1}^{n_{pq}} \frac{\eta_{pq}}{1 - \eta_{pq}} \left(\langle \psi | \rho_{(pq,i)} | \psi \rangle + \langle 0 | \mathcal{E}_{U^\dagger}^{(pq,i)} (|\psi\rangle\langle\psi|) | 0 \rangle \right) + \dots \right) \quad (\text{C.16})$$

While there is evidence that F_{LE} contains some of the information about the errors involved in the computation of F , this relationship is still too general to draw any specific comparisons. So to prove Proposition 22 relating F_{LE} and F , we instantiate a noise model for which both local channels $\mathcal{N}_p$ and $\mathcal{M}_j$ are symmetric depolarizing channels (Eq. 7.15) over one and two qubits respectively:

$$\mathcal{N}_p \rightarrow \mathcal{D}_2(\epsilon_p), \quad \mathcal{M}_{pq} \rightarrow \mathcal{D}_4(\eta_{pq}) \quad (\text{C.17})$$

This simple model involving local depolarization is sometimes used in the analysis of random circuits [237, 270] and will be sufficient to demonstrate interesting behavior in the more general problem of qubit assignment. Each depolarizing operation may be commuted with the preceding unitary gate without affecting the behavior of the global state. Specifically, for the case of single-qubit depolarization where

$$\mathcal{N}'_p = \left(\underbrace{\mathbb{I}_2 \otimes \dots \otimes \mathbb{I}_2}_{i-1} \otimes \mathcal{D}_2(\epsilon_p) \otimes \underbrace{\mathbb{I}_2 \otimes \dots \otimes \mathbb{I}_2}_{n-i} \right) \quad (\text{C.18})$$

$$\mathcal{U}'_i = \left(\underbrace{\mathbb{I}_2 \otimes \dots \otimes \mathbb{I}_2}_{i-1} \otimes \mathcal{U}_i \otimes \underbrace{\mathbb{I}_2 \otimes \dots \otimes \mathbb{I}_2}_{n-i} \right) \quad (\text{C.19})$$

then for each single-qubit gate and channel we have that

$$\mathcal{N}'_k \circ \mathcal{U}'_k \circ |\psi\rangle\langle\psi| = \mathcal{U}'_k \circ \mathcal{N}'_k \circ |\psi\rangle\langle\psi| \quad (\text{C.20})$$

and a similar relation holds for each two-qubit gate acted on by $\mathcal{D}_4$. From the freedom in the ordering of each error and each gate occurring in the noisy adjoint circuit $\mathcal{E}_{U^\dagger}$ it follows that

$$\langle 0 | \mathcal{E}_{U^\dagger}^{(p,i)} (|\psi\rangle\langle\psi|) | 0 \rangle = \langle \psi | \rho_{(p,i)} | \psi \rangle \quad (\text{C.21})$$

$$\langle 0 | \mathcal{E}_{U^\dagger}^{(pq,i)} (|\psi\rangle\langle\psi|) | 0 \rangle = \langle \psi | \rho_{(pq,i)} | \psi \rangle \quad (\text{C.22})$$

Rescaling each $\epsilon \rightarrow \frac{3\epsilon}{4}$ and $\eta \rightarrow \frac{15\eta}{16}$ to account for a trivial pauli string in the Kraus representation of $\mathcal{D}_d$ and substituting Eq. C.21 into Eq. C.16 yields

$$\begin{aligned} F_{\text{LE}} &\rightarrow F_0^2 \left(1 + 2 \left[\sum_{p,i} \frac{\epsilon_p}{1 - \epsilon_p} \langle \psi | \rho_{(p,i)} | \psi \rangle \right. \right. \\ &\quad \left. \left. + \sum_{\{p,q\},i} \frac{\eta_{pq}}{1 - \eta_{pq}} \langle \psi | \rho_{(pq,i)} | \psi \rangle \right] + \dots \right) \end{aligned} \quad (\text{C.23})$$

$$= F_0 (F_0 + 2F_1 + \dots) \quad (\text{C.24})$$

We therefore arrive at a relationship between F_{LE} and F in the single-error limit assuming local depolarizing noise. The final step to arrive at Proposition 22 is to take the small error rate limit. We recall that F_{LE} , F , and F_0 are implicitly dependent on a hardware circuit $\mathcal{C}_p = (\mathbf{g}, \mathbf{v}_L, \mathbf{e}_L)$ containing complete information about which qubits and pairs of qubits are acted on by any gate in the gate sequence $\mathbf{g}$. We let $\mathbf{n}$ to be the length- $|V_p|$ vector of gate counts applied to each logical qubit $p \in V_L$ with $(\mathbf{n})_p = n_p = |\{k : 1 \leq k \leq m', (\mathbf{v}_L)_k = p\}|$, and $\mathbf{m}$ to be the length- $|E_p|$ vector of gate counts applied to each available edge $\{p, q\} \in E_p$ with $(\mathbf{m})_{pq} = n_{pq} = |\{k : 1 \leq k \leq m', (\mathbf{e}_L)_k = \{p, q\}\}|$. We similarly define $(\boldsymbol{\epsilon})_p = \epsilon_p$ and $(\boldsymbol{\eta})_{pq} = \eta_{pq}$ to contain the corresponding single- and two-qubit error rates. For some fixed ordering of $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$, the circuit assignment problem (without introducing SWAP overhead) can then be understood as the procedure of ordering the elements of $\mathbf{n}$ and $\mathbf{m}$ under the constraint that the resulting circuit corresponds to the unitary U and respects the hardware device connectivity.

We now study this optimization by taking the limit that $(n_p \epsilon_p) \ll 1$ and $(n_{pq} \eta_{pq}) \ll 1$ for all $p \in V_p$, $\{p, q\} \in E_p$, and arrive at the perturbative approximation

$$F_0 \approx 1 - \mathbf{n} \cdot \boldsymbol{\epsilon} - \mathbf{m} \cdot \boldsymbol{\eta} \quad (\text{C.25})$$

To repeat this procedure for F and F_{LE} , we define

$$(\mathbf{f})_p = \sum_{i=1}^{n_p} \langle \psi | \rho_{(p,i)} | \psi \rangle, \quad p \in V_p \quad (\text{C.26})$$

$$(\mathbf{g})_{pq} = \sum_{p=1}^{n_{pq}} \langle \psi | \rho_{(pq,i)} | \psi \rangle, \quad \{p, q\} \in E_p \quad (\text{C.27})$$

The arrays $\mathbf{f}$ and $\mathbf{g}$ are explicitly dependent on $\mathbf{n}$ and $\mathbf{m}$ respectively, and are therefore also dependent on the choice of qubit assignment. Noting that from Eq. C.8 we then have that $|(\mathbf{f})_p| \leq n_p$, and similarly $|(\mathbf{g})_{pq}| \leq n_{pq}$. This guarantees that terms of the form $(\epsilon_p(\mathbf{f})_p)^2$ vanish at least as quickly as $(\epsilon_p n_p)^2$ and therefore permits the perturbative expansion

$$F_1 \approx \mathbf{f} \cdot \boldsymbol{\epsilon} + \mathbf{g} \cdot \boldsymbol{\eta} \quad (\text{C.28})$$

$$F \approx 1 + (\mathbf{f} - \mathbf{n}) \cdot \boldsymbol{\epsilon} + (\mathbf{g} - \mathbf{m}) \cdot \boldsymbol{\eta} \quad (\text{C.29})$$

$$F_{LE} \approx 1 + 2(\mathbf{f} - \mathbf{n}) \cdot \boldsymbol{\epsilon} + 2(\mathbf{g} - \mathbf{m}) \cdot \boldsymbol{\eta} \quad (\text{C.30})$$

Then under the conditions we imposed, it is clear that under *any* choice of qubit assignment constraints we have

$$S^* = \arg \max_{\mathbf{f}, \mathbf{g}, \mathbf{n}, \mathbf{m}} F = \arg \max_{\mathbf{f}, \mathbf{g}, \mathbf{n}, \mathbf{m}} F_{LE} = S_{LE}^* \quad (\text{C.31})$$

and therefore under any specific constraints that each choice of $\mathbf{n}, \mathbf{m}$ respects hardware connectivity and realizes the unitary U , F and F_{LE} similarly share a maximum. This demonstrates a specific case of Proposition 22 for the local depolarizing noise model.

A number of remarks are necessary to put this result into context. First, the result of Eq. C.31 would not necessarily hold without the assumption of local depolarizing noise, reemphasizing that this result works only for a restricted family of noise models. Notably, this kind of noise model can be enforced on a given hardware circuit using randomized compiling techniques [299], by which the insertion of some redundant single-qubit operations has the effect of converting systematic errors into stochastic depolarizing errors. Furthermore, this result was attained under a restrictive choice of perturbative limit in which

the typical number of errors (ϵn) is very small (though typically a similar assumption is necessary for fault tolerance required by most forms of quantum error correction). We also note that the terms of the form $\mathbf{f} \cdot \boldsymbol{\epsilon}$ may contribute significantly less to the magnitude of F and F_{LE} than $\mathbf{n} \cdot \boldsymbol{\epsilon}$, depending on how the intermediate states in the implementation of U are affected by the various Pauli operators imposed by $\mathcal{D}$.

However, even under these restrictions we have that an optimization based on F_0 will not necessarily arrive at S^* . Even in the case where an experimentalist has perfect knowledge of the depolarizing error rates in a device (computed using RB for example), more detailed knowledge about the effects of the errors (contained in the terms $\mathbf{f}$ and $\mathbf{g}$) is necessary to probe for the optimal qubit assignment. This shortcoming can be demonstrated with a simple example where we suppose G_N has exactly $n - 1$ noiseless qubits to choose from (with the remainder being acted on by maximally depolarizing channels) and U is the circuit preparing the state

$$\frac{1}{\sqrt{2}} (|0^{n-1}\rangle + |1^{n-1}\rangle) |0\rangle \quad (\text{C.32})$$

In this case, F_0 is flat over the space of qubit assignments that contain the $n - 1$ noiseless qubits, but F is maximized only when n^{th} logical qubit is assigned to a noisy device qubit since depolarizing any of the qubits in the entangled subsystem of the state will completely eliminate the coherence of the entangled state.

The truncated expansions for F , F_{LE} , and F_0 can be validated via numerical simulation. Comparing Eq. C.23 to Eq. C.16 gives a single-error approximation for F as

$$F = \frac{1}{2} \left(\frac{F_{\text{LE}}}{F_0} + F_0 \right) + (\text{multiple-error terms}) \quad (\text{C.33})$$

which allows one to approximate F having knowledge of only F_{LE} and F_0 given this specific noise model.

This relation suggests that it may be possible to estimate F via extrapolation techniques if one can reliably compute F_0 , for instance using randomized circuits constructed using the same number of gates acting on each qubit and edge as U . As this extrapolation assumes an underlying noise model consisting of depolarizing noise, an accurate characterization of performance using the relationship of Eq. C.33 might require techniques such as randomized compiling [299].

C.1.2 Alternative noise models

We now discuss the optimization behavior of F , F_0 , and F_{LE} in the context of other noise models. Appendix C.1.2.1 will describe a simple noise model for which all three metrics share an optimal qubit assignment, demonstrating that under relaxed conditions knowledge of F_0 is sufficient to optimize a qubit assignment. Appendix C.1.2.2 will describe a noise model for which F and F_{LE} are entirely uncorrelated.

C.1.2.1 Global depolarizing noise

In contrast to Appendix C.1.1, here we study a noise model for which the optimal qubit assignment according to F_0 , F_{LE} , and F is identical. We consider global depolarizing noise acting on the system with probability proportional to the error rate of the most recently executed gate. While such a model is highly unrealistic for most choices of U , it is well motivated in the case where U is a random circuit drawn uniformly with respect to the Haar measure on $U(2^n)$. Without loss of generality we may decompose U as $U = \prod_{i=1}^m U_i$ where each $U_i \in G_p$ is a member of the hardware gate set for simplicity. Each gate U_i is understood to act on the vertices and edges $(\mathbf{i}, e_i)$. Then for each U_i we apply a depolarizing channel (Eq. 7.15 with $d = 2^n$) acting over all n qubits so that the noisy implementation of U acting on $|0\rangle\langle 0|$ is given by

$$\mathcal{E}_U(|0\rangle\langle 0|) = \mathcal{D}_m \circ \mathcal{U}_m \circ \cdots \circ \mathcal{D}_1 \circ \mathcal{U}_1(|0\rangle\langle 0|) \quad (\text{C.34})$$

where the map $\mathcal{U}_i : \rho \rightarrow U_i \rho U_i^\dagger$ is adopted for convenience. Each channel $\mathcal{D}_i$ has a depolarizing probability given by

$$\epsilon_p = \begin{cases} w_V(i) & \text{if } e_i = \emptyset \\ w_E(e_i) & \text{otherwise} \end{cases} \quad (\text{C.35})$$

where w_V and w_E denote the weights for a vertex or edge respectively according to a fixed noise graph G_N . We can now directly compute F and F_{LE} . Starting from an initial state $|0\rangle\langle 0|$ the result of applying the noisy process $U_{\mathcal{E}}$ can be computed recursively as

$$\mathcal{E}_U(|0\rangle\langle 0|) = \prod_{k=1}^m (1 - \epsilon_k) \rho_m + \left(1 - \prod_{k=1}^m (1 - \epsilon_k)\right) \frac{\mathbb{I}}{d} \quad (\text{C.36})$$

$$= F_0 \rho_m + (1 - F_0) \frac{\mathbb{I}}{d} \quad (\text{C.37})$$

where we have substituted F_0 defined in Eq. 7.9 with $\eta_{pq} = 0$:

$$F_0 = \prod_{k=1}^m (1 - \epsilon_k) = \prod_{k: v_k \in V} (1 - \epsilon_k)^{n_k} \quad (\text{C.38})$$

where n_k counts the multiplicity of each qubit or edge occurring in the instruction set (so that $\sum_k n_k = m$) and $|V| = n$. This is a slight generalization of the result derived in [300]. Assuming that each operation U_i and $U_i^\dagger$ are depolarized by the same amount, we apply $\mathcal{E}_{U^\dagger}$ and then measure the probability of recovering the all zeros bitstring to be

$$F_{\text{LE}} = \frac{d-1}{d} F_0^2 + \frac{1}{d} \quad (\text{C.39})$$

whereas the fidelity of the state prepared by $\mathcal{E}_U$ using Eq. C.37 as

$$F = \frac{d-1}{d} F_0 + \frac{1}{d} \quad (\text{C.40})$$

Then, defining $S^* = \arg \max_{S \subseteq V_p} F_0$ and using the monotonicity of x^2 for $x > 0$ we immediately find

$$S^* = \arg \max_{S \subseteq V_p} F = \arg \max_{S \subseteq V_p} F_{\text{LE}} = S_{\text{LE}}^* \quad (\text{C.41})$$

which is the desired result. Furthermore, the fact that S^* was defined as the qubit assignment maximizing F_0 means that for this noise model, the optimal qubit assignment can be found by maximizing any one of F , F_{LE} , or F_0 . This result is expected since the noise model does not incorporate any information about the structure of U (only the number of gates acting on each qubit at some point in time), and so the assumption of a digital

error model holds and F_0 is an effective proxy for the circuit fidelity. This result further highlights that in the case of sufficiently random circuits F_0 may be an effective optimizer over qubit assignments.

In the above analyses we have defined F_0 in terms of depolarizing channel parameters ϵ_k and η_{jk} , even though though it is not immediately clear that this is the correct parameter to optimize over. In contrast, a depolarizing model can be viewed in terms of a Pauli error e_p by defining the Kraus operators as [265]

$$E_\ell = \sqrt{\frac{e_p}{2^n - 1}} P_\ell, \quad \ell \neq 0 \quad (\text{C.42})$$

$$E_0 = \sqrt{1 - e_p} \mathbb{I} \quad (\text{C.43})$$

$$P_\ell = \sigma_{\ell_1} \otimes \cdots \otimes \sigma_{\ell_n} \quad (\text{C.44})$$

where σ_ℓ is a Pauli matrix with $\ell \in \{0, 1, 2, 3\}^n$. Comparing this action of this channel to a global depolarizing channel with probability p (Eq. 7.15 for $d = 2^n$) gives the relationship

$$e_p = p(1 - 4^{-n}) \quad (\text{C.45})$$

Therefore computation of F_0 from experimentally measured diagnostic data is subject to ambiguities arising from the interpretation of error rates. The quantity $F_0(\hat{\epsilon}, \hat{\eta})$ is equally valid when its arguments represent individual gate infidelities (which are combined in a multiplicative fashion into a total circuit fidelity by F_0) or when its arguments represent probabilities for depolarization after each gate. The parameters in these two error models are related by a constant scalar, but it is straightforward to show that the ranking of qubit assignments with respect to $F_0(c\hat{\epsilon}, c\hat{\eta})$ will not necessarily be preserved with respect to a rescaling factor c applied to the error rates. Define the function

$$f(x, c) = \prod_{i=1}^d (1 - cx_i)^{n_i} \quad (\text{C.46})$$

for any $x \in \mathbb{R}_+^d$, $n \in \mathbb{Z}_+^d$ and $0 \leq c \leq 1$ (the upper bound on c may be loosened as long as $cx_i \leq 1 \forall i$). Then letting the vectors ϵ and ϵ' be the per-qubit error rates for assignments S and S' respectively, it is easy to show that there exist choices ϵ , ϵ' , c , and c' such that $f(\epsilon, c) > f(\epsilon', c)$ but $f(\epsilon, c') < f(\epsilon', c')$. Then letting $c = 1$, $c' = (1 - 4^{-n})$, one can generally

find a distribution of errors such that the optimal assignment differs depending on whether F_0 is computed with respect to the Pauli error rate e_p or the depolarization parameter p . This further complicates the problem of optimal assignments using F_0 computed from calibration data, as there is no a priori choice of scaling that best suits the assignment problem with respect to a given unitary U .

C.1.2.2 Gate-dependent unitary noise

In contrast to Appendix C.1.1, is straightforward to construct a noise model for which $S^* \neq S_{LE}^*$. Consider if $\mathcal{E}_U$ is strictly unitary, such that $\mathcal{E}_U(\rho) = \tilde{U}\rho\tilde{U}^\dagger$ and

$$\tilde{U} = \prod_{i=1}^m W_i U_i \quad (\text{C.47})$$

If the (unitary) noise affecting the reverse circuit is its own adjoint,

$$\mathcal{E}_{U^\dagger}(\rho) = \tilde{U}^\dagger \rho \tilde{U} \quad (\text{C.48})$$

then $F_{LE} = 1$ for any choice of $\{W_i\}$ while $\mathcal{F}(U)$ can take on arbitrary values in $[0, 1]$. Such a model might be relevant to some forms of unitary control error and highlights the need for some contextual understanding of the noise processes involved in a specific hardware when using the technique we introduced in this work. For instance in [301] it was observed that similar unitary error occurred for both $(\text{iSWAP})^{1/2}$ and $(\text{iSWAP})^{-1/2}$ which would result in accumulated errors reflected in F being possibly cancelled out during the computation of F_{LE} .

C.2 Readout error mitigation

We discuss the effect of readout error on $F_{LE}(U)$ and techniques for efficient readout error mitigation. In this case let $\mathcal{E}_U \rightarrow \mathcal{U}$ and $\mathcal{E}_{U^\dagger} \rightarrow \mathcal{U}^\dagger$, but replace the perfect measurement operation with a noisy one. A noisy computational basis measurement can be represented by the POVM $\{P_j\}$ with $j = 0 \dots 2^n - 1$ with each element P_j given by

$$P_j = \sum_k p_{jk} |k\rangle\langle k| \quad (\text{C.49})$$

where the restriction that $\sum_j p_{jk} = 1$ ensures that $\sum_j P_j = I$. The choice $P_j = |j\rangle\langle j|$ recovers a noiseless computational basis measurement. For a fixed j , each p_{jk} in Eq. C.49 can be interpreted as the probability of measuring computational basis state $|k\rangle$ but then recording that result as $|j\rangle$. Substituting P_0 for $|0\rangle\langle 0|$ in Eq. 7.5 of the main text we find that

$$F_{\text{LE}}(U) = \text{Tr}(P_0 U U^\dagger |0\rangle\langle 0| U^\dagger U) \quad (\text{C.50})$$

$$= \text{Tr}(P_0 |0\rangle\langle 0|) \quad (\text{C.51})$$

$$= p_{00} \quad (\text{C.52})$$

which shows that this fidelity estimate will only depend on the probability that the bitstring 0^n is measured correctly by the noisy measurement, and the qubit set $\tilde{Q}^*$ resulting from optimization of $F_{\text{LE}}(U)$ in the presence of readout error will be biased towards maximizing p_{00} . Clearly this does not fully characterize the effect of readout error on a measurement applied to $\mathcal{E}_U(|0\rangle\langle 0|)$, and so $\tilde{Q}^*$ will generally differ from a qubit set Q^* that minimizes the effect of readout error after preparing, say, $\mathcal{E}_U(|0\rangle\langle 0|)$. In an experimental setting this bias can be removed by applying readout error mitigation which can efficiently recover the probability of the all-zeros bitstring from the output of a quantum device [4].

In exploratory experiments we determined that readout error correction assuming uncorrelated bit flip errors for each qubit closely matched the performance of readout error correction assuming correlated bit flips. We performed readout error correction for all experiments by first determining the conditional error probability $p_{01} = p(0|1)$ of observing “0” given a pre-measurement state $|1\rangle$ and $p_{10} = p(1|0)$ of observing “1” given a pre-measurement state $|0\rangle$ for every qubit. We then ran experiments to empirically estimate the probability $p'(j)$ of observing each bitstring $j = 0, \dots, 2^n$ with readout error. Since $p'(j)$ represents a joint probability over single bit outcomes, we then estimated the true probability $p(j)$ of measuring bitstring j using the tensor product of a set of linear maps used to correct each pair of single-bit marginal probabilities $p'(0)$ and $p'(1)$, each having the form

$$\begin{pmatrix} p(0) \\ p(1) \end{pmatrix} = \begin{pmatrix} p(0|0) & p(0|1) \\ p(1|0) & p(1|1) \end{pmatrix}^{-1} \begin{pmatrix} p'(0) \\ p'(1) \end{pmatrix}. \quad (\text{C.53})$$

The error propagated by a linear equation of the form $\mathbf{p} = Q^{-1}\mathbf{p}'$ increases with increasing condition number $\kappa(Q) = \|Q\|\|Q^{-1}\|$, and therefore greater readout error rates $p(x|\neg x)$

will lead to a breakdown in the performance of readout error correction. We therefore rejected qubit assignments for which $\max(p(0|1), p(1|0)) > 0.15$, typically fewer than 5% of assignments on Rainbow but $\sim 27\%$ assignments on Weber. We found that the relationship between F and F_{LE} was consistently strengthened by performing readout error correction. This is partly due to the tendency of asymmetric readout error rates (i.e. $p(0|1) > p(1|0)$) to increase F_{LE} (since the experimentalist infers greater survival likelihood from a larger population of observed 0^n bitstrings) while at the same time decreasing F (in which case the effect of mostly random readout error with high probability increases the distance of the observed bitstring distribution to the bitstring distribution associated with the output of DFE).

C.3 Mapping logical circuits to physical circuits

We present a more detailed description of the qubit assignment problem. We consider a map acting on a logical circuit that returns a hardware circuit performing the same operation as the logical circuit:

$$M : (\mathcal{G}_L^m, V_L^m, E_L^m) \rightarrow (\mathcal{G}_p^{m'}, V_p^{m'}, E_p^{m'}) \quad (\text{C.54})$$

and assume that $\mathcal{C}_L \in (\mathcal{G}_L^m, V_L^m, E_L^m)$ uses exactly n unique qubits taken from V_L . We further subject M to the constraint that $U(\mathcal{C}_p)|0\rangle$ executed in a noiseless setting will prepare the same state as executing the logical circuit. We should allow for the use of additional operations or qubits in the execution of $\mathcal{C}_p$ that do not affect the unitary $U(\mathcal{C}_L)$, and so we let $\mathcal{C}_p \in (\mathcal{G}_p^{m'}, V_p^{m'}, E_p^{m'})$ use exactly N qubits taken from V_p and constrain the map M such that

$$U(\mathcal{C}_L)|0\rangle\langle 0|U(\mathcal{C}_L)^\dagger = \text{Tr}_A (U(\mathcal{C}_p)|0^N\rangle\langle 0^N|U(\mathcal{C}_p)^\dagger) \quad (\text{C.55})$$

for some register A consisting of at most $(N - n)$ qubits. We will consider implementing a specific unitary U . To allow for a layer of abstraction between the logical circuit and physical implementations, we fix a map M and consider the sets of permissible logical and physical circuits to be

$$\mathbf{C}_L = \{\mathcal{C}_L : \mathcal{C}_L \in (\mathcal{G}_L^m, V_L^m, E_L^m), U(\mathcal{C}_L) = U\} \quad (\text{C.56})$$

$$\begin{aligned} \mathbf{C}_p &= \{\mathcal{C}_p : \mathcal{C}_p \in (\mathcal{G}_p^{m'}, V_p^{m'}, E_p^{m'}), \\ &\quad (\exists \mathcal{C}_L \in (\mathcal{G}_L^m, V_L^m, E_L^m) : \mathcal{C}_p = M(\mathcal{C}_L))\} \end{aligned} \quad (\text{C.57})$$

To implement Simulated Annealing we fix the set of possible states of the annealer to be $S = \mathbf{C}_p$, and the algorithm proceeds by iteratively evaluating the cost function for elements drawn from S and updating to another element of S accordingly. We now discuss how the definition of $\mathbf{C}_p$ with respect to a mapping M complicates the implementation of SA.

In our hardware experiments we chose M to be the identity mapping and set $V_p = V_L$, $E_p = E_L$, and $\mathcal{G}_p = \mathcal{G}_L$. As a result, $\mathbf{C}_p = \mathbf{C}_L$. This effectively separates compilation of logical operations into physical operations from the qubit assignment. If we define an update rule such that every element of $\mathbf{C}_L$ may be reached in a finite number of update steps from any other element, then every element of S may similarly be reached after a finite number of steps and so the algorithm is guaranteed to converge by the conditions presented in Ref. [273].

However in general, compilation of logical circuits to hardware will modify the contents of a logical circuit $\mathcal{C}_L$ such that $\mathbf{C}_L \neq \mathbf{C}_p$. Alternatively, error mitigation strategies such as dynamical decoupling [302, 303] or zero noise extrapolation [304, 305] may introduce additional operations to $\mathcal{C}_p$ that do not affect the realized operation $U(\mathcal{C}_p)$. Similarly, “wait gates” which implement the identity operation with fidelity ≤ 1 on hardware can be used to probe T_1 and T_2^* . Under these considerations, a reasonable choice of M might allow any of the following mappings:

$$([R_x(\theta)], [(v_3)], [\emptyset]) \quad (\text{C.58})$$

$$\rightarrow ([R_x(\theta), X, X], [(v_3), (v_1), (v_1)], [\emptyset, \emptyset, \emptyset])$$

$$([R_x(\theta), X, X], [(v_3), (v_1), (v_1)], [\emptyset, \emptyset, \emptyset]) \quad (\text{C.59})$$

$$\rightarrow ([R_x(\theta)], [(v_3)], [\emptyset])$$

Oftentimes logical circuits using elements of V_L with higher degree than those available in V_p , sometimes resulting in the need for SWAP networks to realize an operation $U(\mathcal{C}_L)$ on

hardware, for example:

$$([\text{CNOT}], [\emptyset], [(v_1, v_3)]) \quad (\text{C.60})$$

$$\begin{aligned} &([\text{SWAP}, \text{CNOT}, \text{SWAP}], \\ \rightarrow &[\emptyset, \emptyset, \emptyset], \\ &[(v_1, v_2), (v_2, v_3), (v_1, v_2)]) \end{aligned} \quad (\text{C.61})$$

Allowing for mappings of this form complicates the implementation of SA. Compiling operations from $\mathcal{G}_L$ to $\mathcal{G}_p$ or simplifying the implementation of $\mathcal{C}_L$ (e.g. Eq. C.59) to reduce the depth or number of operations in a hardware circuit involves a compilation problem that is generally NP-hard [306]. On the other hand, allowing for arbitrary maps of the form shown in Eqs. C.58 and C.60 will result in a state space with infinite configurations. In practice this means that the set S^* of optimal qubit assignments may not be accessible by SA or any optimization over qubit assignments.

C.3.1 Assignment of circuits with unconstrained logical connectivity

Algorithm 2 describes the implementation of SA for qubit assignment given a neighborhood function $S : \mathbf{C}_p \rightarrow \mathbf{C}_p$ and a fixed temperature schedule $T : \mathbb{Z} \rightarrow \mathbb{R}$.

In our simulated experiments involving the assignment of logical circuits with all-to-all logical connectivity to a hardware graph with restricted connectivity, we first assumed $V_L = V_p$ and used a transpiler [258] to construct an intermediate logical circuit $\mathcal{C}_L$ respecting (simulated) hardware connectivity. Then we defined the neighborhood $S(\mathcal{C}_L)$ around $\mathcal{C}_L$ for unconstrained logical connectivity (Sec. 7.3.4) according to Algorithm 3. Despite not having access to the complete state space of qubit assignments, this update scheme was able to reliably find high scoring qubit assignments in simulation.

Fig. C.2 demonstrates results of running SA for all-to-all connectivity logical circuits transpiled onto a nearest-neighbor connectivity simulated hardware device. While this approach was successful over random assignment of similar circuits, we emphasize that transpilation by default represents a one-to-many mapping and so further work is necessary to constrain the map M induced by transpilation such that S^* remains accessible by SA. One potential approach is to add a counting argument to M that enumerates over all choices of SWAP networks such that $U(\mathcal{C}_p) = U$, and to incorporate this counter into the SA update step. More generally, the task of noise-aware quantum compilation involving

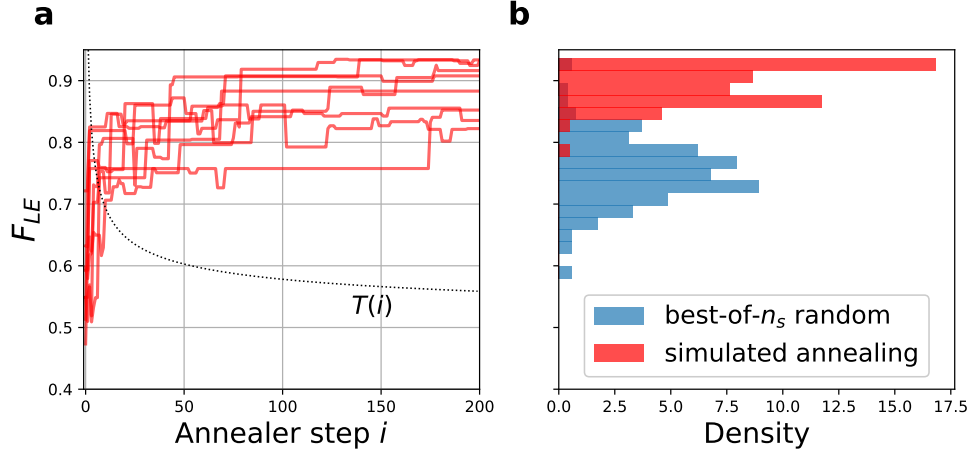


Figure C.2: **a** Sample qubit assignment score at each step of SA for a random 4-qubit circuit with unconstrained logical connectivity. Each F_{LE} is computed for a qubit assignment initially placed onto four simulated qubits before being transpiled [258] into a circuit configuration satisfying nearest-neighbor connectivity on a 5×5 2D grid. A maximum score F_{LE}^* was not identified due to the number of assignment in the state space (the $\binom{25}{4}4!$ possible all-to-all connectivity initial assignments on the grid result in an unknown number of possible transpiled circuits satisfying nearest neighbor connectivity). Experiments were performed with a tuned logarithmic cooling schedule (arbitrarily rescaled for presentation), $T(i) = T_0/(1 + \log(1 + i))$ with $T_0 = 0.08$. **b** 100 trials of SA with transpiled all-to-all connectivity significantly outperforms best-of- n_s random assignment where n_s represents the number of unique states queried by each annealer.

Algorithm 2: Simulated annealing for qubit assignment, N iterations.

Result: $\mathcal{C}_p \in S$

```

1 initialize  $T \leftarrow T(0)$ ,  $\mathcal{C}_p$  ;
2 compute cost  $C = C_{LE}(\mathcal{C}_p)$  ;
3 for  $i = 1 \dots N$  do
4   sample  $\mathcal{C}'_p \in S(\mathcal{C}_p)$  ;
5    $C' = C_{LE}(\mathcal{C}'_p)$  ;
6   if  $C' < C$  then
7     | update  $\mathcal{C}_p \leftarrow \mathcal{C}'_p$ ,  $C \leftarrow C'$  ;
8   else
9     | sample  $x \sim \text{Uniform}(0, 1)$  ;
10    | if  $x < \exp((C' - C)/T)$  then
11      | update  $\mathcal{C}_p \leftarrow \mathcal{C}'_p$ ,  $C \leftarrow C'$  ;
12    | end
13  end
14  update  $T \leftarrow T(i)$  ;
15 end
```

optimization over qubit assignments on noisy hardware with a state space $\mathbf{C}_p$ induced by arbitrary M remains an open problem that we leave to future work.

C.4 Device specifications and additional experiments

We provide detail on the circuits implemented and the hardware devices used for hardware execution. We accessed the Weber and Rainbow superconducting qubit devices through Google’s Quantum Computing Service. Fig. C.3 shows the layout and connectivity of the qubits on each of these devices, colored according to gate fidelity metrics queried from the processor.

Each estimator accumulates error due to finite sample effects, gate infidelity in the DFE measurement subcircuit, and residual readout error after readout error correction has been applied. Without detailed knowledge of these effects we cannot completely characterize the error in each diagnostic, and so we assume every experimental diagnostic is affected equally by these systematic errors. Figs. C.6-C.7 show the results of diagnostic runs for the (random) Clifford conjugation circuits and SWAPnet circuits respectively. For a specific

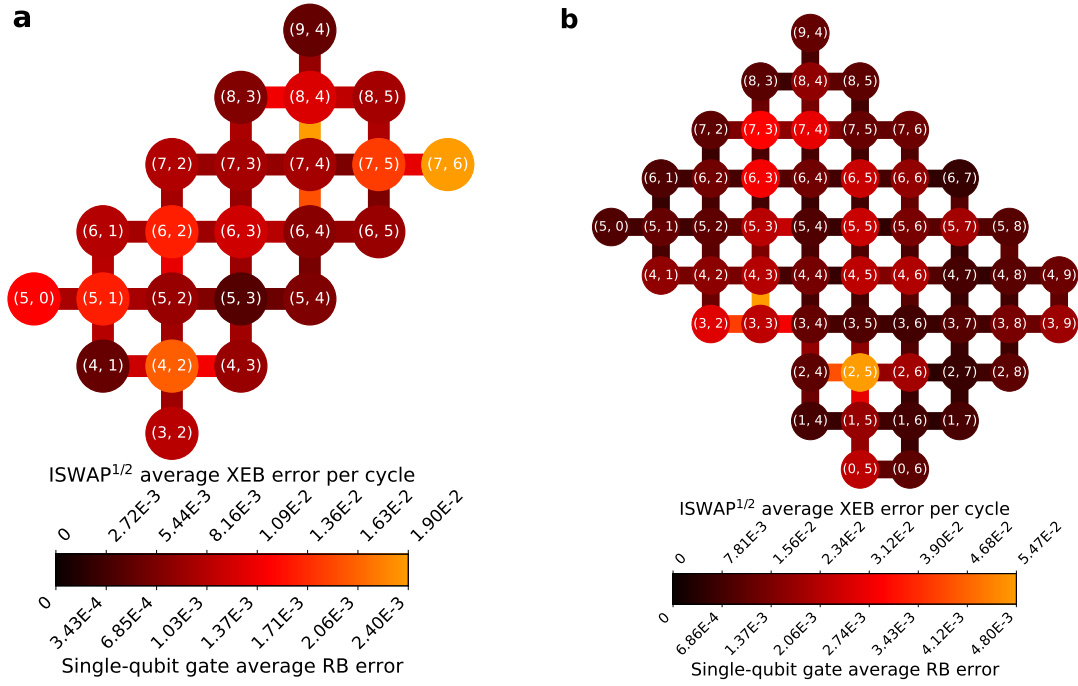


Figure C.3: Hardware layouts for the **a** “Rainbow” 23-qubit device and **b** “Weber” 53-qubit device accessed via Google’s Quantum Computing Service. The color of each node shows single-qubit RB error for gates executed on the corresponding qubit, while the color of each edge shows XEB error per cycle for two-qubit gates executed on the corresponding edge, with darker colors corresponding to lower error rates. Typical error rates were 0.1% and 1% for RB and XEB respectively, and these values were used to construct $\hat{\epsilon}$ and $\hat{\eta}$ to compute F_0 in hardware experiments.

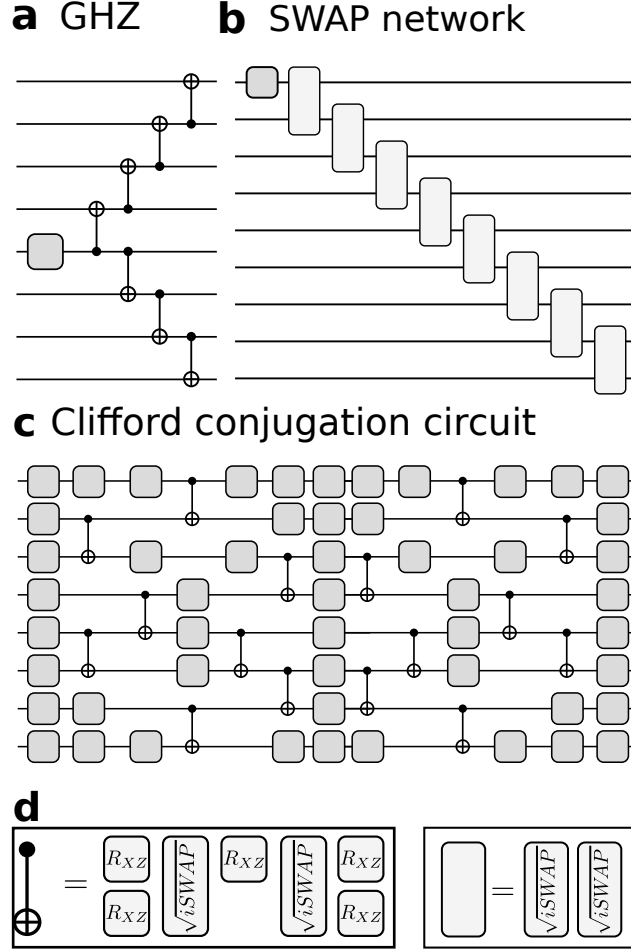


Figure C.4: Circuits implemented for hardware diagnostic experiments for F_{LE} . **a.** The GHZ state $|\psi_{\text{GHZ}}\rangle = |0^n\rangle + |1^n\rangle$ can be prepared on a linear connectivity in $O(n)$ depth. **b.** The SWAP network circuit prepares a state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ that is then swapped into a new register using $4m$ iSWAP gates for some integer m . The fidelity of this circuit can be computed in constant time by performing the projective measurement with elements $\{\mathbb{I} - |\psi\rangle\langle\psi|, |\psi\rangle\langle\psi|\}$. **c.** A sample Clifford conjugation circuit: An arbitrary, random Clifford circuit will typically require significant resources to certify. However the output for circuits of the form $UPU^\dagger$ with $U \in \text{Cl}(2^n)$ and $P \in \{I, X, Y, Z\}^n$ can be verified in constant time. **d.** Decompositions of entangling gates using hardware native gateset for both Rainbow and Weber devices. CNOT gates are decomposed into $\sqrt{i\text{SWAP}}$ and the PhasedXZ(a, b, c) gate defined as $RXZ = R_z(a)R_x(b)R_z(c)$ and implemented with depth-1 on hardware.

Algorithm 3: An implicit definition for the neighborhood around $\mathcal{C}_L$ in terms of an algorithm applied at each iteration of SA.

Result: Sample $\mathcal{C}'_L \in S(\mathcal{C}_L)$

- 1 $\mathcal{C}'_L \leftarrow \mathcal{C}_L$ construct an ordering of qubits $\mathbf{v}$ in $\mathcal{C}_L$;
 - 2 sample $x \in \{1, 2, 3\}$;
 - 3 **if** $x = 1$ **then**
 - 4 | Swap two elements $v_i, v_j \in \mathbf{v}$;
 - 5 **else if** $x = 2$ **then**
 - 6 | “Shift” all $v_i \in \mathbf{v}$ by one row or column
 - 7 **else if** $x = 3$ **then**
 - 8 | Swap an element $v_i \in \mathbf{v}$ with a qubit in $V_p - \{v_k\}_{k=1}^n$;
 - 9 Transpile $\mathcal{C}'_L$ onto hardware topology;
-

family of random Clifford circuits we find that F_{LE} and $\langle F_{\text{LE}}^{\text{rand}} \rangle$ capture the behavior of the fidelity F equally well when the target circuit is itself a random circuit. In contrast, the fidelity of a state transported through a SWAP network is only poorly modeled by F_{LE} computed on random circuits. In both cases, attempting to find a qubit assignment that maximizes F using knowledge of F_0 alone is generally no better than random guessing, and in some cases this strategy is worse than random guessing (Fig. C.7c).

C.4.1 Quantum Fourier Transform

In addition to the circuits described in the main text, we performed SA for qubit assignment for an n qubit Quantum Fourier Transform (QFT) circuit U_{QFT} using $n = 3$ nearest-neighbor connectivity qubits. In general, the state resulting from applying U_{QFT} an arbitrary initial superposition state will have support on an exponential number of Pauli strings when represented in the Pauli basis, so that DFE can not be performed with better than exponential cost. To work around this, we chose to implement U_{QFT} applied to a specific computational basis state $|j\rangle$, $j \in [0, 2^n - 1]$. This requires only a constant number of experiments since the result of this computation a separable state:

$$|\tilde{j}\rangle = U_{\text{QFT}}|j\rangle = \frac{1}{\sqrt{2^n}} \sum_{k=0}^{2^n-1} \exp(-i2\pi jk2^{-n}) \quad (\text{C.62})$$

$$= \frac{1}{\sqrt{2^n}} \bigotimes_{p=1}^n (|0\rangle + \exp(-i2\pi j2^{-p}) |1\rangle) \quad (\text{C.63})$$

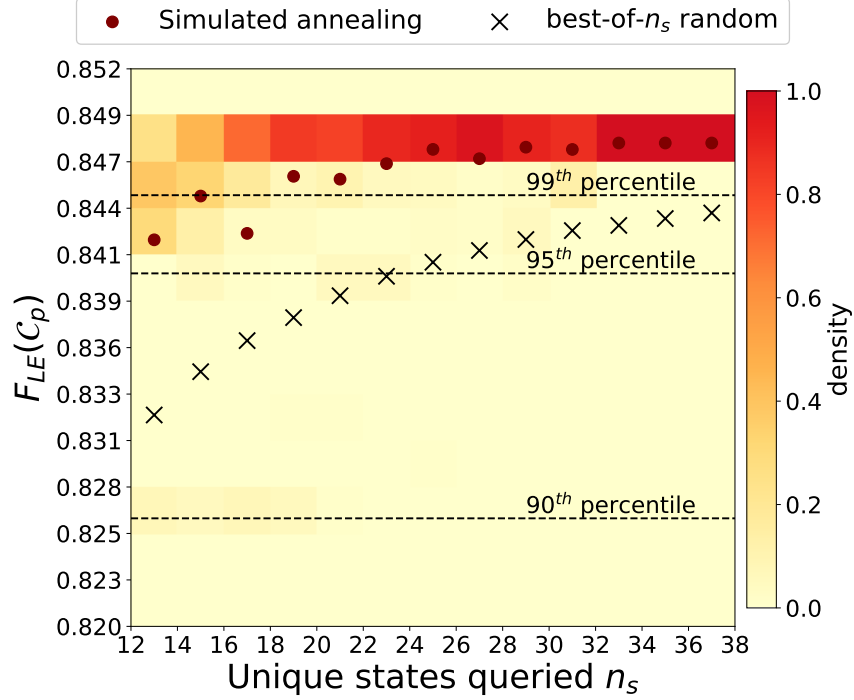


Figure C.5: 2D histogram of best F_{LE} values found over 1000 trials of Simulated Annealing performed on F_{LE} values computed for $U_{QFT}|j\rangle$ with $n = 3$ for the complete state space of 148 circuit assignments on Rainbow. Each F_{LE} is the average over all choices of $j \in [0, 3, 4, 7]$. We used an exponential temperature decay $T_i = T_0 \alpha^i$ ($\alpha = 0.988$, $T_0 = 0.05$). Averaged over all choices of n_s the final F_{LE} achieved by SA outperformed the random sampling scheme by 1.4%.

Defining $|a_p\rangle = 2^{-1/2}(|0\rangle + \exp(-i2\pi j 2^{-p})|1\rangle)$ and setting the measurement parameters

$$\theta_p = 2\pi j 2^{-p} \quad (\text{C.64})$$

$$U_p = R_z(-\theta_p) \quad (\text{C.65})$$

$$M_p = U_p H \sigma_z H U_p^\dagger \quad (\text{C.66})$$

$$(\text{C.67})$$

gives the Pauli decomposition for the state over each local system as

$$|a_p\rangle\langle a_p| = \frac{1}{2}(I + M_p) \quad (\text{C.68})$$

Then given a state ρ output by running QFT on hardware with input $|j\rangle$ we can directly estimate the fidelity of the prepared state as

$$\mathcal{F}(\rho, |\tilde{j}\rangle\langle\tilde{j}|) = \frac{1}{2^n} \text{Tr} \left(\rho \bigotimes_{p=1}^n (I + M_p) \right) \quad (\text{C.69})$$

$$= \frac{1}{2^n} \sum_{\ell \in \{0,1\}^n} \langle \mathbf{M}^\ell \rangle \quad (\text{C.70})$$

where $\mathbf{M}^\ell = M_1^{\ell_1} \otimes M_2^{\ell_2} \otimes \dots \otimes M_n^{\ell_n}$. Therefore a single measurement configuration for the operator $M_1 \otimes M_2 \otimes \dots \otimes M_n$ contains all of the information needed to estimate F up to postprocessing. For each qubit assignment we evaluated F_{LE} and F were evaluated for $j \in [0, 3, 4, 7]$ for the $n = 3$ $U_{QFT}|j\rangle$ states on all possible qubit assignments on the Rainbow device. As with other experiments, F_{LE} with $\tau_b(F_{LE}, F) = 0.72$ outperforms other metrics considered ($\tau_b(\langle F_{LE}^{rand} \rangle, F) = 0.45$, $\tau_b(F_0, F) = 0.15$) when used as a proxy for maximizing F over qubit assignments. Fig. C.5 shows SA for qubit assignment on the state space of length three simple paths on the 23 qubit Rainbow device (148 possible assignments). The performance of SA over a comparable random sampling scheme was less significant compared to the $n = 4$ GHZ qubit assignment. This is likely due to the small size of the state space, for which random guessing is much more likely to produce a performant assignment.

C.4.2 GHZ State Direct Fidelity Estimation

Here we provide measurement setting specifications for performing DFE on the GHZ state according to [277]. The state $|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})$ can be expressed as a linear combination of Pauli matrices using the equalities

$$|0\rangle\langle 0| = \frac{1}{2}(1 + \sigma_z) \quad (\text{C.71})$$

$$|1\rangle\langle 1| = \frac{1}{2}(1 - \sigma_z) \quad (\text{C.72})$$

$$|0\rangle\langle 1| = \frac{1}{2}(\sigma_x + i\sigma_y) \quad (\text{C.73})$$

$$|1\rangle\langle 0| = \frac{1}{2}(\sigma_x - i\sigma_y). \quad (\text{C.74})$$

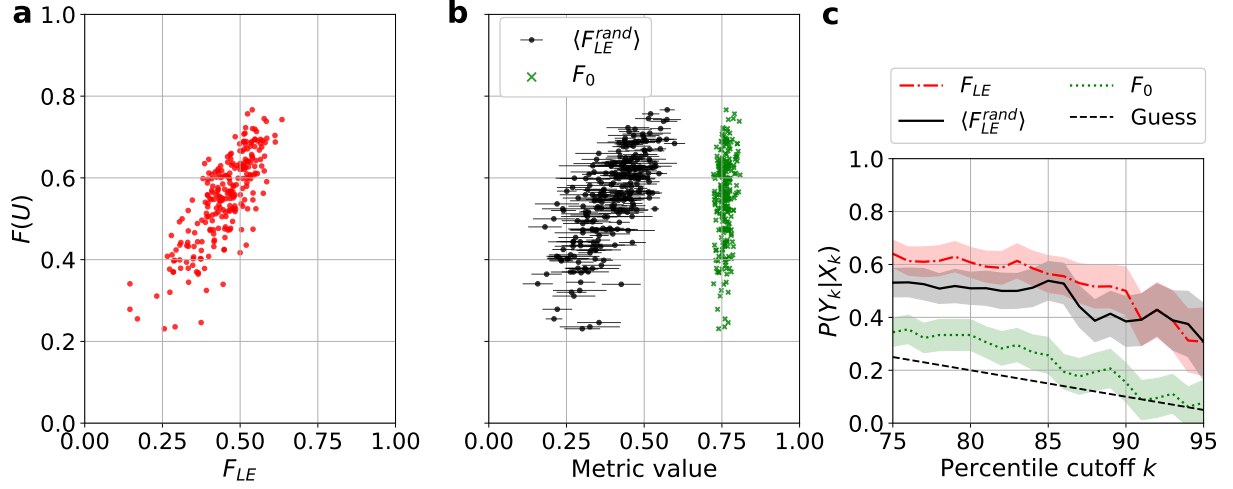


Figure C.6: In the case of random Clifford conjugation circuits, computing F_{LE} for the specific instance of random circuit does not provide significant advantage over computing $\langle F_{LE}^{rand} \rangle$ using a different kind of random circuit shown in **b**. Notably, both metrics still greatly outperform F_0 which was computed using RB and XEB diagnostic error rates. F_0 computed using RB error rates ostensibly describes the fidelity of a random circuit since the effect of Clifford twirling is understood to result in local depolarizing errors, though XEB fidelities do not have a similar straightforward interpretation. **c** $P(F > k | F_{LE} > k)$ and $P(F > k | \langle F_{LE}^{rand} \rangle > k)$ do not differ significantly, demonstrating that either metric is equally useful for picking a qubit assignment to optimize F .

The goal is to measure $F = \langle \psi | \rho | \psi \rangle$ for some experimentally prepared ρ . This fidelity may be expressed as the sum of two terms, $F \equiv f_{XY} + f_Z$, where we define

$$f_Z = \frac{1}{2} \text{Tr} (\rho (|0\rangle\langle 0|^{\otimes n} + |1\rangle\langle 1|^{\otimes n})) \quad (\text{C.75})$$

$$= \frac{1}{2} \text{Tr} \left(\rho \frac{1}{2^n} (1 + \sigma_z)^{\otimes n} \right) + \frac{1}{2} \text{Tr} \left(\rho \frac{1}{2^n} (1 - \sigma_z)^{\otimes n} \right) \quad (\text{C.76})$$

$$= \frac{1}{2^{(n+1)}} \left(\sum_{\ell \in \{0,1\}^n} \langle Z^\ell \rangle (1 + (-1)^{|\ell|}) \right). \quad (\text{C.77})$$

Here, $Z^\ell = \sigma_z^{\ell_1} \otimes \sigma_z^{\ell_2} \otimes \cdots \otimes \sigma_z^{\ell_n}$ with $\ell \in \{0,1\}^n$ is a Pauli operator, and the prefactor of $1/2$ is because ψ contains $\frac{1}{2}(|0\rangle\langle 0|^{\otimes n} + |1\rangle\langle 1|^{\otimes n})$. To estimate f_Z it is therefore sufficient

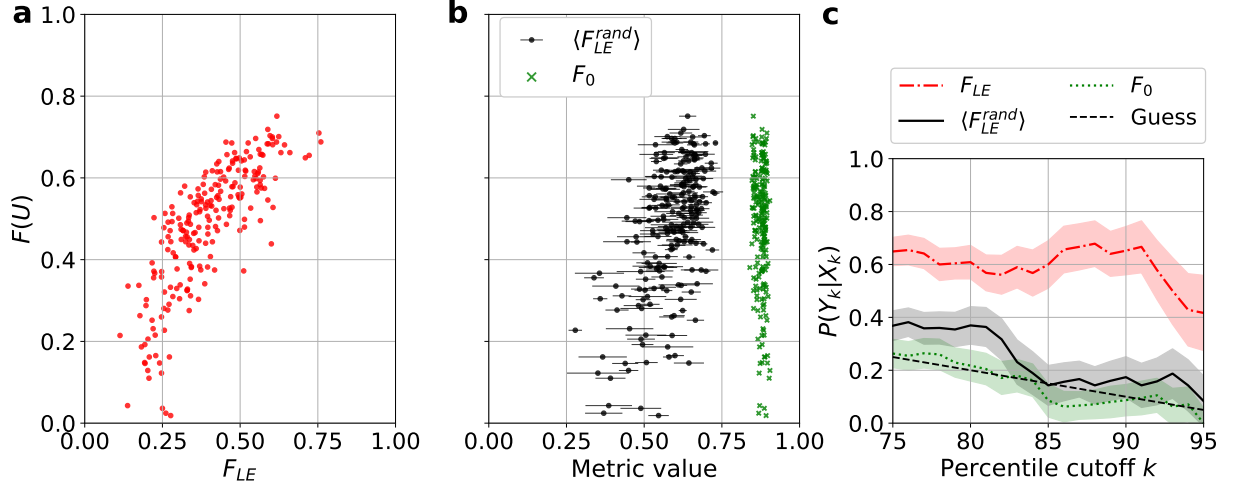


Figure C.7: **a.** Comparison of F to F_{LE} for a 9-qubit SWAP network experiment. Similarly to the GHZ experiment, the fidelity of the state transmitted by the SWAP network was strongly correlated with F_{LE} computed on the same qubit assignment. **b** F_{LE}^{rand} computed using random circuits and F_0 computed using calibration data were not strongly correlated with actual device performance. **c** The conditional success probability $P(Y_k|X_k)$ further emphasizes the performance of F_{LE} compared to F_0 and $\langle F_{LE}^{rand} \rangle$.

to measure in the computational basis for each qubit and perform classical postprocessing. To estimate f_{XY} , we begin by defining

$$M_k = \left(\cos\left(\frac{k\pi}{n}\right) \sigma_x + \sin\left(\frac{k\pi}{n}\right) \sigma_y \right)^{\otimes n} \quad (\text{C.78})$$

$$= U_k H \sigma_z H U_k^\dagger \quad (\text{C.79})$$

where $k = 1, \dots, n$ and $U_k := R_z\left(\frac{k\pi}{n}\right)$. Then one can show that

$$f_{XY} = \frac{1}{2} \text{Tr} \left(\rho (|0\rangle\langle 1|^{\otimes n} + |1\rangle\langle 0|^{\otimes n}) \right) \quad (\text{C.80})$$

$$= \frac{1}{2n} \sum_{k=1}^n (-1)^k \text{Tr}(M_k \rho). \quad (\text{C.81})$$

Therefore to estimate f_{XY} it is sufficient to estimate $\langle M_k \rangle$ using pre-measurement rotations $HU_k^\dagger$ for each choice of $k = 1, \dots, n$. As a consequence we are able to estimate fidelity of the GHZ state $|\psi\rangle$ using $n + 1$ experiments.