Minimax Optimal Estimation of Expectation Values

by

Akshay Seshadri

B.Tech., Indian Institute of Technology Madras, 2016

M.Tech., Indian Institute of Technology Madras, 2016

M.S., University of Colorado Boulder, 2022

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Physics

2025

Committee Members:

Emanuel Knill, Chair

Stephen Becker

Ana Maria Rey

Murray Holland

Oliver DeWolfe

Seshadri, Akshay (Ph.D., Physics)

Minimax Optimal Estimation of Expectation Values

Thesis directed by Dr. Emanuel Knill

There is growing interest in constructing quantum devices that control increasingly large numbers of quantum systems. There is a tremendous need for methods that can measure the quality and other properties of these devices. Such measurements often amount to learning the expectation value of a quantum observable with respect to the quantum state of the device. Learning expectation values is also a key component in many well-known applications such as quantum machine learning and quantum optimization algorithms. Because of the large system sizes involved, it is essential to find methods for learning expectation values that are efficient with respect to the system size. Existing methods that have rigorous guarantees and are practical to implement are often not optimal for observables of interest. Other methods are heuristic, or require the use measurement protocols that are challenging to implement experimentally with current technology.

In this study, we propose an estimation procedure, $\mathbb{T}$he $\mathbb{O}$ptimal $\mathbb{O}$bservable expectation value $\mathbb{L}$earner or $\mathbb{TOOL}$, that can learn the expectation values of observables using the outcomes of any given measurement protocol. We show that there is a seminorm on the set of all observables, which we call the minimax norm, that characterizes the smallest possible estimation error for learning a given observable using the outcomes of a given non-adaptive measurement protocol to a specified confidence level in the worst case over all states. We prove that $\mathbb{TOOL}$ is minimax optimal for every observable by showing that it can achieve an estimation error to within a small constant factor of the minimax norm.

For many applications, one wishes to learn the expectation value of more than one observable from the same experiment. A popular method for learning the expectation values of one or many observables with rigorous guarantees is classical shadows. Classical shadows has near-optimal performance in the worst case over all observables. We prove that $\mathbb{TOOL}$ always performs at least

as well as classical shadows. Moreover, we show by example that TOOL dramatically outperforms classical shadows for many observables of interest. This highlights the need to characterize the optimal performance for the task of simultaneously learning the expectation values of many observables. Under a mild assumption, we give such a characterization using the minimax norm and prove that TOOL is nearly minimax optimal for this task.

We also study the applications of TOOL to fidelity estimation. Using experimental data from a trapped-ion quantum computer, we show that TOOL performs well in practice and matches the estimates obtained from Maximum Likelihood Estimation (MLE), but with rigorous guarantees on the estimation error unlike MLE. We also compare TOOL with another popular method called direct fidelity estimation, which estimates the fidelity by judiciously sampling Pauli observables and measuring them. We show that there is a different importance sampling scheme for Pauli measurements for which TOOL performs as well as, or better than, direct fidelity estimation.

Since TOOL constructs an estimator using only the observable, the measurement protocol, and the confidence level, it provides the flexibility to perform estimation for experiments that have already been performed and experiments that will be performed in the future. Similarly, since the minimax norm can be computed beforehand, it can be used to compare the performance of different measurement protocols and allow minimax optimal design of experiments.

Dedication

To my parents

Acknowledgements

I would like to thank my advisors, Emanuel Knill, Stephen Becker, and Graeme Smith, for their guidance throughout my graduate studies. I thank them for their patience in answering all my questions, and for supporting and encouraging me to work on problems that I found interesting. I would like to thank Scott Glancy and Yanbao Zhang, who have always helped me. I would also like to that Arul Lakshminarayan, Vaibhav Madhok, and R. I. Sujith who mentored me while I was at IIT Madras. I would like to thank all my friends, colleagues, and collaborators, without whom the journey would not have been as fruitful or fun. Last, but not the least, I would like to thank my parents and my extended family who have always stood by me in all my endeavours.

Contents

Chapter

1 Introduction **1**

1.1 Overview and motivation 1

1.2 Summary of the main results 7

2 Preliminaries **14**

2.1 Linear algebra 15

2.2 Quantum states and measurements 20

2.3 Probability theory 25

2.4 Statistics 30

2.5 Convex analysis and optimization 33

3 Classical and quantum distance measures **39**

3.1 Bhattacharyya distance and quantum fidelity 40

3.2 Other distance measures 49

3.3 Relation of Bhattacharyya distance with other distance measures 53

4 Statistical problem **59**

4.1 Mathematical formulation 59

4.2 Juditsky and Nemirovski's estimation procedure 65

4.3 Simplified estimation procedure 68

5 TOOL: A minimax optimal procedure for learning expectation values **84**

5.1 Mathematical formulation . . . 84

5.2 Estimation procedure . . . 90

5.3 Properties of the estimator . . . 95

5.3.1 Minimax optimality . . . 96

5.3.2 Bias . . . 102

5.4 Optimization algorithm . . . 105

6 Application to fidelity estimation **108**

6.1 Comparison with maximum likelihood estimation . . . 109

6.2 Comparison with direct fidelity estimation . . . 111

7 Lower bounds on learning expectation values **118**

7.1 Minimax norm . . . 118

7.2 Lower bound on the error for a given measurement protocol . . . 130

7.3 General lower bound on the estimation error . . . 138

7.4 Lower and upper bounds on the error for shadow tomography . . . 143

8 Miscellaneous applications of the lower bounds **151**

8.1 Randomized measurements . . . 151

8.2 Two no-go theorems . . . 168

Bibliography **174**

Tables

Table

1 A brief comparison of TOOL and classical shadows. M denotes the total number of POVM elements, d denotes the system dimension, and N denotes the number of samples. By optimal, we mean minimax optimal in the worst case over all states. . . 13

2 A dictionary mapping the quantities for the statistical problem given in Sec. 4.1 to the corresponding quantities for the quantum problem of estimating expectation values. The index i varies from 1 to L, where L denotes the number of measurement settings. 91

3 Fidelity estimates and estimation error for a 4-qubit GHZ state, W state, and a cluster state obtained from experimental data for a confidence level of 95%. Estimates are calculated using TOOL and MLE. The error for MLE is obtained from Monte-Carlo (MC) resampling. 110

Figures

Figure

1 A schematic of the procedure for learning the expectation value of an observable $\mathscr{O}$, with respect to the quantum state prepared by the device. Measurements are made on the unknown state, and the measurement outcomes are processed by an estimation procedure to give an estimate for the expectation value. 3

2 Interpretation of the minimax norm as the support function of the constraint-difference set $\Delta\mathscr{C}(\mathfrak{M}, \delta)$. The observable $\mathscr{O}$ is normalized such that $\|\mathscr{O}\|_{\mathrm{HS}} = 1$. The minimax norm measures the distance of the supporting hyperplane $H_{\mathscr{O}} = \{\mathscr{O}' \in \mathbb{S}_d \mid \mathrm{Tr}(\mathscr{O}'\mathscr{O}) = \|\mathscr{O}\|_{\mathfrak{M},\delta}\}$ of $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ from the origin. 127

3 Plot of the rescaled minimax norm (defined in Eq. (8.46)) and the shadow norm of $|Z^n, 0^n\rangle\langle Z^n, 0^n| - \mathbb{I}/2^n$ for uniformly random Pauli measurements as a function of the number of qubits n. The minimax norm is rescaled by a factor of $\sqrt{1-(\delta/2)^{2/N}}$. The analytically computed upper bound of $\sqrt{(9/8)^n}$ on the rescaled minimax norm (dashed line) and lower bound of $\sqrt{(3/2)^n} - 1/2^n$ on the shadow norm (dot-dash line) are plotted for reference. 166

Chapter 1

Introduction

1.1 Overview and motivation

Quantum information science presents an opportunity to perform tasks that are inefficient or sometimes impossible to perform classically. This includes a wide range of tasks, spanning multiple fields of science such as computation [73, 38], communication [5, 18], metrology [40], and cryptography [41]. This has engendered a great amount of research and investment into developing quantum technologies, in particular the development of a general-purpose quantum computer. While it may take some time to build a large-scale fault-tolerant quantum computer that can outperform classical computers, we can still benefit from the theoretical and technological advancements that are achieved on the road to building such a computer.

In order for a quantum protocol to obtain an advantage over a classical protocol, it is necessary to utilize a resource that is quantum mechanical. For example, using an entangled state, one can create nonlocal correlations that no classical system can produce [9, 51, 93, 42]. Such states can be harnessed for tasks like device-independent quantum key distribution [81], quantum teleportation [10, 80], and quantum metrology [40]. Because real quantum devices are noisy, the state prepared by the device can be very different from the resource state that was necessary for implementing the desired quantum information task, which can lead to poor performance or even failure in implementing the task. This motivates us to learn what quantum state was prepared by the device. Unfortunately, it turns out that for learning a d-dimensional quantum state to an error

For the most up-to-date version of this thesis, please consult the arXiv or contact the author.

ε in trace distance with high probability, every procedure needs at least $\Omega(d^2/\varepsilon^2)$ copies of the state in the worst case [46]. For a system of n qubits, the dimension $d = 2^n$ scales exponentially with the number of qubits. Thus, even the most efficient quantum tomography procedures [78, 46] become intractable for large system sizes.

Fortunately, we rarely need the fully reconstructed quantum state in practice. Usually, it suffices to learn some properties of the quantum state depending on the application. One such property, which is focus of our study, is the expectation value of an observable with respect to the quantum state prepared by the device. We focus on the task of learning the expectation values of observables because it is an integral component in several applications such as the characterization of quantum systems [29, 49], entanglement verification using an entanglement witness [23], quantum optimization algorithms [99, 14], quantum machine learning [13, 86, 105], and quantum chemistry [19, 7]. Consequently, there is a vast literature on estimation of expectation values, including both rigorous and heuristic approaches as well as applications to experiments. We refer the reader to recent review articles [4, 37] on learning properties of quantum systems for references and details. In our discussion below, we focus on some recent results on the complexity of estimating expectation values.

There are two steps involved in learning the expectation values of observables: (1) perform measurements on the quantum state, and (2) process the measurement outcomes to obtain estimates of the expectation values. See Fig. 1 for a schematic of this process. Since we only need to learn the expectation value, we can considerably reduce the number of samples required to perform the estimation to a fixed precision. Indeed, for learning the fidelity with a pure quantum state, which is a special case of estimating expectation values, Ref. [33, 26] proposed a randomized Pauli measurement protocol and an estimation procedure that can learn the fidelity to an error ε with high probability using $O(d/\varepsilon^2)$ samples in the worst case. Since the dimension $d = 2^n$ grows exponentially with the number of qubits n, this provides an exponential improvement over quantum tomography.

Building on the idea of performing randomized measurements, Ref. [54] proposed the now well-known technique of classical shadows for simultaneously learning many observables. The classical

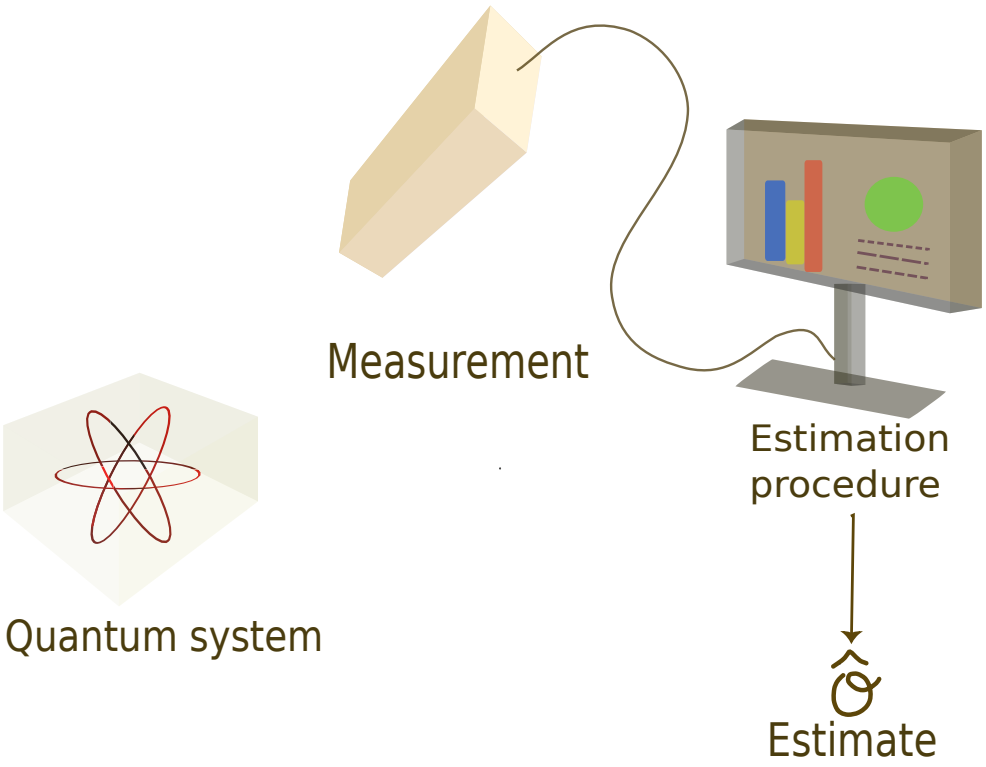

Figure 1: A schematic of the procedure for learning the expectation value of an observable $\mathcal{O}$, with respect to the quantum state prepared by the device. Measurements are made on the unknown state, and the measurement outcomes are processed by an estimation procedure to give an estimate for the expectation value.

shadows method involves randomly selecting a unitary operator from a fixed ensemble of unitary operators, rotating the state by the sampled unitary, performing computational basis measurements, and processing the measurement outcomes to estimate many observables simultaneously. Ref. [54] showed that there is a norm $\|\cdot\|_{\text{shadow}}$ called the shadow norm on the set of observables that depends on the chosen unitary ensemble, such that $O(\max_{1\leq i\leq R}\|\mathcal{O}_i\|^2_{\text{shadow}}\log(R)/\varepsilon^2)$ copies are sufficient to simultaneously learn the expectation values of $\mathcal{O}_1,\ldots,\mathcal{O}_R$. If the unitary ensemble is the set of all global Clifford unitaries, then $\|\mathcal{O}\|_{\text{shadow}}$ scales as the Hilbert-Schmidt norm of $\mathcal{O}$. On the other hand, if $\mathcal{O}$ is a k-local observable (which is an operator that acts non-trivially on at most k qubits) and the unitary ensemble is the set of all local Clifford unitaries, then $\|\mathcal{O}\|_{\text{shadow}}$ is bounded above by 4^k times the operator norm of $\mathcal{O}$. In particular, classical shadows can simultaneously

learn the expectation values of exponentially many low-weight Pauli observables efficiently using local measurements. Similar results were obtained by [31, 25, 16, 60] for learning low-weight Pauli observables or reduced density matrices. This motivated further research on randomized measurement protocols [30], and several generalization and applications of classical shadows to different problems of interest[55, 47, 45, 58, 2, 57].

If we wish to learn the expectation value of a weight-n Pauli observable, which is a Pauli observable that acts non-trivially on n qubits, to an error ε with high probability, classical shadows requires $\Theta(2^n/\varepsilon^2)$ samples using unitary ensemble of global/local Clifford operators. Consequently, for both global and local random Clifford measurements, we need at least $\Omega(2^n/\varepsilon^2)$ samples to simultaneously learn all the Pauli observables to an error ε using classical shadows. This prompts the question of whether it is possible to *efficiently* learn the expectation values of many observables simultaneously. The general problem of simultaneously estimating the expectation values of many observables is called *shadow tomography*, and was introduced by Aaronson [1]. [1] showed that performing entangled measurements on $\widetilde{O}((\log(R))^4 \log(d)/\varepsilon^4)$ copies of the d-dimensional state are sufficient to simultaneously learn the expectation values of $0 \leq \mathcal{O}_1, \ldots, \mathcal{O}_R \leq \mathbb{I}$ to an error ε with high probability, where by writing $\widetilde{O}$ we hide additional logarithmic factors $\log\log(R)$, $\log\log(d)$, and $\log(1/\varepsilon)$. In particular, we can efficiently learn the expectation values of all 4^n Pauli observables using $\mathrm{poly}(n)$ copies of the state. [1] also showed that one needs at least $\Omega(\min\{d^2, \log(R)\}/\varepsilon^2\}$ copies of the state to simultaneously learn the expectation values of R observables, in the worst case over all observables satisfying $0 \leq \mathcal{O}_1, \ldots, \mathcal{O}_R \leq \mathbb{I}$. Aaronson's results have subsequently been improved and generalized [6, 44, 96].

While the results of [1] are appealing from a theoretical standpoint, performing entangled measurements on a large number of copies is incredibly challenging with current technological capabilities. Towards remedying this situation, it was recently shown by [65, 21] that one can simultaneously learn all n-qubit Pauli observables using $\widetilde{O}(\log(d)/\varepsilon^4)$ copies of the state, where entangled measurements are performed only on two copies of the state at a time. This is about as good as we can do, because it was shown that without using entangled measurements, one needs

at least $\Omega(d/\varepsilon^2)$ copies of the state to simultaneously learn all the Pauli observables [56, 20, 21]. This shows that entangled measurements can provide fundamental advantages over performing unentangled measurements. Another avenue to reduce the number of copies needed for estimation is by adaptively choosing the next measurement to perform depending on the outcome observed in the previous experiments. Similar to entangled measurements, adaptive measurements can provide an advantage over non-adaptive measurements for shadow tomography [22, 95, 103, 28].

While both entangled and adaptive measurements are theoretically appealing, performing non-adaptive measurements on a single copy of the state at a time remains the most practical with current technology. Therefore, it is important to know how well one can do for non-adaptive measurements. Towards this end, [54] proved that one needs at least $\Omega(B^2 \log(R)/\varepsilon^2)$ copies of the state to learn the expectation values of R observables to an error ε with high probability in the worst case over all observables satisfying $\max_{1\leq i\leq R} \|\mathcal{O}_i\|_{\text{shadow}} \leq B$ (see Thm. 8.2 and Thm. 8.3 for a precise statement of their result). Later, [69, Thm. (6.3)] derived the lower bound of $\Omega(d \min\{d^2, \log(R)\}/(\varepsilon^2(1+\log(L)/d)))$ on the number of copies of the state needed for shadow tomography using L non-adaptive measurements, in the worst case over all observables with a bound of $d/2$ on the Hilbert-Schmidt norm.

The lower bounds discussed above are for the worst case over all states (since the estimation error ε must be valid no matter what state is prepared by the quantum device), *and* the worst case over all observables with a fixed bound on a norm. Due to the presence of noise or experimental imperfections, it can happen that the state prepared by the device is very different from what we intended to prepare. Since there is no way for us to know what state has been prepared by the device except by performing measurements on the state, it is reasonable to study the worst-case performance over all states. This will inform us on how the number of copies required to perform estimation will scale as a function of the dimension and error, no matter what state is prepared by the device. In contrast, for most applications in quantum information, the observables whose expectation values we want to learn *are known* to us either before or after the measurements are performed. Therefore, it is important to know what is the optimal performance for learning the

expectation values of the specific observables we are interested in, and not the worst case over all observables.

Furthermore, the lower bounds discussed above are derived by allowing a large class of measurement protocols, such as all non-adaptive measurements. While this is an important question from a theoretical standpoint, the measurements that achieve the lower bound may be hard to implement experimentally. For example, global Clifford measurements achieve the worst-case lower bound of [54], but they are challenging to implement for large system sizes with current technology. In practice, the measurements that are implementable/implemented in an experiment depend on several factors such as the observables of interest, the architecture of the quantum computer, current technological limitations, and noise. Hence, it is useful to know how well one can learn the expectation value of a given observable using the outcomes of a measurement protocol that is implementable in an experiment.

We are, therefore, motivated to answer the following basic problem.

Learning Quantum Expectations (LQE):

(1) Given an observable $\mathcal{O}$ and a measurement protocol $\mathfrak{M}$, what is the smallest possible error (over all estimation procedures) for learning the expectation value of $\mathcal{O}$ using the outcomes of $\mathfrak{M}$ with probability greater than $1-\delta$ for all states?

(2) Is there a constructive estimation procedure that can achieve this estimation error to within a constant factor?

LQE asks for a quantification of the "optimal performance" *as a function of* $\mathcal{O}$, $\mathfrak{M}$, and δ, in the worst case over all states. Observe that we quantify the performance in terms of the estimation error instead of the number of copies of the state. This is necessary because **LQE** allows one to specify an arbitrary measurement protocol as an input, and as a result, the number of copies of the state used for estimation is fixed by $\mathfrak{M}$. That said, for many measurement protocols of interest, one can translate between the "smallest estimation error for a fixed number of copies of the state" and the "minimum number of copies of the state needed for a fixed error".

Finding answers to **LQE** is also helpful from a theoretical standpoint, because we can study the performance for other cases of interest. For example, if we know the optimal performance for every observable, we can compute or bound the optimal performance of simultaneously learning many observables. Similarly, if we know the optimal performance for every measurement protocol, we can compute or bound the optimal performance when allowing one to implement a measurement protocol from a given set of measurement protocols.

1.2 Summary of the main results

In this section, we summarize the main results of our study. We begin by providing an answer to **LQE** for non-adaptive measurements. By a (non-adaptive) measurement protocol, we mean a list of positive operator-valued measures (POVMs), along with the number of times each POVM is repeated (see Def. 2.1). We show in Thm. 7.13 that for each measurement protocol $\mathfrak{M}$ and each confidence level $1-\delta \in (0.75, 1)$, there is a seminorm $\|\cdot\|_{\mathfrak{M},\delta}$ on the set of observables, such that for every observable $\mathcal{O}$, $\|\mathcal{O}\|_{\mathfrak{M},\delta}$ gives the optimal estimation error for learning the expectation value of $\mathcal{O}$ to within a factor of $1/c(\delta)$. The constant $c(\delta)$ is defined in Eq. (7.32). For confidence levels greater than or equal to 95%, we have $1/c(\delta) < 5$, and therefore, Thm. 7.13 gives a fairly tight bound on the optimal estimation error.

The precise definition of the seminorm $\|\cdot\|_{\mathfrak{M},\delta}$ is given in Def. 7.1. Intuitively, $\|\mathcal{O}\|_{\mathfrak{M},\delta}$ measures how far apart the expectation value of $\mathcal{O}$ can be with respect to states that are "close enough". Since we only have access to the states through the measurements we perform, we measure the distance between two states through the distance between the probability distributions over measurement outcomes determined by the states. It turns out that the "correct" distance measure to look at is the average Bhattacharyya distance determined by $\mathfrak{M}$ (Def. 3.4). How close the states need to be depends on the chosen confidence level δ and the total number of samples used by $\mathfrak{M}$.

The reason $\|\cdot\|_{\mathfrak{M},\delta}$ is not a norm, and only a seminorm, is that for learning the expectation value of an observable $\alpha\mathbb{I}$ that is a multiple of the identity, the optimal estimation error is zero. This is because the expectation value of $\alpha\mathbb{I}$ with respect to every state is equal to α, and therefore

there is nothing to learn. We show that if we "mod out" all the multiples of identity from the set of all observables, then $\|\cdot\|_{\mathfrak{M},\delta}$ defines a norm (Prop. 7.2). For this reason, we refer to $\|\cdot\|_{\mathfrak{M},\delta}$ as the *minimax norm*, where "minimax" alludes to the fact that $\|\cdot\|_{\mathfrak{M},\delta}$ characterizes the best ("min") performance in the worst case ("max") over all states.

The minimax norm satisfies several desirable properties. Importantly, it can be calculated by convex optimization (Prop. 7.5.3). It is invariant under measurement symmetries (Prop. 7.8.4), and satisfies the data-processing inequality (Prop. 7.9). A more comprehensive list of properties of the minimax norm, including a geometric interpretation, can be found in Sec. 7.1.

The other important aspect of Thm. 7.13 is that there is a constructive estimation procedure, which we call The Optimal Observable expectation value Learner or TOOL, that can achieve an estimator error to a small constant factor of $\|\mathcal{O}\|_{\mathfrak{M},\delta}$. Given $\mathcal{O}$, $\mathfrak{M}$, and $1-\delta$ as inputs, TOOL constructs an estimator for the expectation value of $\mathcal{O}$ using convex optimization (see Box 3). The estimator so constructed is an affine function of the observed frequencies (Prop. 5.12), and can efficiently compute estimates from the data as a result. Since the construction procedure itself does not depend on the experimental data, the estimator can be constructed *either before or after* the measurements are performed. This gives us the flexibility to perform estimation for experiments that will take place in the future, as well as those that have already been completed.

TOOL was introduced in [91, 92] in the context of fidelity estimation, and is obtained by adapting results from statistics [62, 43, 61] to the problem of learning expectation values of observables. For the general statistical problem studied by Juditsky & Nemirovski [62], we present a simplified, less computationally intensive procedure to construct an estimator (Box 2). The estimator we construct satisfies all the guarantees of [62] (Thm. 4.14), and is more amenable to theoretical analysis. Additional results on the estimation procedure and estimation error for the general statistical problem can be found in Sec. 4.3.

We study the application of TOOL to fidelity estimation in Ch. 6. Since the fidelity with a pure state is equal to the expectation value of the projector onto that pure state, fidelity estimation (for a pure target state) is a special case of estimating expectation values. We find that TOOL performs

well on experimental data obtained from a trapped-ion quantum computer. The estimates computed using $\mathbb{TOOL}$ agree well with maximum likelihood estimation (MLE) [53]. On the other hand, through numerical simulations, we show that using (a variant of) bootstrap confidence intervals for MLE can sometimes give erroneous results, unlike $\mathbb{TOOL}$ which is guaranteed to be correct. We also compare $\mathbb{TOOL}$ with direct fidelity estimation (DFE) [33, 26]. We use a slightly different importance sampling scheme for Pauli measurements, and show that for this sampling scheme, $\mathbb{TOOL}$ gives the same or better sample complexity than DFE depending on the target state.

Since the minimax norm characterizes the optimal performance for every observable and every measurement protocol, and $\mathbb{TOOL}$ achieves an estimation error to within a small constant factor of the minimax norm, we can bound the minimax norm to understand how well one can do for different problems of interest. A *lower* bound on the minimax norm will give us a limit on how well every estimation procedure can do, whereas an *upper* bound will show that $\mathbb{TOOL}$ can achieve that error to within a constant factor. We use this strategy to answer a few questions of interest.

We begin with the following question: what is the minimum number of samples needed to learn the expectation value of $\mathcal{O}$? Is there a measurement protocol that achieves this lower bound on the number of samples? The answer to these questions is what one intuitively expects – measuring in the eigenbasis of the observables $\mathcal{O}$ gives the optimal performance. In Thm. 7.15, we derive a lower bound on the number of samples, and in Prop. 7.16, we show that measuring in the eigenbasis is sufficient to achieve this lower bound to within a constant factor.

Next, we fix the measurement protocol $\mathfrak{M}$, and ask what are the observables whose expectation value we can learn to within an arbitrarily small error using outcomes of $\mathfrak{M}$. The answer to this question is also intuitive and familiar to many – we can only learn the expectation values of those observables that are in the linear span of the POVMs in $\mathfrak{M}$. In Prop. 8.12, we give an explicit lower bound on the estimation error for observables that lie outside the linear span of $\mathfrak{M}$.

Next, we study focus on randomized measurement protocols, which have received much attention in the recent literature [30]. All randomized measurements can be expressed using a single *effective* POVM, as explained in Sec. 2.2. Therefore, it suffices to study measurement protocols

where a single POVM is measured many times. Classical shadows, while originally proposed for $\mathscr{U}$-random unitary measurements for unitaries sampled from the set $\mathscr{U}$, were later generalized to all informationally complete POVMs [2, 57]. This includes, for example, single-setting measurements such as SIC-POVMs [83]. The definition of shadow norm for informationally complete POVMs is given in Def. 8.4. Since $\mathbb{TOOL}$ is optimal for every observable and every measurement protocol, it must, in particular, match the performance of classical shadows. We explicitly show in Cor. 8.6 that $\mathbb{TOOL}$ performs at least as well as classical shadows for learning the expectation value of one or many observables.

Since the classical shadows protocol is optimal in the worst case over observables with a fixed bound on the shadow norm [54], there is some observable satisfying this bound on the shadow norm for which no estimation procedure can do better than classical shadows. On the other hand, $\mathbb{TOOL}$ is optimal for every observable. Therefore, we can ask the question if there are observables for which $\mathbb{TOOL}$ performs better than classical shadows. For answering this question, we focus on local non-adaptive measurements on a system of n qubits, as they are one of the easiest types of measurement one can perform in an experiment. By local measurements, we mean that each qubit is measured separately. For local non-adaptive measurements, [54] showed that $\Omega(3^k B^2 \log(R)/\varepsilon^2)$ measurements are necessary to simultaneously learn the expectation values of R k-local observables, in the worst case over all k-local observables whose operator norm is bounded above by B (see Thm. 8.3 for a precise statement of their result). [54] also showed that by choosing the unitary ensemble $\mathscr{U} = \mathrm{Cl}_1^{\otimes n}$ to be the set of local Clifford unitaries on n qubits, classical shadows can simultaneously estimate the expectation values R k-local observables to an error of ε using at most $O(4^k B^2 \log(R)/\varepsilon^2)$ samples. For this reason, we focus on the case where $\mathscr{U}$ is the set of local Clifford unitaries. Since performing a $\mathscr{U}$-random unitary measurement is equivalent to uniformly sampling a weight-n Pauli operator and measuring in its eigenbasis, we refer to such measurements as *uniformly random Pauli measurements*. For uniformly random Pauli measurements, we prove in Cor. 8.8 that there are many observables of interest for which $\mathbb{TOOL}$ can perform exponentially better than classical shadows. This result also holds for learning the expectation values of many observables

simultaneously. Thus, $\mathbb{TOOL}$ can perform exponentially better than both classical shadows as well as the worst-case lower bound of $\Omega(3^k \log(R)/\varepsilon^2)$ for local non-adaptive measurements obtained by [54].

A frequently mentioned feature of classical shadows is that one can choose the observables whose expectation values we wish to estimate after performing the measurements. An important aspect to note here is that while the observables can depend on the measurement protocol (i.e., the $\mathscr{U}$-random unitary measurement that was performed), they must not depend on the measurement data. We remark that this feature is not unique to classical shadows or $\mathscr{U}$-random unitary measurements, and is a common property of many statistical procedures. For example, one can perform tomography using any informationally complete measurement, and store a classical description of the reconstructed state for future use. One can later choose the observables (independent of the observed data) whose expectation values needs to be estimated, and estimate the expectation values of these observables using the reconstructed state. $\mathbb{TOOL}$ shares the same feature, and works much better than tomography for estimating expectation values. One can implement an arbitrary measurement protocol in an experiment, and store the measurement outcomes for future use. The observables can be chosen later (independent of the observed outcomes), and $\mathbb{TOOL}$ can give optimal performance for all the chosen observables. This contrasts with classical shadows that can be far from optimal for many observables.

Because classical shadows and the worst-case lower bounds derived in [54] can give sub-optimal results for many observables, we are motivated to derive bounds on the optimal performance of shadow tomography using non-adaptive measurements. First, we note that every estimation procedure that can learn the expectation values of any given observable, and in particular $\mathbb{TOOL}$, can be extended to simultaneously learn the expectation values of many observables following the strategy in Box 7. Box 7 is a simple application of the union bound, and many estimation protocols used in quantum information (such as classical shadows) use such a strategy to simultaneously learn the expectation values of many observables. We show in Prop. 7.18 that for estimation procedures that simultaneously learn the expectation values of the observables $\mathscr{O}_1, \ldots, \mathscr{O}_R$ using the

outcomes of a given measurement protocol $\mathfrak{M}$ to a confidence level of $1-\delta \in (0.5, 1)$ using the union bound, $\max_i \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}$ characterizes the optimal estimation error. Furthermore, $\mathbb{TOOL}$ achieves this estimation error to within a factor of $1/\mathscr{c}(\delta)$, showing that $\mathbb{TOOL}$ is also minimax optimal for shadow tomography amongst estimation procedures that use the union bound (Prop. 7.17). Similarly, we characterize the optimal estimation error for simultaneously learning the expectation values of many observables by allowing one to implement a measurement protocol from a given set of measurement protocols in Prop. 7.19.

Since the minimax norm can be used to characterize the optimal estimation error for learning one or many observables for any given measurement protocol, it can be used to compare the performance of different measurement protocols. One can also use the minimax norm to perform minimax optimal experimental design, by optimizing the minimax norm over a given set of measurement protocols.

The main drawback of $\mathbb{TOOL}$ is the *computational* complexity for constructing the estimator and computing the minimax norm (estimation error). If M denotes the total number of POVM elements in the measurement protocol and d denotes the system dimension, then in the worst case scenario, our implementation of $\mathbb{TOOL}$ given in Sec. 5.4 needs $O(Md^2) + O(d^3)$ time and $O(Md^2)$ memory to perform the optimization to construct the estimator and compute the minimax norm for a single observable. Once the estimator is constructed, the estimator can compute estimates from N measurement outcomes in $O(N)$ time. Since the estimator can be reused as many times as necessary (for a given $\mathcal{O}$, $\mathfrak{M}$, and $1-\delta$), the costly computation only needs to be performed once.

While the worst-case computational complexity of $\mathbb{TOOL}$ is bad, it is possible to improve the computational complexity for special cases of interest. For example, for fidelity estimation, there is a 2-outcome POVM (which models a large class of measurement protocols), for which $\mathbb{TOOL}$ can construct the estimator and compute the minimax norm in $O(1)$ time and memory, *independent* of the system dimension (see Prop. 6.1). We leave the problem of devising efficient algorithms for constructing the estimator using $\mathbb{TOOL}$ for other cases of interest as a problem for future research.

On the other hand, the worst-case computational complexity of classical shadows for a single

observable is $O(Nd^2)$ time and $O(Nd^2)$ memory, since the $d \times d$ observable as well as the shadows need to be stored and expectation values need to be computed. If we focus on the unitary ensembles of global or local Clifford unitaries, then the memory can be reduced to $O(d^2)$ instead of $O(Nd^2)$ since the shadows are stabilizer states and can be stored efficiently, and only the observable needs to be stored. The time complexity remains $O(Nd^2)$ in this case because the observable whose expectation values needs to be estimated may not have any classically efficient description. However, when the observable can be written as a linear combination of polynomially many (in the number of qubits) projectors onto stabilizer states, classical shadows can be implemented in $O(N \operatorname{polylog}(d))$ time and memory.

We present a brief comparison of TOOL and classical shadows in Tab. 1.

	TOOL	Classical shadows [54]
Optimal for every observable	✔	✗
Optimal for shadow tomography	✔[a]	✗
Optimal for every measurement protocol	✔	✗
Estimation error	$\sim \lVert\cdot\rVert_{\mathfrak{M},\delta}$	$\sim \lVert\cdot\rVert_{\text{shadow}} \sqrt{\frac{\log(2/\delta)}{N}}$ [b]
Worst-case computational complexity	$O(Md^2) + O(d^3)$ once, $O(N)$ afterwards	$O(Nd^2)$

[a] Optimal amongst estimation protocols that use union bound to simultaneously learn in l_∞-norm
[b] For N repetitions of an informationally complete POVM

Table 1: A brief comparison of TOOL and classical shadows. M denotes the total number of POVM elements, d denotes the system dimension, and N denotes the number of samples. By optimal, we mean minimax optimal in the worst case over all states.

Chapter 2

Preliminaries

In this chapter, we review some definitions, concepts, and results that are used in this thesis. We also introduce some notation that is used throughout this thesis.

Sets are denoted by the calligraphic upper case letters such as $\mathscr{X}, \mathscr{Y}, \mathscr{Z}$. Random variables are denoted by upper case roman letters such as $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$, while lower case letters such as x, y, z denote the values taken by random variables. Linear and affine maps are also denoted by upper case roman letters such as L and A. The set of natural numbers (excluding zero) is denoted by $\mathbb{N}$. The set of real numbers is denoted by $\mathbb{R}$, while the set of non-negative real numbers is denoted by $\mathbb{R}_+$. The set of extended real-valued numbers is denoted by $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. The set of complex numbers is denoted by $\mathbb{C}$. Given a complex number $z \in \mathbb{C}$, we denote z^* to be its complex conjugate. All the logarithms appearing in study are with respect to base e unless specified otherwise. For any $M \in \mathbb{N}$, we denote $[M] = \{1, \ldots, M\}$. The **Kronecker delta function** is defined as $\delta_{ij} = 1$ if $i = j$ and 0 if $i \neq j$. Given some statement P on a set Ω, we denote $\{P\} = \{\omega \in \Omega \mid P(\omega) \text{ is true}\}$ to be the set of elements of Ω where the statement P holds. For example, if f is a real-valued function on Ω and $a \in \mathbb{R}$ is some number, we write $\{f \leq a\} = \{\omega \in \Omega \mid f(\omega) \leq a\}$. We denote an **indexed family** as $\{O_i\}_{i \in \mathscr{I}}$ or $(O_i)_{i \in \mathscr{I}}$, where the elements/objects O_i (which could be vectors, matrices, etc.) are indexed by elements of a set $\mathscr{I}$. Formally, an indexed family is a function from $\mathscr{I}$ to the set $\{\mathscr{O}_i \mid i \in \mathscr{I}\}$. Indexed families help to keep track of the order of elements (when $\mathscr{I}$ is ordered) and allow for repetitions of elements, in contrast with sets which are unordered and contain no repeated elements. Next, we define the asymptotic order notation. Let f, g be non-negative functions on

$\mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, where for all $i \in [N]$, $\mathcal{X}_i$ is either $\mathbb{N}$ or $(0, \infty)$. We say $f = O(g)$ if there are positive numbers B and C such that $\min_i x_i \geq B$ implies $f(x_1, \ldots, x_N) \leq Cg(x_1, \ldots, x_N)$. We say $f = \Omega(g)$ if there are positive numbers B and C such that $\min_i x_i \geq B$ implies $f(x_1, \ldots, x_N) \geq Cg(x_1, \ldots, x_N)$. We say $f = \Theta(g)$ if $f = O(g)$ and $f = \Omega(g)$. Our definitions concerns the asymptotic behavior of f and g as all the parameters are approaching infinity, and this may differ from the definitions in the literature for the order notation for multi-parameter functions. Finally, we note that we use the abbreviation "iff" to mean "if and only if".

2.1 Linear algebra

We call a vector space over $\mathbb{R}$ a real vector space and a vector space over $\mathbb{C}$ a complex vector space. In the following discussion, $\mathbb{K}$ is either $\mathbb{R}$ or $\mathbb{C}$. An **inner product** on a vector space $\mathscr{V}$ over $\mathbb{K}$ is a function $\langle \cdot, \cdot \rangle : \mathscr{V} \times \mathscr{V} \to \mathbb{K}$ such that $\langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$, $\langle u, \alpha v \rangle = \alpha \langle u, v \rangle$, and $\langle u, v \rangle^* = \langle v, w \rangle$ for all $u, v, w \in \mathscr{V}$ and $\alpha \in \mathbb{K}$. When convenient (e.g., when working with pure quantum states), we will use the Dirac notation: vectors are denoted by $|\psi\rangle$, and inner product between $|\phi\rangle$ and $|\psi\rangle$ is denoted by $\langle \phi | \psi \rangle$. A vector space equipped with an inner product is called an **inner product space**.

Given a subset $\mathscr{U} \subseteq \mathscr{V}$, the **span** of $\mathscr{U}$ is $\operatorname{span} \mathscr{U} = \{\sum_{i=1}^n \alpha_i u_i \mid (\forall n \in \mathbb{N})(\forall i \in [n]) \alpha_i \in \mathbb{K}, u_i \in \mathscr{U}\}$. A (Hamel) **basis** $\mathscr{B}$ of $\mathscr{V}$ is a minimal spanning set (i.e., $\operatorname{span} \mathscr{B} = \mathscr{V}$ and for any $\mathscr{B}' \subsetneq \mathscr{B}$, $\operatorname{span} \mathscr{B}' \subsetneq \mathscr{V}$). The number of basis vectors is called the **dimension** of $\mathscr{V}$, denoted by $\dim \mathscr{V} = |\mathscr{B}|$. It can be shown that the dimension is independent of the choice of basis. $\mathscr{V}$ is said to be finite-dimensional if $\dim \mathscr{V} < \infty$. All vector spaces in this study are assumed to be finite dimensional, unless stated otherwise. Any vector $v \in \mathscr{B}$ can be written as a unique linear combination of basis vectors. A collection of vectors $v_1, \ldots, v_n$ are said to be linearly independent if $\sum_{i=1}^n \alpha_i v_i = 0$ implies $\alpha_i = 0$ for all $i \in [n]$. It can be verified that a basis $\mathscr{B}$ is a maximal linearly independent set (i.e., every set strictly containing $\mathscr{B}$ is linearly dependent). Given an n-dimensional inner product space $\mathscr{V}$, an **orthonormal** basis of $\mathscr{V}$ is a collection of n vectors $\mathscr{B} = \{e_1, \ldots, e_n\}$ that satisfies $\langle e_i, e_j \rangle = \delta_{ij}$ for all $i, j \in [n]$. It can be verified that any set of n vectors satisfying

this property form a basis. Since $\mathscr{B}$ is a basis, there are unique numbers $v_1, \ldots, v_n \in \mathbb{K}$ such that $v = \sum_{i=1}^{n} v_i e_i$. The number v_i is said to be the ith component of v with respect to the basis $\{e_i\}_{i=1}^{n}$, and we write $v = (v_1, \ldots, v_n)$ when the basis is understood. We use the same terminology even if $\mathscr{B}$ is not orthonormal.

Given two vector spaces $\mathscr{V}$ and $\mathscr{W}$ over $\mathbb{K}$, their **direct sum** $\mathscr{V} \oplus \mathscr{W}$ is a vector space of tuples (v, w) with $v \in \mathscr{V}$ and $w \in \mathscr{W}$, with addition and scalar multiplication defined component-wise. Inner products on $\mathscr{V}$ and $\mathscr{W}$ induce an inner product on $\mathscr{V} \oplus \mathscr{W}$ as $\langle (v, w), (v', w') \rangle = \langle v, v' \rangle + \langle w, w' \rangle$ for $v, v' \in \mathscr{V}$ and $w, w' \in \mathscr{W}$. The **tensor product** of $\mathscr{V}$ and $\mathscr{W}$ is denoted by $\mathscr{V} \otimes \mathscr{W}$. The technical definition of a tensor product is not needed for this study, and we refer the interested reader to [11, Sec. (I.4)] for details. In practice, it suffices to look at Kronecker products as they give a concrete way to compute the tensor product of two finite-dimensional vectors. If $\{e_1, \ldots, e_n\}$ is a basis of $\mathscr{V}$ and $\{f_1, \ldots, f_m\}$ is a basis of $\mathscr{W}$, then $\{e_i \otimes f_j \mid i \in [n], j \in [m]\}$ is a basis for $\mathscr{V} \otimes \mathscr{W}$. We define the components of vectors in $\mathscr{V} \otimes \mathscr{W}$ with respect to this basis as follows. If $v \in \mathscr{V}$ and $w \in \mathscr{W}$, then $(v \otimes w)_{ij} = v_i w_j$ are the components of $v \otimes w$ for $i \in [n]$ and $j \in [m]$. Inner products on $\mathscr{V}$ and $\mathscr{W}$ induce an inner product on $\mathscr{V} \otimes \mathscr{W}$ as $\langle v \otimes w, v' \otimes w' \rangle = \langle v, v' \rangle \langle w, w' \rangle$ for $v, v' \in \mathscr{V}$ and $w, w' \in \mathscr{W}$. A **linear subspace** $\mathscr{U}$ of a vector space $\mathscr{V}$ is a subset of $\mathscr{V}$ that satisfies $u + \alpha v \in \mathscr{U}$ for all $u, v \in \mathscr{U}$ and all $\alpha \in \mathbb{K}$. It follows that $\mathscr{U}$ is itself a vector space under the addition and scalar multiplication inherited from $\mathscr{V}$. Given a subspace $\mathscr{U} \subseteq \mathscr{V}$, the **quotient space** $\mathscr{V}/\mathscr{U}$ consists of elements (called cosets) $[v] = \{v + u \mid u \in \mathscr{U}\}$ for $v \in \mathscr{V}$, with addition defined as $[v_1] + [v_2] = [v_1 + v_2]$ and scalar multiplication defined as $\alpha[v] = [\alpha v]$ for $v_1, v_2, v \in \mathscr{V}$ and $\alpha \in \mathbb{K}$. It can be verified that the quotient space is itself a linear vector space under the addition and scalar multiplication defined above. Given any subset $\mathscr{U} \subseteq \mathscr{V}$, the **orthogonal complement** of $\mathscr{U}$ is defined as $\mathscr{U}^{\perp} = \{v \in \mathscr{V} \mid \langle u, v \rangle = 0 \ \forall u \in \mathscr{U}\}$. If $\mathscr{U}$ is a subspace, then $\mathscr{U}^{\perp}$ is also a subspace, and we have $\mathscr{V} = \mathscr{U} + \mathscr{U}^{\perp}$, where $\mathscr{A} + \mathscr{B} = \{a + b \mid a \in \mathscr{A}, b \in \mathscr{B}\}$ is the **Minkowski sum** of the sets $\mathscr{A}, \mathscr{B} \subseteq \mathscr{V}$.

A linear map L betwen two vector spaces $\mathscr{V}$ and $\mathscr{W}$ is a function $\mathrm{L} \colon \mathscr{V} \to \mathscr{W}$ that satisfies $\mathrm{L}(u + v) = \mathrm{L}(u) + \mathrm{L}(v)$ and $\mathrm{L}(\alpha v) = \alpha \mathrm{L}(v)$ for all $u, v \in \mathscr{V}$ and $\alpha \in \mathbb{K}$. We will denote $\mathrm{L}(v)$ as $\mathrm{L}v$

when no confusion arises. The adjoint of a linear map L is a function $\mathrm{L}^\dagger \colon \mathscr{W} \to \mathscr{V}$ that satisfies $\langle w, \mathrm{L}v\rangle = \langle \mathrm{L}^\dagger w, v\rangle$ for all $v \in \mathscr{V}$ and $w \in \mathscr{W}$. A linear map $\mathrm{L} \colon \mathscr{V} \to \mathscr{V}$ is said to be **Hermitian** or **self-adjoint** if $\mathrm{L}^\dagger = \mathrm{L}$. If $\mathscr{V}$ has dimension n and $\mathscr{W}$ has dimension m, then L can be written as an $m \times n$ matrix by choosing a basis for $\mathscr{V}$ and $\mathscr{W}$. When no confusion arises, we will use the same notation L for the linear map as well as the matrix. In terms of matrices, $\mathrm{L}^\dagger = (\mathrm{L}^*)^{\mathrm{T}}$ is the conjugate transpose of L. A linear map $\mathrm{L} \colon \mathscr{V} \to \mathscr{V}$ is **positive semidefinite (PSD)** if $\langle v, \mathrm{L}v\rangle \geq 0$ for all $v \in \mathscr{V}$. For $\mathbb{K} = \mathbb{C}$, it can be shown that a PSD map is necessarily Hermitian [24, Prop. (2.12)]. The **kernel** of a linear map L is $\ker(L) = \{v \in \mathscr{V} \mid Lv = 0\}$. The kernel is always a linear subspace of $\mathscr{V}$. The **range** of a linear map is $\mathrm{range}(L) = \{Lv \mid v \in \mathscr{V}\}$. The image is always a linear subspace of $\mathscr{W}$. The **identity map** is a linear map $\mathbb{I}_{\mathscr{V}} \colon \mathscr{V} \to \mathscr{V}$ defined as $\mathbb{I}_{\mathscr{V}}(v) = v$ for all $v \in \mathscr{V}$. When $\mathscr{V}$ is clear from the context, we denote $\mathbb{I}_{\mathscr{V}}$ as $\mathbb{I}$. A linear map L is said to be an **isomorphism** if it is bijective. All isomorphisms are invertible, that is, there is a linear map $\mathrm{L}^{-1} \colon \mathscr{W} \to \mathscr{V}$ such that $\mathrm{L}^{-1} \circ \mathrm{L} = \mathbb{I}_{\mathscr{V}}$ and $\mathrm{L} \circ \mathrm{L}^{-1} = \mathbb{I}_{\mathscr{W}}$. A linear map L between two inner product spaces is said to be an **isometry** if it preserves inner products ($\langle \mathrm{L}u, \mathrm{L}v\rangle = \langle u, v\rangle$ for all $u, v \in \mathscr{W}$). A linear map $\mathrm{U} \colon \mathscr{V} \to \mathscr{V}$ is said to be **unitary** if it an isometric isomorphism. This can be shown to be equivalent to the condition $\mathrm{U}^\dagger\mathrm{U} = \mathrm{U}\mathrm{U}^\dagger = \mathbb{I}$. Consequently, we have $U^{-1} = U^\dagger$. Given two linear maps $\mathrm{L}^{(1)} \colon \mathscr{V}^{(1)} \to \mathscr{W}^{(1)}$ and $\mathrm{L}^{(2)} \colon \mathscr{V}^{(2)} \to \mathscr{W}^{(2)}$, their direct sum $\mathrm{L} = \mathrm{L}^{(1)} \oplus \mathrm{L}^{(2)}$ is the linear map $\mathrm{L} \colon (\mathscr{V}^{(1)} \oplus \mathscr{V}^{(2)}) \to (\mathscr{W}^{(1)} \oplus \mathscr{W}^{(2)})$ defined as $\mathrm{L}((v^{(1)}, v^{(2)})) = (\mathrm{L}^{(1)}(v^{(1)}), \mathrm{L}^{(2)}(v^{(2)}))$. In matrix form, we can write L as the block matrix

$$\mathrm{L} = \begin{pmatrix} \mathrm{L}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathrm{L}^{(2)} \end{pmatrix}, \tag{2.1}$$

where $\mathbf{0}$ is a matrix of zeros of appropriate size.

A **linear functional** on a vector space $\mathscr{V}$ over $\mathbb{K}$ is a linear map $\mathrm{L} \colon \mathscr{V} \to \mathbb{K}$. Given a vector space $\mathscr{V}$, the dual space $\mathscr{V}^*$ is the set of all linear functions on $\mathscr{V}$. The Riesz representation theorem

ensures that for every $v \in \mathscr{V}$, there is a unique linear functional $L_v \in \mathscr{V}^*$ such that $L_v(w) = \langle v, u \rangle$ for all $u \in \mathscr{V}$ [24, Thm. (3.4)]. In Dirac notation, we write $|v\rangle$ for elements of $\mathscr{V}$ and $\langle v|$ elements of $\mathscr{V}^*$. If we write $|v\rangle$ as a column vector, $\langle v|$ is the row vector obtained by taking the conjugate transpose of $|v\rangle$. Given $v \in \mathscr{V}$ and $w \in \mathscr{W}$, we will denote $|w\rangle\langle v| : \mathscr{V} \to \mathscr{W}$ to be the linear map $|w\rangle\langle v|(u) = \langle v, u\rangle w$ for $u \in \mathscr{V}$. This works particularly well with the Dirac notation, where $|w\rangle\langle v||u\rangle = \langle v|u\rangle|w\rangle$.

A **seminorm** on a vector space $\mathscr{V}$ is a function $\|\cdot\| : \mathscr{V} \to \mathbb{R}$ that is (1) absolutely homogeneous ($\|\alpha v\| = |\alpha|\,\|v\|$ for all $\alpha \in \mathbb{K}$, $v \in \mathscr{V}$), and (2) satisfies the triangle inequality ($\|u + v\| \leq \|u\| + \|v\|$). The above properties imply that $\|0\| = 0$ and $\|v\| \geq 0$ for all $v \in \mathscr{V}$. A **norm** is a seminorm that satisfies $\|v\| = 0$ if and only if $v = 0$. A norm induces a metric on $\mathscr{V}$ according to $\mathscr{d}(u, v) = \|u - v\|$. Every inner product induces a norm according to $\|v\| = \sqrt{\langle v, v\rangle}$. Note, however, that there are norms that are not induced by an inner product.

A **Hilbert space** $\mathscr{H}$ over $\mathbb{K}$ is a complete inner product space. Recall that $\mathscr{H}$ is said to be complete if every Cauchy sequence with respect to the norm induced by the inner product converges to a point in $\mathscr{H}$. All finite-dimensional inner product spaces are complete, and therefore, Hilbert spaces. Thus, we will use the terminology Hilbert space for an inner product space in this study.

The set of n-dimensional vectors with entries from $\mathbb{K}$ is denoted by $\mathbb{K}^n$, while the set of $m \times n$ matrices with entries from $\mathbb{K}$ is denoted by $\mathbb{K}^{m\times n}$. The set of $n \times n$ complex, self-adjoint (Hermitian) matrices is denoted by $\mathbb{S}_n$. It can be verified that $\mathbb{S}_n$ is a *real* vector space of dimension n^2. Every matrix $A \in \mathbb{S}_n$ has a **spectral decomposition**, i.e., we can write $A = \sum_{i=1}^n \lambda_i |\lambda_i\rangle\langle\lambda_i|$, where $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ are called the eigenvalues of A and $|\lambda_1\rangle, \ldots, |\lambda_n\rangle$ are the corresponding eigenvectors (not necessarily unique). The normalized eigenvectors of an $n \times n$ Hermitian matrix form an orthonormal basis for $\mathbb{C}^n$. For any matrix $A \in \mathbb{S}_n$, we denote $\lambda(A) = (\lambda_1(A), \ldots, \lambda_n(A))$ to be the vector of eigenvalues of A. Furthermore, we denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to be the maximum and minimum eigenvalues of A, respectively. Given a Hermitian matrix $A \in \mathbb{S}_n$, we define its **support** as the span of eigenvectors corresponding to non-zero eigenvalues. Equivalently, the support of A is the orthogonal complement of the kernel of A. The singular values of $A \in \mathbb{K}^{m\times n}$ are the eigenvalues

of $\sqrt{A^\dagger A}$, where we note that $A^\dagger A$ is Hermitian for any (possibly rectangular) matrix A. It can be verified that if A is a Hermitian matrix, then the singular values of A are just the absolute values of the eigenvalues of A. For any matrix $A \in \mathbb{K}^{m\times n}$, we denote $\sigma(A) = (\sigma_1(A), \ldots, \sigma_n(A))$ to be the vector of singular values of A. We denote $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ to be the maximum and minimum singular values of A, respectively.

Given any Hermitian matrix A and a function $f\colon \mathbb{R} \to \mathbb{R}$, we define $f(A) = \sum_{i=1}^n f(\lambda_i) \left|\lambda_i\right\rangle \left\langle\lambda_i\right|$. The domain of the function f can be restricted to a subset of the real line depending on the scenario. For example, we can define the square-root of a PSD matrix $A \in \mathbb{S}_n$ as $\sqrt{A} = \sum_{i=1}^n \sqrt{\lambda_i} \left|\lambda_i\right\rangle \left\langle\lambda_i\right|$. The trace of a matrix $A \in \mathbb{K}^{n\times n}$ is defined as $\mathrm{Tr}(A) = \sum_{i=1}^n \langle e_i, Ae_i\rangle$, where $\{e_1, \ldots, e_n\}$ is an orthonormal basis of $\mathbb{K}^d$. It can be verified that trace is the same irrespective of the choice of orthonormal basis, and we have $\mathrm{Tr}(A) = \sum_{i=1}^n \lambda_i$. The **rank** of $A \in \mathbb{K}^{m\times n}$ is the dimension of range(A). It can be verified that rank 1 matrices can be written as $\left|w\right\rangle \left\langle v\right|$ for some $v \in \mathbb{K}^n$ and $w \in \mathbb{K}^m$. We say $A \in \mathbb{K}^{n\times n}$ is **full rank** if its rank is equal to n. It can be verified that full rank matrices are invertible.

The Euclidean or standard inner product on $\mathbb{K}^n$ is defined as $\langle u, v\rangle = u^\dagger v = \sum_{i=1}^n u_i^* v_i$, where $u = (u_1, \ldots, u_n), v = (v_1, \ldots, v_n) \in \mathbb{K}^n$ with respect to some fixed orthonormal basis. The Euclidean norm on $\mathbb{K}^n$ is defined as $\left\|v\right\|_2 = \sqrt{\langle v, v\rangle} = \sqrt{v^\dagger v}$. More generally, for $p \in [1, \infty)$, the **p-norm** on $\mathbb{K}^n$ is defined as $\left\|v\right\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$, whereas $\left\|v\right\|_\infty = \max_{i\in[n]} |v_i|$. **Hölder's inequality** states that $|\left\langle u, v\right\rangle| \leq \left\|u\right\|_p \left\|v\right\|_q$ for any $p, q \in [1, \infty]$ satisfying $1/p + 1/q = 1$. For the case of $p = q = 2$, we obtain the **Cauchy-Schwarz inequality** $|\left\langle u, v\right\rangle| \leq \left\|u\right\|_2 \left\|v\right\|_2$.

The Hilbert-Schmidt (HS) inner product on $\mathbb{K}^{n\times n}$ is defined as $\langle A, B\rangle = \mathrm{Tr}(A^\dagger B) = \sum_{i,j=1}^n A_{ij}^* B_{ij}$, where $A = (A_{ij})$ and $B = (B_{ij})$ with respect to some fixed orthonormal basis of $\mathbb{K}^n$. The Hilbert-Schmidt or Frobenius norm on $\mathbb{K}^{n\times n}$ is defined as $\left\|A\right\|_{\mathrm{HS}} = \sqrt{\mathrm{Tr}(A^\dagger A)}$. The **Schatten p-norm** on $\mathbb{K}^{n\times n}$ is defined to be $\left\|A\right\|_p = \left\|\sigma(A)\right\|_p$, where $p \in [1, \infty]$ and $\sigma(A) = (\sigma_1, \ldots, \sigma_n)$ denotes the vector of singular values of A. If A is Hermitian, it holds that $\left\|A\right\|_p = \left\|\lambda(A)\right\|_p$, where $\lambda(A) = (\lambda_1, \ldots, \lambda_n)$ denote the vector of eigenvalues of A. The Schatten-2 norm is just the Hilbert-Schmidt norm. Of particular interest is the Schatten-1 norm, also called the **trace norm**,

which can be written as $\|A\|_1 = \mathrm{Tr}\sqrt{A^\dagger A}$. Also, the Schatten-$\infty$ norm coincides with the **operator norm** or the **spectral norm** of A, which is given as $\|A\|_\infty = \max_{\|v\|_2 \leq 1} \|Av\|_2 = \sigma_{\max}(A)$. All Schatten-$p$ norms are unitarily invariant, i.e., $\left\|UAU^\dagger\right\|_p = \|A\|_p$ for all unitaries U. For matrices, we can derive a Hölder's inequality for Schatten norms. If $A, B \in \mathbb{K}^{n\times n}$, we have $|\mathrm{Tr}(A^\dagger B)| \leq \langle \sigma(A), \sigma(B)\rangle \leq \|\sigma(A)\|_p \|\sigma(B)\|_q = \|A\|_p \|B\|_q$ for any $p, q \in [1, \infty]$ satisfying $1/p + 1/q = 1$. The first inequality is a consequence of von Neumann's trace inequality [72] and the fact that $\sigma(A^\dagger) = \sigma(A)$, while the second inequality is the usual Hölder's inequality.

Finally, we make a remark on notation involving inequalities involving vectors and matrices. For a vector $x \in \mathbb{R}^n$, inequalities such as $x \geq 0$ are interpreted component-wise. For matrices $A, B \in \mathbb{S}_n$, the inequality $A \geq B$ means $A - B$ is positive semidefinite. It can be verified that these definitions define a partial order on $\mathbb{R}^n$ and $\mathbb{S}_n$, respectively.

2.2 Quantum states and measurements

In this section, we review some basic definitions and results in quantum information that are used in our study. We refer the reader to [77] for a comprehensive introduction to quantum information theory.

A d-dimensional **quantum state** $\sigma \in \mathbb{C}^{d\times d}$ is a Hermitian, positive semi-definite matrix with trace 1. A quantum state is said to be **pure** if it is rank 1. Equivalently, $\rho \in \mathbb{C}^{d\times d}$ is pure if there is a vector $|\psi\rangle \in \mathbb{C}^d$ such that $\rho = |\psi\rangle\langle\psi|$. An **observable** is mathematically a Hermitian matrix. Depending on the situation, an observable can describe some physical property of the system such as energy. Given an observable $\mathcal{O} \in \mathbb{C}^{d\times d}$, its **expectation value** with respect to the state ρ is defined as $\langle \mathcal{O}\rangle = \mathrm{Tr}(\mathcal{O}\rho)$.

In practice, we do not know the underlying state ρ. Instead, we typically have access to outcomes obtained by measuring the state ρ. Our goal in this study is to learn or verify properties of ρ using these measurement outcomes. By a property of a state, we mean any function of the state, as for example the expectation value of an observable.

To obtain a measurement outcome, we need to measure the quantum state according to

a chosen measurement procedure defined by a measurement setting. A measurement setting is described mathematically by a **positive operator-valued measure (POVM)**, which is an indexed family of positive semi-definite operators $\boldsymbol{E} = \{E_i\}_{i=1}^{M}$ that sum to identity. We call an element $m \in [M]$ a **label** corresponding to the POVM element E_m that can be observed upon a measurement. Here, M denotes the total number of labels for the POVM $\boldsymbol{E}$. If the underlying state is ρ, the probability of observing the label $m \in [M]$ upon measuring $\boldsymbol{E}$ is given by Born's rule as $p_{\boldsymbol{E},\rho}(m) = \mathrm{Tr}(E_m\rho)$. When the POVM is understood, we denote $p_{\boldsymbol{E},\rho}$ by p_ρ. If there are multiple POVMs $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(L)}$, then the kth element of the ith POVM is denoted by $E_k^{(i)}$, and we write $p_{\boldsymbol{E}^{(i)},\rho} = p_\rho^{(i)}$ when the POVMs are understood.

We denote 1-qubit Pauli observables as X, Y, Z. The eigenstates of X with eigenvalues $+1$ and -1 are denoted by $|+\rangle, |-\rangle$ respectively, while the eigenstates of Z with eigenvalues $+1$ and -1 are denoted by $|0\rangle$ and $|1\rangle$. The eigenbasis of $Z^{\otimes n}$ in an n-qubit system is called the computational basis. For an n-qubit Pauli P, the POVM that measures the eigenvalue of P is $\{(\mathbb{I}+P)/2, (\mathbb{I}-P)/2\}$.

A peculiarity of quantum mechanics is that after a measurement, the quantum state is disturbed. For this reason, to obtain several measurement outcomes, one needs to prepare many copies of the state of interest ρ. Ideally, one seeks to prepare many independent and identical copies of the state ρ, which are then measured one at a time. We refer to assumption that independent and identical copies of the states are prepared as the *iid* assumption, in line with the independent and identically distributed assumption used in classical statistics. While it is hard to satisfy the iid assumption exactly, in many experiments, the iid assumption is reasonable, at least over short time scales. Moreover, the iid assumption greatly simplifies the statistical analysis of data, especially for the purposes of learning or verifying the properties of quantum system. For this reason, we will work with the iid assumption in this study.

In the situation where we perform multiple measurements, there are broadly three measurement strategies one can implement: non-adaptive, adaptive, and entangled measurements. The definitions given below follow [108]. A measurement is said to be non-adaptive if we fix the POVMs a priori, and each POVM is implemented on a single copy of the state. A measurement is said to be adaptive

if we measure one copy of the state at a time, but the POVM at any given time can depend on the past measurement outcomes. Finally, entangled measurements corresponds to jointly measuring many copies of the state at one time. We remark that one can, in principle, mix and match these different types of measurements. For example, one can choose to jointly measure two copies of a state at a given time, and use the previous outcomes to inform the POVM to be implemented in the next time step.

In this study, we focus on non-adaptive measurements. Thus, when we say a "measurement protocol", we mean a non-adaptive measurement protocol, unless stated otherwise. A non-adaptive measurement protocol is simply a list of POVMs along with the number of times each POVM is repeated. We formally define this below. Note that we assume all the measurements are performed independently.

Definition 2.1 (Measurement protocol). A (non-adaptive) measurement protocol $\mathfrak{M}$ is a list of pairs, where each pair consists of a POVM along with the number of times that POVM is repeated, i.e.,

$$\mathfrak{M} = \left\{ \left(\boldsymbol{E}^{(i)}, N_i \right) \right\}_{i=1}^{L} \tag{2.2}$$

Here, the ith POVM $\boldsymbol{E}^{(i)} = \{E_1^{(i)}, \ldots, E_{M_i}^{(i)}\}$ has M_i labels, and it is repeated N_i times. L denotes the total number of POVMs implemented by the measurement protocol. We assume that all the POVMs are distinct.

The total number of samples $\sum_{i=1}^{L} N_i$ used by the measurement protocol $\mathfrak{M}$ is denoted by $N(\mathfrak{M})$ or simply N when $\mathfrak{M}$ is clear from context.

Given that the underlying state is ρ, we denote the joint probability distribution over the labels determined by the measurement protocol $\mathfrak{M}$ as per Born's rule by the shorthand $\mathbb{P}_{\mathfrak{M},\rho}$. □

Another equivalent way of thinking about a measurement protocol is as a finite sequence of POVMs, where elements in the sequence may be repeated. The definition of a measurement protocol given in Eq. (2.2) avoids this repetition by listing the distinct POVMs that were/will be measured, and saying how many times each POVM was repeated.

To illustrate the above definition, consider the following example. Suppose that we measure Pauli X two times and Pauli Z one time on a one-qubit state ρ. The POVM defining X measurement is $\boldsymbol{E}^{(1)} = \{|+\rangle\langle+|, |-\rangle\langle-|\}$ (projection onto eigenvectors of X), and similarly, the POVM defining Z measurement is $\boldsymbol{E}^{(2)} = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}$. The measurement protocol is then described by the set $\mathfrak{M} = \{(\boldsymbol{E}^{(1)}, 2), (\boldsymbol{E}^{(2)}, 1)\}$. The labels of an experiment implementing this measurement protocol are tuples (or strings) of the form (i, j, k) where $i, j \in \{+, -\}$ and $k \in \{0, 1\}$ (since we measure X twice and Z once). Since we assume that the states are prepared independently and identically, the joint probability distribution $\mathbb{P}_{\mathfrak{M},\rho}$ is a product distribution, i.e., $\mathbb{P}_{\mathfrak{M},\rho}((i, j, k)) = \mathrm{Tr}(\rho|i\rangle\langle i|)\mathrm{Tr}(\rho|j\rangle\langle j|)\mathrm{Tr}(\rho|k\rangle\langle k|)$ for any given label (i, j, k).

The above examples describe performing a fixed set of measurements. Sometimes, however, it is advantageous to implement a randomized measurement protocol. For example, we can randomly sample $\boldsymbol{E}^{(1)}$ or $\boldsymbol{E}^{(2)}$ and implement the sampled POVM. We formally define a randomized measurement protocol below.

Definition 2.2 (Randomized measurement protocol). A randomized measurement protocol consists of L POVMs $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(L)}$, a probability distribution p over $[L]$, and a positive integer N, wherein one samples the ith POVM with probability p_i and measures it, and this procedure is repeated N times. □

Randomized measurement protocols can be described using a single *effective POVM*. The most general situation is where one stores both the index i of the POVM that was sampled and the outcome $j \in [M_i]$ that was observed upon measuring this POVM. In this case, the effective POVM is given as $\{p_i E_j^{(i)} \mid j \in [M_i], i \in [L]\}$. This situation occurs, for example, in direct fidelity estimation [33, 26] and classical shadows [54]. Another situation is when $M_i = M_\circ$ for all $i \in [L]$ and one only records outcome observed upon measurement and not the POVM that was sampled. This situation occurs, for example, in the randomized Pauli measurement protocol discussed in [92]. In this case, the effective POVM is given by $\{\sum_{i=1}^{L} p_i E_j^{(i)} \mid j \in [M_\circ]\}$. Thus, when referring to the measurement protocol $\mathfrak{M}$ for a randomized measurement, it is sufficient to specify the effective

POVM along with the number of times this POVM is measured.

Another type of measurement protocol that is important in practice is an informationally complete measurement protocol. Such measurement protocols contain enough information to completely reconstruct the state, and thus, are important for quantum state tomography. We formally define these measurement protocols below.

Definition 2.3 (Informationally complete measurement protocol). A measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ is said to be informationally complete (IC) if for all states ρ, σ with $\rho \neq \sigma$, there is some POVM $\boldsymbol{E}^{(i)}$ for $i \in [L]$ and an index $k \in [M_i]$ such that $\mathrm{Tr}(E_k^{(i)}\rho) \neq \mathrm{Tr}(E_k^{(i)}\sigma)$. □

While the above definition might look mathematically unwieldy, it turns out that a measurement protocol is IC if and only if it spans the set of all Hermitian matrices of appropriate dimension, giving a mathematically simple characterization of IC measurements. Although this result is known in the literature [88], we include a short proof below. We use the notation $\operatorname{span}\mathfrak{M} = \operatorname{span}\{E_k^{(i)} \mid k \in [M_i], i \in [L]\}$ for any measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ in the proof.

Proposition 2.4. *A measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ is informationally complete if and only if $\{E_k^{(i)} \mid k \in [M_i], i \in [L]\}$ spans the set of all $d \times d$ Hermitian matrices.*

Proof. Let $(\operatorname{span}\mathfrak{M})^{\perp}$ denote the orthogonal complement of $\operatorname{span}\mathfrak{M}$.

Suppose that $\mathfrak{M}$ does not span $\mathbb{S}_d$. Then, we can write $\mathbb{S}_d = \operatorname{span}\mathfrak{M} \oplus (\operatorname{span}\mathfrak{M})^{\perp}$, where $(\operatorname{span}\mathfrak{M})^{\perp}$ contains at least one non-zero element A. Since $\mathbb{I} \in \operatorname{span}\mathfrak{M}$, we must have $\langle \mathbb{I}, A\rangle = \mathrm{Tr}(A) = 0$. Thus, after rescaling if necessary, we can assume that $\mathrm{Tr}(A) = 0$ and $\|A\|_\infty \leq 1$, where $\|A\|_\infty$ denotes the Schatten-∞ norm of A. Then, $\rho = (\mathbb{I} + A)/d$ and $\sigma = (\mathbb{I} - A)/d$ are density matrices that are not equal. However, $\mathrm{Tr}((E_k^{(i)}(\rho - \sigma)) = 2\mathrm{Tr}(E_k^{(i)}A)/d = 0$ for all $k \in [M_i]$ and all $i \in [L]$, so that $\mathfrak{M}$ is not informationally complete.

Next, suppose that $\mathfrak{M}$ spans $\mathbb{S}_d$. Then, there is a subset $\mathscr{B} \subseteq \{E_k^{(i)} \mid k \in [M_i], i \in [L]\}$ that is a basis of $\mathbb{S}_d$. Denote the elements of $\mathscr{B}$ as $W_1, \ldots, W_{d^2}$. Consequently, given any two density matrices ρ, σ, we can write $\rho - \sigma = \sum_j \alpha_j W_j$ for some unique numbers $\alpha_1, \ldots, \alpha_{d^2} \in \mathbb{R}$. If

$\langle \rho - \sigma, W_i \rangle = \mathrm{Tr}((\rho - \sigma)W_i) = 0$ for all i, then $0 = \sum_j \alpha_j \langle \rho - \sigma, W_j \rangle = \langle \rho - \sigma, \rho - \sigma \rangle = \|\rho - \sigma\|_{\mathrm{HS}}^2$. This implies $\rho = \sigma$, showing that $\mathfrak{M}$ is informationally complete. □

A special type of informationally completely measurements, called **symmetric informationally complete (SIC)** measurements, are of importance in quantum information [88]. A POVM $\boldsymbol{E}$ is said to be a SIC-POVM if it is informationally complete, has exactly d^2 elements all rank one, and $\mathrm{Tr}(E_j E_k) = (d\delta_{jk} + 1)/(d^2(d+1))$ for all $j, k \in [d^2]$, where δ_{jk} denotes the Kronecker delta function. Since SIC-POVMs are informationally complete, they can be used for quantum tomography [83]. Such measurements protocols are sometimes referred to as "single-setting", since we only measure a single POVM.

2.3 Probability theory

In this section, we review some basic concepts in measure-theoretic probability. These concepts are useful for studying the general statistical problem in Ch. 4. That said, when we apply this statistical framework to the problem of learning observables, we can do away with most of the underlying measure-theoretic details.

We start with some basic concepts from measure theory (see [48] for an introduction to the subject). Given a non-empty set Ω, a **$\boldsymbol{\sigma}$-algebra** $\mathscr{F}$ on Ω is defined to be a collection of subsets of Ω that is closed under countable unions and relative complements. That is, given $A_1, A_2, \ldots \in \mathscr{F}$, we have $\cup_n A_n \in \mathscr{F}$, and given $A, B \in \mathscr{F}$, $A \setminus B \in \mathscr{F}$. We work with the convention that union of an empty collection of sets is equal to the empty set $\varnothing$. This implies that we always have $\varnothing, \Omega \in \mathscr{F}$. We call the pair $(\Omega, \mathscr{F})$ a **measurable space**, and the elements of $\mathscr{F}$ are called $\mathscr{F}$-measurable sets or simply measurable sets if the σ-algebra is clear from context. Given any collection $\mathscr{A}$ of subsets of Ω, the σ-algebra generated by $\mathscr{A}$ is the smallest σ-algebra on Ω containing $\mathscr{A}$, and is denoted by $\sigma(\mathscr{A})$. In probability theory, a special class of σ-algebras, called the Borel σ-algebras, plays an important role. To define a Borel σ-algebra, we need the idea of a topology on Ω. Moreover, we will work with a specific type of topological spaces called Polish spaces in this study. For this reason, we

review some basic concepts from topology.

A **topology** on a non-empty set Ω is a collection τ of subsets of Ω that is closed under arbitrary unions and finite intersections. The pair (Ω, τ) is called a **topological space** (see [74] for an introduction to topology and metric spaces). The elements of τ are called open sets, and a complement of an open set is called a closed set. Given any subset $A \subseteq \Omega$, the **interior** of A, denoted $\operatorname{int} A$, is the largest open set contained in A, while the **closure** of A, denoted $\operatorname{cl} A$, is the smallest closed set containing A. A subset $A \subseteq \Omega$ is **compact** if for every collection of open sets whose union contains A, there is a finite subcollection whose union contains A. An important result in analysis, called the Heine-Borel theorem, states that a set in a Euclidean space is compact iff it is closed and bounded. A subset $\Omega_0 \subseteq \Omega$ is said to be **dense** in Ω if $\operatorname{cl} \Omega_0 = \Omega$. A topological space is said to be **separable** if it has a countable dense subset. Given a family $\mathscr{A}$ of subsets of Ω, the topology generated by $\mathscr{A}$ is the smallest topology on Ω that contains $\mathscr{A}$.

While topological spaces can be very abstract, we will mainly deal with spaces that are generated by a metric. A **metric** on Ω is a function $\mathscr{d} \colon \Omega \times \Omega \to \mathbb{R}$ that (1) satisfies $\mathscr{d}(\omega_1, \omega_2) = 0$ iff $\omega_1 = \omega_2$ for all $\omega_1, \omega_2 \in \Omega$, (2) is symmetric ($\mathscr{d}(\omega_1, \omega_2) = \mathscr{d}(\omega_1, \omega_2)$ for all $\omega_1, \omega_2 \in \Omega$), and (3) satisfies the triangle inequality ($\mathscr{d}(\omega_1, \omega_2) \leq \mathscr{d}(\omega_1, \omega_2) + \mathscr{d}(\omega_2, \omega_3)$ for all $\omega_1, \omega_2, \omega_3 \in \Omega$). Using the above requirements, it can be shown that a metric must always be non-negative. A metric gives a way to measure distances between points of Ω. The pair $(\Omega, \mathscr{d})$ is called a **metric space**. $(\Omega, \mathscr{d})$ is said to be **complete** if every Cauchy sequence in Ω converges to a point in Ω. For any $r > 0$, the set $B(\omega, r) = \{\omega' \in \Omega \mid \mathscr{d}(\omega, \omega') < r\}$ is called an **open ball** of radius r around $\omega \in \Omega$. The topology on Ω generated by the open balls is called the metric topology, or the topology induced by the metric $\mathscr{d}$. A topological space (Ω, τ) is said to be **metrizable** if there is a metric $\mathscr{d}$ that induces the topology τ. A **Polish space** is a complete separable metrizable topological space. Polish spaces are important in the study of probability theory since they provide a way to unify commonly encountered spaces such as discrete spaces (for "discrete probability distributions") and Euclidean spaces (for "continuous probability distributions").

A **Borel** σ-algebra on a topological space (Ω, τ) is the σ-algebra generated by the topology τ.

We denote the Borel σ-algebra as $\mathscr{B}(\Omega) = \sigma(\tau)$ when τ is clear from context. The pair $(\Omega, \mathscr{B}(\Omega))$ is called a **Borel space**. When working with a topological space (Ω, τ) that is a Polish space, we will refer to the corresponding Borel space $(\Omega, \mathscr{B}(\Omega))$ also as a Polish space.

A function $f\colon \Omega_1 \to \Omega_2$ between two measurable spaces $(\Omega_1, \mathscr{F}_1)$ and $(\Omega_2, \mathscr{F}_2)$ is said to be **measurable** if the preimage of measurable sets are measurable, i.e., $f^{-1}(A) \equiv \{\omega_1 \in \Omega_1 \mid f(\omega_1) \in A\} \in \mathscr{F}_1$ for all $A \in \mathscr{F}_2$. Similarly, a function f between two topological spaces is said to be **continuous** if the preimage of opens sets are open. A function is said to be **Borel measurable** or Borel if it is a measurable function between two Borel spaces. It follows from the definitions that all continuous functions are Borel measurable.

A **measure** m on a measurable space $(\Omega, \mathscr{F})$ is a non-negative function on $\mathscr{F}$ that is countably additive, i.e., for any $A_1, A_2, \ldots \in \mathscr{F}$ that are mutually disjoint, we have $m(\cup_n A_n) = \sum_n m(A_n)$. We call the triple $(\Omega, \mathscr{F}, m)$ a measure space. A measure is said to be finite if $m(\Omega) < \infty$, and it is said to be σ-finite if there is some sequence of measurable sets $A_1, A_2, \ldots \in \mathscr{F}$ that satisfies $\Omega \subseteq \cup_n A_n$ and $m(A_n) < \infty$. A Borel measure is a measure defined on a Borel space.

Given a real-valued measurable function f, we denote the Lebesgue integral of f over $A \in \mathscr{F}$ with respect to m as $\int_A f dm$. A measurable function f is said to be **integrable** if $\int_\Omega |f| dm < \infty$. A measure m_1 is said to be **absolutely continuous** with respect to another measure m_2, denoted $m_1 \ll m_2$, if $m_2(A) = 0$ implies $m_1(A) = 0$ for all $A \in \mathscr{F}$. An important theorem from measure theory, called the Radon-Nikodym theorem, says that if m_1 and m_2 are σ-finite and $m_1 \ll m_2$, then there is a non-negative integrable function f such that $m_1(A) = \int_A f dm_2$. Moreover, f is unique up to a set of m_2-measure 0 in the sense that if g is another function that satisfies the above properties, then $m_2(\{f \neq g\}) = 0$. The function f is called the Radon-Nikodym derivative of m_1 with respect to m_2, and denoted by dm_1/dm_2.

We are now ready to use these basic definitions to define probability spaces and random variables. A **probability measure** $\mathbb{P}$ on a measurable space $(\Omega, \mathscr{F})$ is a finite measure that satisfies $\mathbb{P}(\Omega) = 1$. We sometimes refer to a probability measure as a **probability distribution** or simply distribution. The triple $(\Omega, \mathscr{F}, \mathbb{P})$ is called a **probability space**. The elements of $\mathscr{F}$ are called

events in the context of probability theory. It follows from the definitions that if (A_n) is a finite or countable sequence of events, then $\mathbb{P}(\cup_n A_n) \leq \sum_n \mathbb{P}(A_n)$. This is called the **union bound**. A statement is said to hold **almost surely**, or **a.s.** for short, if it holds on an event of probability 1. If $\mathscr{F} = \mathscr{B}(\Omega)$ is a Borel σ-algebra on Ω, then we call $(\Omega, \mathscr{B}(\mathbb{P}), \mathbb{P})$ a Borel probability space and $\mathbb{P}$ a Borel probability measure. Every probability measure is σ-finite since it is finite. If $\mathscr{m}$ is a σ-finite measure on $(\Omega, \mathscr{F})$, called the **reference measure**, and if $\mathbb{P} \ll \mathscr{m}$, then by the Radon-Nikodym theorem, there is a non-negative, integrable function p on Ω such that $\mathbb{P}(A) = \int_A p d\mathscr{m}$. The function p is called the **probability density function**, or simply probability density, of $\mathbb{P}$ with respect to $\mathscr{m}$. Since $\mathbb{P}(\Omega) = 1$, we must have $\int_\Omega p d\mathscr{m} = 1$. Observe that instead of specifying the distribution $\mathbb{P}$, one may as well specify the probability density p to define the distribution $\mathbb{P}$ implicitly.

A **random variable** is a measurable function between two measurable spaces. If $\mathrm{X}\colon \Omega_1 \to \Omega_2$ is a random variable and $\mathbb{P}_1$ is a probability distribution on $(\Omega_1, \mathscr{F}_1)$, we call $\mathbb{P}_2 = \mathrm{X}_* \mathbb{P}_1$, defined as $\mathbb{P}_2(B) = \mathbb{P}_1(\mathrm{X}^{-1}(B))$ for $B \in \mathscr{F}_2$, as the distribution on $(\Omega_2, \mathscr{F}_2)$ **induced** by the random variable X. Thus, when we talk about the distribution of a random variable taking values in Ω_2, there is some underlying probability space $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$, and we mean the distribution $\mathbb{P}_2 = \mathrm{X}_* \mathbb{P}_1$ on $(\Omega_2, \mathscr{F}_2)$. If we only care about the distribution $\mathbb{P}_2$, we may omit the underlying space $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$ from our discussion. If $\mathscr{m}_2$ is a σ-finite reference measure on $(\Omega_2, \mathscr{F}_2)$ and $\mathbb{P}_2 \ll \mathscr{m}_2$, we refer to the probability density function $d\mathbb{P}_2/d\mathscr{m}_2$ as the probability density of X. The **expected value** or the expectation value of a real-valued random variable X on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is defined as the integral $\mathbb{E}[X] = \int_\Omega X d\mathbb{P}$.

A special class of random variables, the discrete random variables, is of great importance in statistics and also in our study. First, we recall that the **discrete σ-algebra** on Ω is the power set $2^\Omega = \{A \mid A \subseteq \Omega\}$ of Ω. Clearly, the discrete σ-algebra is the largest σ-algebra one can put on Ω. Often, this σ-algebra is too large for most cases of interest (e.g., when $\Omega = \mathbb{R}$). However, if Ω happens to be a finite or countably infinite set, the discrete σ-algebra is the most natural choice of σ-algebra on Ω (which perhaps motivates the terminology "discrete" for this σ-algebra). In the same vein, 2^Ω is also a topology on Ω called the **discrete topology**. The σ-algebra generated by

a discrete topology is discrete. Observe that if $\mathbb{P}$ is any distribution on $(\Omega, 2^\Omega)$ and Ω is finite or countable, its action on 2^Ω is completely specified by its action on elements of Ω. In other words, if we know $\mathbb{P}(\omega) \equiv \mathbb{P}(\{\omega\})$ for all $\omega \in \Omega$, then we can calculate the probability of any event $A \subseteq \Omega$. Such distributions are called **discrete distributions**. With this in mind, a random variable is said to be **discrete** if it is a measurable function taking values in $(\Omega, 2^\Omega)$, where Ω is either a finite or a countably infinite set. It can be verified that the distribution on Ω induced by a discrete random variable is always a discrete distribution. Furthermore, observe that if $\mathscr{m}$ is the counting measure on $(\Omega, 2^\Omega)$ and Ω is finite or countable, then $\mathbb{P} \ll \mathscr{m}$ for all distributions $\mathbb{P}$ on $(\Omega, 2^\Omega)$. In this case, the probability density $d\mathbb{P}/d\mathscr{m}$ is just the function that maps $\omega \in \Omega$ to $\mathbb{P}(\omega)$. Thus, when working with discrete distributions, we sometimes refer to $d\mathbb{P}/d\mathscr{m}$ as the distribution instead of probability density. In the discrete case, we call Ω the **alphabet**, and the elements of Ω are called **symbols**. If Ω has d symbols, we can take $\Omega = [d]$ without loss of generality, by relabelling the symbols if necessary. The set of probability distributions on $[d]$ is the **standard simplex** in d dimensions, defined as $\Delta_d = \{p \in \mathbb{R}^d \mid p \geq 0, \sum_{i=1}^d p_i = 1\}$. We define the **support** of a distribution p as set $\operatorname{supp} p = \{i \in [d] \mid p_i > 0\}$.

Finally, we discuss the notion of an f-divergence, which quantifies the distance/dissimilarity between two distributions. They satisfy many desirable properties, and it can be shown that many known divergences or metrics are in fact f-divergences. This makes them important in probability theory. We borrow the following definition from [82, Ch. 7].

Definition 2.5 (f-divergence). Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions on $(\Omega, \mathscr{F})$ with densities p and q respectively with respect to a σ-finite reference measure $\mathscr{m}$. Then, the **f-divergence** between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$D_f(\mathbb{P}, \mathbb{Q}) = \int_{\{q>0\}} qf\left(\frac{p}{q}\right) d\mathscr{m} + f'(\infty)\mathbb{P}(\{q = 0\}), \tag{2.3}$$

where $f'(\infty) = \lim_{x\downarrow 0} xf(1/x)$. In this definition, we use the convention that $0 \times \infty = 0$, and $f(0) = \lim_{x\downarrow 0} f(x)$. □

Note that a reference measure m satisfying $\mathbb{P}, \mathbb{Q} \ll m$ always exists (e.g., by taking $m = (\mathbb{P} + \mathbb{Q})/2$). Importantly, the f-divergence between $\mathbb{P}$ and $\mathbb{Q}$ does not depend on the choice of the reference measure [82, Rem. 7.2]. Now, observe that if $\mathbb{P} \ll \mathbb{Q}$, then $\mathbb{P}(\{q = 0\}) = 0$ since $\mathbb{Q}(\{q = 0\}) = 0$. As a result, for $\mathbb{P} \ll \mathbb{Q}$, we have

$$D_f(\mathbb{P}, \mathbb{Q}) = \int_{\{q>0\}} qf\left(\frac{p}{q}\right) dm = \int_\Omega f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q}. \tag{2.4}$$

We refer the reader to [82, Ch. 7] for a list of properties satisfied by f-divergences.

An important special case is $f(x) = x\log(x)$. This gives rise to the well-known Kullback-Leibler (KL) divergence. If $\mathbb{P}, \mathbb{Q}$ are two distributions, then

$$\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \int_\Omega \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P} = \int_\Omega p\log\left(\frac{p}{q}\right) dm \tag{2.5}$$

if $\mathbb{P} \ll \mathbb{Q}$ and ∞ otherwise.

2.4 Statistics

Estimating parameters is an important task in physics, where one might wish to learn parameters of a physical model from experimental data. The problem of estimating parameters using observed data is an important topic in statistics, and has garnered a lot of attention in the recent past due to interest in machine learning. In this section, we will define what we mathematically mean by estimation, and also introduce concepts that are important in our study.

Suppose that we have an underlying probability space $(\Omega, \mathscr{F}, \mathbb{P}_{\text{true}})$, and we have access to outcomes in Ω sampled according to the distribution $\mathbb{P}_{\text{true}}$. The distribution $\mathbb{P}_{\text{true}}$ is not known to us, but we are given the promise that $\mathbb{P}_{\text{true}}$ lies in a known set of probability distributions $\mathscr{P}_0$. We are given a function $\mathfrak{p}\colon \mathscr{P}_0 \to \Theta$ taking values in some set Θ, and the quantity that we wish to estimate is $\mathfrak{p}(\mathbb{P}_{\text{true}})$. In statistics, the function $\mathfrak{p}$ is called, perhaps confusingly, a parameter (hence the symbol $\mathfrak{p}$). To avoid potential confusion, we will not refer to $\mathfrak{p}$ as a parameter in this study. To illustrate

the meaning of $\mathfrak{p}$, consider the following examples. Say $\mathscr{P}_0$ is the set of Gaussian distributions on the real line, where each $\mathbb{P}_{\mu,\sigma^2} \in \mathscr{P}_0$ is parametrized by the mean μ and variance σ^2. Then we can define $\mathfrak{p}(\mathbb{P}_{\mu,\sigma^2}) = (\mu, \sigma^2)$ to be the function that maps the distribution to its parameters. More generally, we can suppose that each distribution $\mathbb{P}_x \in \mathscr{P}_0$ is parameterized (not necessarily uniquely) by some vector $x \in \mathbb{R}^d$. Then, we can define $\mathfrak{p}(\mathbb{P}_x) = f(x)$ to be some function of the parameter x characterizing the distribution. This scenario will be important in our study. We remark that the definition of $\mathfrak{p}$ covers more general situations, beyond the scenario where distributions in $\mathscr{P}_0$ are parametrized by vectors in $\mathbb{R}^d$. For example, if $\mathscr{P}_0$ is the set of *all* distributions on the real line with finite mean, we can define $\mathfrak{p}(\mathbb{P})$ to be the mean of $\mathbb{P} \in \mathscr{P}_0$.

The typical goal in estimation is to learn the value of $\mathfrak{p}(\mathbb{P}_{\text{true}})$. Since we only have access to the outcomes in Ω sampled according to $\mathbb{P}_{\text{true}}$, and we don't know the distribution $\mathbb{P}_{\text{true}}$ itself, we need to use these outcomes to learn the value of $\mathfrak{p}(\mathbb{P}_{\text{true}})$. In the remainder of this section, we assume that Θ is a metric space with the metric $\mathscr{d}$, endowed with the Borel σ-algebra. A **point estimator** for $\mathfrak{p}\colon \mathscr{P}_0 \to \Theta$ is a measurable function $\widehat{\mathfrak{p}}\colon \Omega \to \Theta$. The idea here is that given an outcome $\omega \in \Omega$, sampled according to $\mathbb{P}_{\text{true}} \in \mathscr{P}_0$, the value $\widehat{\mathfrak{p}}(\omega)$ gives an estimate of the true value $\mathfrak{p}(\mathbb{P}_{\text{true}})$. If $\Theta \subseteq \mathbb{R}^d$, then the quantity $\mathfrak{p}(\mathbb{P}_{\text{true}}) - \mathbb{E}_{\text{true}}[\widehat{\mathfrak{p}}]$ is called the **bias** of $\widehat{\mathfrak{p}}$. The estimator $\widehat{\mathfrak{p}}$ is said to be **unbiased** if $\mathbb{E}_{\text{true}}[\widehat{\mathfrak{p}}] = \mathfrak{p}(\mathbb{P}_{\text{true}})$ for all $\mathbb{P}_{\text{true}} \in \mathscr{P}_0$. Point estimators are frequently used for estimation. For example, given n independent and identically distributed samples $X_1, \ldots, X_n$ of a random variable X, $\widehat{\mathfrak{p}} = \sum_{i=1}^n X_i/n$ is an unbiased point estimator of the mean of X.

In practice, due to only having access to a finite number of samples, there is always some error in estimating the true value. Thus, a point estimate by itself is not very useful, as we need to know, in addition, what the estimation error is. This leads us to the notion of a confidence set or a confidence region. A **confidence set assignment** for a confidence level of $1 - \delta \in [0, 1]$ is a *set-valued function* $\mathscr{C}$ from Ω to subsets of Θ, such that $\{\mathfrak{p}(\mathbb{P}_{\text{true}}) \in \mathscr{C}\} \in \mathscr{F}$ and $\mathbb{P}_{\text{true}}(\mathfrak{p}(\mathbb{P}_{\text{true}}) \in \mathscr{C}) > 1 - \delta$ for all $\mathbb{P}_{\text{true}} \in \mathscr{P}_0$. Given an observation $\omega \in \Omega$, the set $\mathscr{C}(\omega)$ is called a **confidence set** or a confidence region. The definition of a confidence set guarantees that no matter what the true distribution $\mathbb{P}_{\text{true}}$ is, the true value $\mathfrak{p}(\mathbb{P}_{\text{true}})$ lies in the confidence set with high probability. Thus, instead of

a point estimate, we output a *region* that contains the quantity of interest with high probability. We colloquially refer to any method that constructs a confidence set assignment as an estimation method/protocol/procedure.

In this study, we are mainly interested in two special cases, which we describe below. First, consider the case when the quantity to be estimated is real-valued. In this case, we construct a **confidence interval assignment** $\mathscr{C}$, such that $\mathscr{C}(\omega)$ is an interval for all $\omega \in \Omega$. If one has a point estimator $\widehat{\mathfrak{p}}$ for $\mathfrak{p}$, and there is some $\varepsilon > 0$ (possibly dependent on the data) such that $\mathbb{P}(|\widehat{\mathfrak{p}} - \mathfrak{p}(\mathbb{P})| \leq \varepsilon) > 1 - \delta$ for all $\mathbb{P} \in \mathscr{P}_0$, then $\mathscr{C} = [\widehat{\mathfrak{p}} - \varepsilon, \widehat{\mathfrak{p}} + \varepsilon]$ defines a confidence interval for $\mathfrak{p}$. For any observation $\omega \in \Omega$, the interval $\mathscr{C}(\omega) = [\widehat{\mathfrak{p}}(\omega) - \varepsilon(\omega), \widehat{\mathfrak{p}}(\omega) + \varepsilon(\omega)]$ is called a **confidence interval**. The second case of interest is estimating a vector-valued quantity with respect to l_∞-norm. Thus, we have $\Theta \subseteq \mathbb{R}^d$ and $\mathscr{d}(x, y) = \|x - y\|_\infty$. As before, if $\widehat{\mathfrak{p}}$ is a point estimator for $\mathfrak{p}$, and there is some $\varepsilon > 0$ (possibly dependent on data) such that $\mathbb{P}(\left\|\widehat{\mathfrak{p}} - \mathfrak{p}(\mathbb{P})\right\|_\infty \leq \varepsilon) > 1 - \delta$ for all $\mathbb{P} \in \mathscr{P}_0$, then $\mathscr{C} = \overline{B}(\widehat{\mathfrak{p}}, \varepsilon)$ is a confidence set assignment, where $\overline{B}(x, r) = \{y \in \Theta \mid \mathscr{d}(x, y) \leq r\}$ is the closed ball of radius $r > 0$ centered around $x \in \Theta$. This can be interpreted as *simultaneously* estimating all the components of $\mathfrak{p}(\mathbb{P}_{\text{true}})$ (which are real numbers) to within error ε with high probability.

In general, the error ε can depend on the observed data. This is often the case when using heuristics to compute the estimation error. For example, experiments in physics sometimes quote a standard deviation or compute bootstrap intervals, both of which compute errors from observed data. These methods are heuristic in the sense that the true value may *not* lie inside the computed interval/set with high probability, and thus, they do not give confidence set assignments in general. When the estimation error can be bounded by a constant, we call the confidence set assignment *minimax*. Formally, we say that $\mathscr{C}$ is **ϵ-minimax** if $\sup_{\omega \in \Omega} \operatorname{diam} \mathscr{C}(\omega) \leq 2\epsilon$, where $\operatorname{diam} A = \sup\{\mathscr{d}(x, y) \mid x, y \in A\}$ is the diameter of the set $A \subseteq \Theta$. In particular, if the diameter does not depend on the data, then the estimation procedure is minimax. For the case when $\mathscr{C} = [\widehat{\mathfrak{p}} - \varepsilon, \widehat{\mathfrak{p}} + \varepsilon]$, we have $\operatorname{diam} \mathscr{C}(\omega) = 2\varepsilon(\omega)$. Thus, if ε does not depend on the data, the estimation procedure is ε-minimax. A similar reasoning holds for the case when $\mathscr{C} = \overline{B}(\widehat{\mathfrak{p}}, \varepsilon)$.

Minimax confidence set assignments, or minimax methods, give worst-case error bounds,

because the same error is returned for all data points. Nevertheless, minimax methods have their advantages. For example, since the error is independent of the data, it can be computed *before* the start of an experiment. This can be particularly helpful in quantum information, where one can determine what measurements to perform before starting an experiment so as to minimize the estimation error. Another advantage is that we can compute the **sample complexity** for the estimation method, which is the number of samples needed to estimate the quantity of interest to a fixed error $\varepsilon > 0$ with probability greater than $1 - \delta$. It is helpful to know a priori the sample complexity of estimation methods in quantum information. This is because the dimension of a quantum system comprised of n qubits scales exponentially as 2^n, and thus, it is useful to know if the chosen estimation method is implementable for large system sizes. We note that many estimation methods proposed in the quantum information literature are minimax methods.

2.5 Convex analysis and optimization

Let $\mathscr{V}$ be a finite dimensional vector space. A set $\mathscr{A} \subseteq \mathscr{V}$ is said to be **affine** if for all $x, y \in \mathscr{A}$ and all $\lambda \in \mathbb{R}$, we have $\lambda x + (1 - \lambda)y \in \mathscr{A}$. It can be shown that an affine set is the translation of a linear subspace in the sense that there is a linear subspace $\mathscr{U} \subseteq \mathscr{V}$ such that $\mathscr{A} = \mathscr{U} + a$ for all $a \in \mathscr{A}$, where $\mathscr{U} + a = \{u + a \mid u \in \mathscr{U}\}$. The **affine hull** of a set $\mathscr{K} \subseteq \mathscr{V}$, denoted $\operatorname{aff} \mathscr{K}$, is the smallest affine subset of $\mathscr{V}$ containing $\mathscr{K}$. It can be shown that $\operatorname{aff} \mathscr{K} = \{\sum_{i=1}^n \lambda_i v_i \mid n \in \mathbb{N}, \lambda_1, \ldots, \lambda_n \in \mathbb{R}, \sum_{i=1}^n \lambda_i = 1, v_1, \ldots, v_n \in \mathscr{K}\}$.

A set $\mathscr{C} \subseteq \mathbb{R}^d$ is said to be **convex** if for all $x, y \in \mathscr{C}$ and $\lambda \in [0, 1]$, we have $\lambda x + (1 - \lambda)y \in \mathscr{C}$. All affine sets are convex, but the converse need not be true. The **convex hull** of a set $\mathscr{K} \subseteq \mathscr{V}$, denoted $\operatorname{conv} \mathscr{K}$, is the smallest convex subset of $\mathscr{V}$ containing $\mathscr{K}$. It can be shown that $\operatorname{conv} \mathscr{K} = \{\sum_{i=1}^n \lambda_i v_i \mid n \in \mathbb{N}, \lambda_1, \ldots, \lambda_n \in [0, 1], \sum_{i=1}^n \lambda_i = 1, v_1, \ldots, v_n \in \mathscr{K}\}$.

An important notion in finite-dimensional convex analysis is the notion of a relative interior. To motivate the definition, consider the d-dimensional standard simplex $\Delta_d = \{x \in \mathbb{R}^d \mid x \geq 0, \sum_{i=1}^d x_i = 1\}$. It can be shown that the interior of Δ_d in $\mathbb{R}^d$ is empty (think of a simplex in 3 dimensions for visualization). However, it still makes sense to look at the interior of Δ_d

with respect to the affine subspace containing Δ_d, and this is exactly the notion of relative interior. Formally, if $\mathscr{K} \subseteq \mathbb{R}^d$ is any set, then we define the **relative interior** of $\mathscr{K}$ as the set $\operatorname{relint} \mathscr{K} = \{x \in \mathscr{K} \mid (\exists \epsilon > 0)\ B(x, \epsilon) \cap \operatorname{aff} \mathscr{K} \subseteq \mathscr{K}\}$, where $B(x, \epsilon)$ is the l_2-ball of radius ϵ around x. Topologically, one can think of the relative interior of $\mathscr{K}$ as the interior of $\mathscr{K}$ with respect to the subspace topology induced by $\operatorname{aff} \mathscr{K}$. $\mathscr{K}$ is said to be **relatively open** if $\mathscr{K} = \operatorname{relint} \mathscr{K}$. It can be shown that $\operatorname{relint} \Delta_d = \{x \in \mathscr{K} \mid x > 0, \sum_{i=1}^d x_i = 1\}$, which we call the relatively open simplex. Importantly, if $\mathscr{C} \subseteq \mathbb{R}^d$ is a non-empty convex set, then $\operatorname{relint} \mathscr{C} \neq \varnothing$ [8, Fact 6.14].

A function $\mathrm{A} \colon \mathscr{V} \to \mathscr{W}$ from a vector space $\mathscr{V}$ to a vector space $\mathscr{W}$ is said to be an **affine function** if for all $x, y \in \mathscr{V}$ and $\lambda \in \mathbb{R}$, we have $\mathrm{A}(\lambda x + (1-\lambda)y) = \lambda \mathrm{A}(x) + (1-\lambda)\mathrm{A}(y)$. It can be shown that an affine function is just the translation of a linear function, that is, there is some linear map $\mathrm{L} \colon \mathscr{V} \to \mathscr{W}$ and a vector $w \in \mathscr{W}$ such that $\mathrm{A}(x) = \mathrm{L}(x) + w$ for all $x \in \mathscr{V}$. An extended real-valued function $f \colon \mathscr{C} \to \overline{\mathbb{R}}$ defined on a convex set $\mathscr{C}$ is said to be a **convex function** if for all $x, y \in \mathscr{C}$ and $\lambda \in [0, 1]$, we have $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. A function f is said to be **concave** if $-f$ is convex, while it is said to be **log-concave** if $\log(f)$ is concave. Every concave function is log-concave but the converse need not be true. A real-valued affine function is both convex and concave.

A function $f \colon \mathscr{K} \to \overline{\mathbb{R}}$ defined on a set $\mathscr{K} \subseteq \mathbb{R}^d$ is said to be a **proper** if f never takes the value $-\infty$, and there is some point $\mathscr{K}$ at which f is finite. A function $f \colon \mathscr{K} \to \overline{\mathbb{R}}$ is said to be **lower semi-continuous** or **lsc** at a point $x_\circ \in \mathscr{K}$ if $\liminf_{x \to x_\circ} f(x) \geq f(x_\circ)$. A function $f \colon \mathscr{K} \to \overline{\mathbb{R}}$ is said to be **upper semi-continuous** or **usc** at $x_\circ \in \mathscr{K}$ if $\limsup_{x \to x_\circ} f(x) \leq f(x_\circ)$. f is said to be lsc/usc if it is lsc/usc at every point in $\mathscr{K}$. A function $f \colon \mathscr{K} \to \overline{\mathbb{R}}$ is said to be **coercive** if for every sequence (x_n) in $\mathscr{K}$ with $\|x_n\| \to \infty$, we have $f(x_n) \to \infty$. It is well-known that if $\mathscr{K}$ is a closed convex set, and f is a proper, lsc, convex, coercive function on $\mathscr{K}$, then f has a minimizer over $\mathscr{K}$ [8, Thm. 11.15].

Given a set $\mathscr{K}$, its **characteristic function** is defined in convex analysis as

$$\chi_{\mathscr{K}}(x) = \begin{cases} 0 & \text{if } x \in \mathscr{K}, \\ \infty & \text{otherwise.} \end{cases} \tag{2.6}$$

We note that the definition of a characteristic (or indicator) function in convex analysis is different from that used in probability theory and representation theory. Observe that $\chi_{\mathscr{K}}$ just encodes the set $\mathscr{K}$ as a function. It can be verified that if $\mathscr{K} \subseteq \mathbb{R}^d$ is a non-empty convex set, then $\chi_{\mathscr{K}} \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is a proper convex function. If $\mathscr{K}$ is a closed set, then $\chi_{\mathscr{K}}$ is lower semi-continuous. Another important function is the support function of a set. Given $\mathscr{K} \subseteq \mathbb{R}^d$, its **support function** is defined as

$$S_{\mathscr{K}}(x) = \sup\{\langle x, y\rangle \mid y \in \mathscr{S}\}. \tag{2.7}$$

It can be verified that $S_{\mathscr{K}}$ is always a convex function, irrespective of whether or not $\mathscr{K}$ is convex. If $\mathscr{K}$ is a closed convex set, then $S_{\mathscr{K}}$ is lsc, and if $\mathscr{K}$ is a bounded convex set, then $S_{\mathscr{K}}$ does not take the value ∞. The characteristic function and the support function of a set are dual to each other in the sense defined below.

Definition 2.6 (Convex conjugate). The **convex conjugate** or the Legendre-Fenchel transform of a function $f \colon \mathscr{X} \to \overline{\mathbb{R}}$ defined on $\mathscr{X} \subseteq \mathbb{R}^d$ is defined as

$$f^*(y) = \sup_{x \in \mathbb{R}^d} (\langle y, x\rangle - f(x)) \tag{2.8}$$

for $y \in \mathbb{R}^d$. □

Convex conjugate generalizes the notion of Legendre transform that is frequently used in physics, especially thermodynamics and classical mechanics. It can be shown that f^* is a convex function even when f is not. If f is a proper lsc convex function, then f^* is also a proper lsc convex function and we have $f^{**} = f$ [8, Cor. (13.38)]. One can verify that if $\mathscr{K}$ is a non-empty closed convex set, then $S_{\mathscr{K}} = \chi^*_{\mathscr{K}}$ and $\chi_{\mathscr{K}} = S^*_{\mathscr{K}}$. Thus, for a closed convex set, its support function

encodes the set as a function, just as the characteristic function does.

Finally, we define another useful transformation that will appear later in our study. Given a proper function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$, its **perspective** is the function $\mathfrak{p}_f\colon \mathbb{R}^d \times (0,\infty) \to \overline{\mathbb{R}}$ defined as $\mathfrak{p}_f(x,t) = tf(x/t)$. It can be shown that if f is a convex function, then $\mathfrak{p}_f$ is also a (jointly) convex function.

We now turn our attention from convex analysis to convex optimization. The discussion below follows the exposition in [17]. Consider the following optimization problem in the so-called standard form:

$$\begin{aligned} (\mathrm{P}) \quad \min_{x\in\mathscr{X}} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i \in [n] \\ & h_j(x) = 0, \quad j \in [m]. \end{aligned} \tag{2.9}$$

The function f_0 is called the objective function. The functions $f_1, \ldots, f_n$ define inequality constraints, while the functions $h_1, \ldots, h_m$ define equality constraints. $\mathscr{X} \subseteq \mathbb{R}^d$ is a set over which all of these functions are well-defined. The optimization problem written above is called the **primal problem**, in contrast with its dual problem we will define below. If the functions $f_0, f_1, \ldots, f_n$ are convex, $h_1, \ldots, h_m$ are affine, and the set $\mathscr{X}$ is convex, the above problem is called a convex optimization problem. This is because it amounts to minimizing the convex function f_0 over a convex set determined by the constraints. The optimal value p^* of (P) is called the **primal optimal value**.

An important property of convex functions is that all local minima of a convex function are also global minima. Thus, it suffices to compute the local minima of convex functions. Usually, this is done by looking at the points where the gradient of f_0 is zero, but this does not account for the constraints in the optimization problem. To remedy this, one defines a function called the Lagrangian that explicitly depends on the functions $f_1, \ldots, f_n$ and $h_1, \ldots, h_m$ defining the constraints. The **Lagrangian** for the primal problem (P) is defined as

$$\mathscr{L}(x;\lambda,\nu) = f_0(x) + \sum_{i=1}^{n} \lambda_i f_i + \sum_{j=1}^{m} \nu_j h_j. \tag{2.10}$$

The variable $x \in \mathbb{R}^d$ is called the **primal variable**, while the variables $\lambda \in \mathbb{R}^n$ and $\nu \in \mathbb{R}^n$ are called **dual variables**. The (Lagrange) **dual function** for the problem (P) is defined as

$$h(\lambda, \nu) = \min_{x \in \mathscr{X}} \mathscr{L}(x; \lambda, \nu). \tag{2.11}$$

The dual function is a concave function of (λ, ν), even if the primal optimization problem is not convex. Importantly, one can show that $h(\lambda, \nu) \leq p^*$ for all $\lambda \geq 0$ and $\nu \in \mathbb{R}^m$. Using this observation, we define the (Lagrange) **dual problem** of the primal problem as

$$\begin{aligned} (\mathrm{D}) \quad \max_{\lambda, \nu} \quad & h(\lambda, \nu) \\ & \lambda \geq 0. \end{aligned} \tag{2.12}$$

The optimal value d^* of (D) is called the **dual optimal value**. Since $h(\lambda, \nu) \leq p^*$ for all $\lambda \geq 0$ and ν, it follows that $d^* \leq p^*$. This is called **weak duality**. The difference $p^* - d^*$ is called the duality gap. When $d^* = p^*$, or equivalently, when the duality gap is zero, we say that **strong duality** holds.

We now describe first-order optimality conditions, called **Karush-Kuhn-Tucker (KKT)** conditions. Suppose that $f_0, \ldots, f_n$ and $h_1, \ldots, h_m$ are differentiable on an open set containng $\mathscr{X}$. Then, the points $x^* \in \mathscr{X}$ and $\lambda^* \in \mathbb{R}^n$, $\nu^* \in \mathbb{R}^m$ are said to satisfy the KKT conditions if the following hold:

(1) (Primal feasibility) $f_i(x^*) \leq 0$ for $i \in [n]$ and $h_j(x^*) = 0$ for $j \in [m]$.

(2) (Dual feasiblity) $\lambda^* \geq 0$.

(3) (Complementary slackness) $\lambda_i^* f_i(x^*) = 0$ for all $i \in [n]$.

(4) (Stationarity) The gradient of the Lagrangian vanishes at $(x^*; \lambda^*, \nu^*)$, i.e.,

$$\nabla f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{m} \nu_j^* \nabla h_j(x^*) = 0. \tag{2.13}$$

It can be shown that if $x^* \in \mathscr{X}$ and $\lambda^* \in \mathbb{R}^n_+$ and $\nu^* \in \mathbb{R}^m$ are primal and dual optimal points with zero duality gap, then $(x^*; \lambda^*, \nu^*)$ necessarily satisfy the KKT conditions. The converse need not hold. That is, KKT conditions are, in general, not sufficient to ensure optimality.

However, if the primal problem (P) is convex, then it can be shown that KKT conditions are sufficient. Thus, for a convex optimization problem, KKT conditions are necessary and sufficient if strong duality holds. A condition that guarantees strong duality for convex problems is Slater's condition. **Slater's condition** says that if the primal problem (P) is convex and there is at least one feasible point $x_\circ \in \operatorname{relint} \mathscr{X}$ satisfying $f_i(x_\circ) < 0$ for all $i \in [n]$, then strong duality holds. Thus, if we can show that Slater's condition holds for a convex problem, we can use KKT conditions to find primal and dual optimal points.

Chapter 3

Classical and quantum distance measures

We begin our study by defining some distance measures for classical probability distributions and quantum states that are relevant to our study. When we say "distance measure", we mean some function that says how close or similar two probability distributions/quantum states are. Given a measurement protocol and a quantum state, one obtains classical probability distributions for the observed outcomes through the Born's rule. We call the distance measure one can define between two quantum states through such classical probability distributions as a classical distance measure on quantum states. A quantum distance measure can be obtained by optimizing the classical distance measure over all measurement protocols.

The main distance measure of interest in our study is the average Bhattacharyya distance between two quantum states determined by a measurement protocol. We will study the relation of this measure to its quantum counterpart, (half) negative log-fidelity, as well as other well-studied classical and quantum distance measures. Many of the results we present in this chapter review known results in the literature, though possibly in a different form. Two results that we would like to highlight in this chapter, which may find applications elsewhere, are (1) a closed-form expression for the convex conjugate of Bhattacharyya distance for a special case of interest (Prop. 3.9), and (2) a continuity bound for quantum fidelity (Prop. 3.17).

3.1 Bhattacharyya distance and quantum fidelity

We start the discussion with the familiar notion of Bhattacharyya coefficient between two classical probability distributions [12].

Definition 3.1 (Bhattacharyya coefficient and classical fidelity). The **Bhattacharyya coefficient** (or Hellinger affinity) between two probability distributions p and q over M symbols is defined as

$$\mathrm{BC}(p,q) = \sum_{i=1}^{M} \sqrt{p_i q_i}. \tag{3.1}$$

The **classical fidelity** between p and q is defined as

$$\mathrm{FC}(p,q) = (B(p,q))^2. \tag{3.2}$$

□

The Bhattacharyya coefficient is the classical counterpart of square-root fidelity [36]. It is a number between 0 and 1, equal to 1 iff $p = q$, and equal to 0 iff p and q have disjoint support. Furthermore, it is a jointly concave function of its arguments [106, Cor. 3.26]. While the Bhattacharyya coefficient itself is not a metric, one can define different metrics using it. In our study, the closely related notion of Bhattacharyya distance [12, 63] is important.

Definition 3.2 (Bhattacharyya distance). The **Bhattacharyya distance** between two probability distributions p and q is defined as

$$\mathrm{BD}(p,q) = -\log(\mathrm{BC}(p,q)). \tag{3.3}$$

□

Note that the Bhattacharyya distance is not a metric. However, it has some useful properties, which we list below.

Proposition 3.3 (Properties of Bhattacharyya distance). *1. For any distributions p, q, we have $0 \leq \mathrm{BD}(p,q) \leq \infty$, with $\mathrm{BD}(p,q) = 0$ iff $p = q$ and $\mathrm{BD}(p,q) = \infty$ iff p and q have disjoint support.*

2. BD *is a proper jointly convex function.*

3. BD *is additive for product distributions. That is, given $p, q \in \Delta_M$ and $p', q' \in \Delta_N$, we have $\mathrm{BD}(p \otimes p', q \otimes q') = \mathrm{BD}(p,q) + \mathrm{BD}(q,q')$.*

Proof. 1. Follows from the definition of BD and properties of BC.

2. Since BC is concave, and hence log-concave, BD is convex. It is proper because $\mathrm{BD} \geq 0$ and $\mathrm{BD}(p,p) = 0$ for any distribution p.

3. Observe that

$$\mathrm{BC}(p \otimes p', q \otimes q') = \sum_i \sqrt{p_i q_i} \sum_j \sqrt{p'_j q'_j} = \mathrm{BC}(p,q)\mathrm{BC}(p',q'). \tag{3.4}$$

The additivity of BD follows by taking negative logarithm on both sides of the above equation. □

Since independent random variables give rise to product distributions on the large (product) space over which all the random variables are defined, the Bhattacharyya distance gives us additivity for distributions obtained from independent measurements. The Bhattacharyya coefficient, on the other hand, is multiplicative. This motivates us the define the following classical distance measures between two quantum states determined by a measurement protocol.

Definition 3.4 (Average Bhattacharyya distance). Given a measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$, the **average Bhattacharyya distance** between two quantum states ρ and σ determined by $\mathfrak{M}$ is defined as

$$\mathrm{BD}_{\mathfrak{M}}(\rho,\sigma) = \sum_{i=1}^{L} \frac{N_i}{N} \mathrm{BD}(p_\rho^{(i)}, p_\sigma^{(i)}), \tag{3.5}$$

where $p_\rho^{(i)}$ is the probability for the ith POVM with respect to the state ρ given by Born's rule, and $N = \sum_{i=1}^{L} N_i$ is the total number of samples.

Similarly, we define the **geometric-average Bhattacharyya coefficient** between ρ and σ determined by $\mathfrak{M}$ as

$$\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = \prod_{i=1}^{L} \left(\mathrm{BC}(p_\rho^{(i)}, p_\sigma^{(i)}) \right)^{N_i/N} . \tag{3.6}$$

□

We define the **geometric-average classical fidelity** determined by $\mathfrak{M}$ as the square of the geometric-average Bhattacharyya coefficient, i.e., $\mathrm{FC}_{\mathfrak{M}}(\rho, \sigma) = \mathrm{BC}^2_{\mathfrak{M}}(\rho, \sigma)$ for $\rho, \sigma \in \mathscr{X}$. Thus, the statements concerning the Bhattacharyya coefficient can be translated to classical fidelity and vice-versa. As for the case with classical probability distributions, the average Bhattacharyya distance and geometric-average Bhattacharyya coefficient do not define a metric on the set of quantum states. Nevertheless, they are closely related to distance measures on quantum states that are pseudometrics. Thm. 7.13 shows that the average Bhattacharyya distance determines the optimal performance one can get for estimating expectation values of observables using outcomes of the measurement protocol $\mathfrak{M}$, which underlines the importance of this distance measure. This motivates us to study its properties and its relation to other distance measures commonly used in the quantum information literature.

We begin by studying the quantum counterparts of $\mathrm{BC}_{\mathfrak{M}}$ and $\mathrm{BD}_{\mathfrak{M}}$. To that end, we recall the definition of fidelity between two quantum states.

Definition 3.5 (Fidelity). The **fidelity** between the quantum states ρ and σ is defined as

$$F(\rho, \sigma) = \left(\mathrm{Tr}(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}) \right)^2 . \tag{3.7}$$

□

We denote the **square-root fidelity** as $\sqrt{F}$ and **log fidelity** as $\log F$. We review some basic properties of fidelity that are well-known in the literature. See [68, 106, 107] for other properties.

Proposition 3.6 (Properties of fidelity). *Let ρ, σ be two quantum states. The following statements hold.*

1. $F(\rho, \sigma) = \left\| \sqrt{\rho}\sqrt{\sigma} \right\|_1^2$.
2. $F(\rho, \sigma)$ *lies between* 0 *and* 1. *It is equal to* 1 *iff* $\rho = \sigma$ *and equal to* 0 *iff* $\mathrm{Tr}(\rho\sigma) = 0$.
3. $F(\rho, \sigma) = F(\sigma, \rho)$.
4. $\sqrt{F}$ *is jointly concave, while* F *is jointly log-concave.*
5. *If* ρ *or* σ *is pure, then* $F(\rho, \sigma) = \mathrm{Tr}(\rho\sigma)$.
6. *If* ρ *and* σ *commute, then* $F(\rho, \sigma) = \mathrm{FC}(\lambda(\rho), \lambda(\sigma))$, *where* $\lambda(\rho), \lambda(\sigma)$ *denote the vector of eigenvalues of* ρ, σ *respectively.*

Proof. 1. Follows from definitions.

2. By the Fuchs-van de Graaf inequality [36] (see Eq. (3.35)), we have $F(\rho, \sigma) = 1$ iff $\|\rho - \sigma\|_1 = 0$ iff $\rho = \sigma$. On the other hand, $F(\rho, \sigma) = 0$ iff $\sqrt{\rho}\sqrt{\sigma} = 0$ iff $\rho\sigma = 0$ iff $\mathrm{Tr}(\rho\sigma) = 0$.

3. This follows from Uhlmann's theorem [102], [77, Thm. 9.4].

4. [106, Cor. 3.26] shows that $\sqrt{F}$ is concave, from which it follows that $\log F$ is concave.

5. If $\rho = |\psi\rangle\langle\psi|$, then $\sqrt{\rho}\sigma\sqrt{\rho} = \langle\psi|\sigma|\psi\rangle\,|\psi\rangle\langle\psi| = \mathrm{Tr}(\rho\sigma)\rho$. It follows that $F(\rho, \sigma) = \mathrm{Tr}(\rho\sigma)$.

6. If ρ and σ commute, then $\sqrt{\rho}\sigma\sqrt{\rho} = \rho\sigma$ and $\sqrt{\rho\sigma} = \sqrt{\rho}\sqrt{\sigma}$. The latter claim can be verified by squaring both sides and using uniqueness of matrix square-root. Then, $F(\rho, \sigma) = (\mathrm{Tr}(\sqrt{\rho}\sqrt{\sigma}))^2$. Since ρ and σ commute, they can be diagonalized in a common orthonormal basis. Evaluating the trace in this basis gives the desired result. □

We saw in Prop. 3.6 that when ρ and σ commute, the fidelity between ρ and σ coincides with the classical fidelity between the spectrums of ρ and σ. When ρ and σ commute, they can be simultaneously diagonalized, and therefore, $\lambda(\rho)$ and $\lambda(\sigma)$ are just the probabilities observed upon measuring ρ and σ in their common eigenbasis. Thus, we have shown that when ρ and σ commute, the fidelity between them can be realized as the classical fidelity between the outcome

probabilities for a particular measurement. Now, the question arises whether this observation can be generalized to the case when ρ and σ don't commute. This question was answered in the affirmative by [35, 34], who showed that the fidelity is the minimum of the classical fidelity between the outcome probabilities over all measurements, and that there is some POVM that achieves this minimum. We can use this observation to give the quantum counterparts of average Bhattacharyya distance and geometric-average Bhattacharyya coefficient.

Proposition 3.7. *If ρ and σ are two quantum states, then the following hold.*

1.

$$\sqrt{F}(\rho,\sigma) = \min_{\mathfrak{M}} \mathrm{BC}_{\mathfrak{M}}(\rho,\sigma). \tag{3.8}$$

2.

$$F(\rho,\sigma) = \min_{\mathfrak{M}} \mathrm{FC}_{\mathfrak{M}}(\rho,\sigma). \tag{3.9}$$

3.

$$-\frac{1}{2}\log F(\rho,\sigma) = \max_{\mathfrak{M}} \mathrm{BD}_{\mathfrak{M}}(\rho,\sigma). \tag{3.10}$$

There is a measurement protocol that achieves the minimum in all the above equations.

Proof. 1. From [35, 34], we know that $\sqrt{F}(\rho,\sigma) = \min_{\boldsymbol{E}} \mathrm{BC}(p_{\boldsymbol{E},\rho}, p_{\boldsymbol{E},\sigma})$, where the minimization is over all POVMs. In particular, $\mathrm{BC}(p_{\boldsymbol{E},\rho}, p_{\boldsymbol{E},\sigma}) \geq \sqrt{F}(\rho,\sigma)$ for every POVM $\boldsymbol{E}$. It then follows from the definition of $\mathrm{BC}_{\mathfrak{M}}$ that $\mathrm{BC}_{\mathfrak{M}}(\rho,\sigma) \geq \sqrt{F}(\rho,\sigma)$ for any measurement protocol. Since there is a POVM $\boldsymbol{E}_*$ such that $\sqrt{F}(\rho,\sigma) = \mathrm{BC}(p_{\boldsymbol{E}_*,\rho}, p_{\boldsymbol{E}_*,\sigma})$ [34], the measurement protocol $\mathfrak{M}_* = \{(\boldsymbol{E}_*, 1)\}$ achieves the minimum in Eq. (3.8).

2. Since $\mathrm{FC}_{\mathfrak{M}}$ is the square of $\mathrm{BC}_{\mathfrak{M}}$ and $\mathrm{BC}_{\mathfrak{M}}$ is non-negative, the result follows by squaring both sides of Eq. (3.8).

3. Since $\mathrm{BD}_{\mathfrak{M}}$ is the negative logarithm of $\mathrm{BC}_{\mathfrak{M}}$, and $-\log(x)$ is a strictly decreasing function, we obtain Eq. (3.10) from Eq. (3.8). Note that both sides of Eq. (3.10) can be infinity. □

Next, we list some basic properties of average Bhattacharyya distance and geometric-average Bhattacharyya coefficient following Prop. 3.6.

Proposition 3.8 (Properties of average Bhattacharyya distance). *Let ρ, σ be quantum states, and let $\mathfrak{M}$ be a measurement protocol. Then, the following statements hold.*

1. $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma)$ *is bounded between* 0 *and* 1. $\rho = \sigma$ *implies* $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = 1$, *while* $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = 0$ *implies* $\mathrm{Tr}(\rho\sigma) = 0$.

2. $\mathrm{BD}_{\mathfrak{M}}(\rho, \sigma)$ *is bounded between* 0 *and* ∞. $\rho = \sigma$ *implies* $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = 0$, *while* $\mathrm{BD}_{\mathfrak{M}}(\rho, \sigma) = \infty$ *implies* $\mathrm{Tr}(\rho\sigma) = 0$.

3. $\mathrm{BD}_{\mathfrak{M}}(\rho, \sigma) = \mathrm{BD}_{\mathfrak{M}}(\sigma, \rho)$ *and* $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = \mathrm{BC}_{\mathfrak{M}}(\sigma, \rho)$.

4. $\mathrm{BD}_{\mathfrak{M}}$ *is a proper convex function, while* $\mathrm{BC}_{\mathfrak{M}}$ *is a log-concave function.*

Proof. 1. $\mathrm{BC}_{\mathfrak{M}}$ is the geometric mean of numbers bounded between 0 and 1, and hence also bounded between 0 and 1. Direct computation shows that $\rho = \sigma$ implies $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma)$. If $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = 0$, then by Prop. 3.7, we have $F(\rho, \sigma) = 0$. By Prop. 3.6, this implies $\mathrm{Tr}(\rho\sigma) = 0$.

2. Follows from (1).

3. Follows from the definitions.

4. Follows from the definitions and Prop. 3.3. □

The converse of Prop. 3.8.1 and Prop. 3.8.2 does not hold in general. It can be shown, however, that if $\mathfrak{M}$ is informationally complete, then $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = 1$ implies $\rho = \sigma$. On the other hand, informational completeness is not sufficient to ensure the claim that $\mathrm{Tr}(\rho\sigma) = 0 \implies \mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) = 0$. To see this, take $\rho = |0\rangle\langle 0|$ and $\sigma = |1\rangle\langle 1|$ for a 1-qubit system. If $\mathfrak{M}$ corresponds to randomly sampling from $\{X, Y, Z\}$ and then measuring it, it can be checked that $\mathrm{BC}_{\mathfrak{M}}(\rho, \sigma) \neq 0$, even though $\mathfrak{M}$ is informationally complete.

Before ending this section, we note down some properties of the convex conjugate of the Bhattacharyya distance. We use the following notations. For any vector $u \in \mathbb{R}^M$, we write

$u_{\max} = \max_i u_i$ and $u_{\min} = \min_i u_i$. We denote $\operatorname{argmax} u = \{i \in [M] \mid u_i = u_{\max}\}$ and $\operatorname{argmin} u = \{j \in [M] \mid u_j = u_{\min}\}$.

Proposition 3.9 (Convex conjugate of Bhattacharyya distance). *The convex conjugate* $\mathrm{BD}^*\colon \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ *of Bhattacharyya distance satisfies the following properties.*

1. *For* $u \in \mathbb{R}^M$, *denoting* $\Upsilon = u_{\max} - u_{\min}$, *we have*

$$\mathrm{BD}^*(u,-u) = \begin{cases} \sqrt{\frac{1+2\Upsilon^2-\sqrt{1+4\Upsilon^2}}{2}} + \frac{1}{2}\log\left(\frac{\sqrt{1+4\Upsilon^2}-1}{2\Upsilon^2}\right) & \text{if } \Upsilon > 0 \\ 0 & \text{if } \Upsilon = 0. \end{cases} \tag{3.11}$$

   Furthermore, $\mathrm{BD}^*(u,-u) \geq 0$ *and* $\mathrm{BD}^*(u,-u)$ *is a convex function of* $u \in \mathbb{R}^M$.

2. *For all* $u, v \in \mathbb{R}^M$, *we have* $\mathrm{BD}^*(u,v) = \mathrm{BD}^*(v,u)$.

3. *For* $u, v \in \mathbb{R}^M$, *we have*

$$\begin{aligned} &\frac{1}{2}\left(\max_{i \in \operatorname{argmax}(u-v)} (u+v)_i + \max_{j \in \operatorname{argmin}(u-v)} (u+v)_j\right) + \mathrm{BD}^*\left(\frac{(u-v)}{2}, -\frac{(u-v)}{2}\right) \\ &\leq \mathrm{BD}^*(u,v) \\ &\leq (u+v)_{\max} + \mathrm{BD}^*\left(\frac{(u-v)}{2}, -\frac{(u-v)}{2}\right). \end{aligned} \tag{3.12}$$

4. *For all* $p, q \in \Delta_M$ *and all* $u, v \in \mathbb{R}^M$, *we have* $\mathrm{BD}(p,q) + \mathrm{BD}^*(u,v) \geq \langle u, p\rangle + \langle v, q\rangle$.

Proof. 1. Since BD is only defined on the set of probability distributions, the expression for BD^* according to Def. 2.6 becomes

$$\begin{aligned} \mathrm{BD}^*(u,-u) &= \sup_{p,q \in \Delta_M} \left(\langle u, p\rangle - \langle u, q\rangle + \log(\mathrm{BC}(p,q))\right) \\ &= \frac{1}{2} \sup_{p,q \in \Delta_M} \left(2\langle u, p\rangle - 2\langle u, q\rangle + \log(\mathrm{FC}(p,q))\right), \end{aligned} \tag{3.13}$$

where we used the fact that $\mathrm{FC} = \mathrm{BC}^2$ to obtain the second equality. Denote $x - a = (x_1 - a, \ldots, x_M - a)$ for $a \in \mathbb{R}$ and $x \in \mathbb{R}^M$. Then, for all distributions p, q, we have $\langle u, p - q\rangle = \langle u - u_{\min}, p - q\rangle$.

Now, denote $\mathscr{F} = \{i \in [M] \mid (p_i - q_i) > 0$. Since $0 \leq u_i - u_{\min} \leq \Upsilon$ for all i, we have $\langle u - u_{\min}, p - q\rangle \leq \sum_{i\in\mathscr{F}}(u_i - u_{\min})(p_i - q_i) \leq \Upsilon \sum_{i\in\mathscr{F}}(p_i - q_i)$. Since $\sum_i (p_i - q_i) = 0$ and $\sum_{i\in\mathscr{F}}(p_i - q_i) - \sum_{i\in[M]\setminus\mathscr{F}}(p_i - q_i) = \|p - q\|_1$, we have $\sum_{i\in\mathscr{F}}(p_i - q_i) = \|p - q\|_1/2$. Thus, we obtain $\langle u - u_{\min}, p - q\rangle \leq \Upsilon \|p - q\|_1/2$. Furthermore, by Fuchs-van de Graaf inequality [36], we have $\|p - q\|_1/2 \leq \sqrt{1 - \mathrm{FC}(p, q)}$. Thus, we obtain the inequality

$$\begin{aligned}\mathrm{BD}^*(u, -u) &\leq \frac{1}{2} \sup_{p,q\in\Delta_M} (2\Upsilon\sqrt{1 - \mathrm{FC}(p, q)} + \log(\mathrm{FC}(p, q))) \\ &\leq \frac{1}{2} \sup_{x\in[0,1]} (2\Upsilon\sqrt{1 - x} + \log(x)),\end{aligned} \tag{3.14}$$

where the last inequality follows from the fact that $\mathrm{FC}(p, q) \in [0, 1]$ for all $p, q \in \Delta_M$. If $\Upsilon = 0$, then the maximum is achieved at $x = 1$, and the maximum value is equal to 0. This can be achieved in Eq. (3.13) by choosing $p, q \in \Delta_M$ with $p = q$, and therefore, $\mathrm{BD}^*(u, -u) = 0$ in this case. Thus, we assume that $\Upsilon > 0$. The function $f(x) = 2\Upsilon\sqrt{1 - x} + \log(x)$ takes the value $-\infty$ at $x = 0$ and the value 0 at $x = 1$. The derivative of f is given by $f'(x) = -\Upsilon/\sqrt{1 - x} + 1/x$, so that $f'(1) = -\infty$. Thus, the maximum cannot occur at either $x = 0$ or $x = 1$. Since f is a strictly concave function, it has a unique maximum in $(0, 1)$, which can be obtained by setting its derivative to 0. Rearranging $f'(x) = 0$, we obtain

$$\Upsilon^2 x^2 + x - 1 = 0. \tag{3.15}$$

After discarding the negative solution, we obtain

$$x^* = \frac{-1 + \sqrt{1 + 4\Upsilon^2}}{2\Upsilon^2}. \tag{3.16}$$

It can be verified that $x^* \in (0, 1)$. Now, choose $p^*, q^* \in \Delta_M$ as follows. Fix $i \in \mathrm{argmax}(u)$ and $j \in \mathrm{argmin}(u)$, and take $p_i^* = (1 + \sqrt{1 - x^*})/2$, $p_j^* = (1 - \sqrt{1 - x^*})/2$, $p_k^* = 0$ for $k \neq i, j$, and $q_i^* = (1 - \sqrt{1 - x^*})/2$, $q_j^* = (1 + \sqrt{1 - x^*})/2$, $q_k^* = 0$ for $k \neq i, j$. For this choice of p^*, q^*, we have $\langle u, p^* - q^*\rangle = \Upsilon\sqrt{1 - x^*}$ and $\mathrm{FC}(p^*, q^*) = x^*$. Thus, substituting p^*, q^* in the objective of Eq. (3.13) gives $(2\Upsilon\sqrt{1 - x^*} + \log(x^*))/2$, ensuring that p^*, q^* attains the maximum in Eq. (3.13).

It also follows that $\mathrm{BD}^*(u,-u) \geq 0$. Since $\mathrm{BD}^*(u,v)$ is a convex function of (u,v) and $u \mapsto (u,-u)$ is a linear function of u, $\mathrm{BD}^*(u,-u)$ is a convex function of u.

2. Follows from the fact that $\mathrm{BD}(p,q) = \mathrm{BD}(q,p)$ for all $p,q \in \Delta_M$.

3. Observe that we can write

$$\mathrm{BD}^*(u,v) = \sup_{p,q\in\Delta_M} \left(\left\langle \frac{(u+v)}{2}, p+q \right\rangle + \left\langle \frac{(u-v)}{2}, p-q \right\rangle + \log(\mathrm{BC}(p,q)) \right). \tag{3.17}$$

Since $\sup(f+g) \leq \sup f + \sup g$ for any real-valued functions f and g, and $\sup_{p,q\in\Delta_M} \langle (u+v), (p+q)/2 \rangle = (u+v)_{\max}$, we obtain

$$\mathrm{BD}^*(u,v) \leq (u+v)_{\max} + \mathrm{BD}^* \left(\frac{(u-v)}{2}, -\frac{(u-v)}{2} \right). \tag{3.18}$$

On the other hand, when $\Upsilon = ((u-v)_{\max} - (u-v)_{\min})/2 > 0$, the choice of p^*, q^* in the proof of Prop. 3.9.1 corresponding to the input $((u-v)/2, -(u-v)/2)$ gives the lower bound

$$\frac{(u+v)_i + (u+v)_j}{2} + \mathrm{BD}^* \left(\frac{(u-v)}{2}, -\frac{(u-v)}{2} \right) \leq \mathrm{BD}^*(u,v) \tag{3.19}$$

for all $i \in \mathrm{argmax}(u-v)$ and $j \in \mathrm{argmin}(u-v)$. When $\Upsilon = 0$, we can take $p_i^* = p_j^* = 1/2$ and $q_i^* = q_j^* = 1/2$ for any $i \in \mathrm{argmax}(u-v)$ and $j \in \mathrm{argmin}(u-v)$.

4. This follows from the definition of convex conjugate. It is called the Fenchel-Young inequality [8, Prop. 13.15] in the general scenario. □

The convex conjugate of Bhattacharyya distance will be used later in our study. We note that the above results can be generalized to obtain the convex conjugate of $-(1/2)\log F(\rho,\sigma)$. Here, the vectors u, v are replaced by observables $\mathcal{O}_1, \mathcal{O}_2$, and for an observable $\mathcal{O}$, we define $\Upsilon = \lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O})$ to be the difference between its maximum and minimum eigenvalues. This can be shown by following the proof above, and using ideas from the proof of Lem. 7.14 to replace the l_1-norm with the Schatten-1 norm and classical fidelity with quantum fidelity. We leave the

details as an exercise to the interested reader. We move on to studying other distance measures and their relation to the average Bhattacharyya distance.

3.2 Other distance measures

We study a few classical distance measures on states that are used in quantum information. We start be reviewing some definitions from classical statistics.

Definition 3.10 (Metrics on probability distributions). Let p and q be two probability distributions over M symbols.

(1) The **total variation distance (TVD)** between p and q is defined as

$$\|p-q\|_{\mathrm{TV}} = \sup_{A\subseteq[M]} \left| \sum_{i\in A} p_i - \sum_{i\in A} q_i \right| = \frac{1}{2}\|p-q\|_1 . \tag{3.20}$$

(2) The **Hellinger distance** between p and q is defined as

$$\mathrm{HD}(p,q) = \sqrt{1-\mathrm{BC}(p,q)} = \frac{1}{\sqrt{2}}\|\sqrt{p}-\sqrt{q}\|_2 . \tag{3.21}$$

(3) The **classical sine distance** between p and q is defined as

$$\mathrm{SDC}(p,q) = \sqrt{1-\mathrm{FC}(p,q)}. \tag{3.22}$$

□

All the above distance measures are metrics on the set of probability distributions on a fixed number of symbols. [39] proved that the sine distance, defined in Eq. (3.31), is a metric on quantum states. That the classical sine distance is a metric on probability distributions follows from this result.

Motivated by the discussion in the previous section, we define these distance measures on quantum states, as determined by a measurement protocol. Unlike the previous section, where

we only studied the average Bhattacharyya distance, we will look at both average and worst-case distance measures in this section.

Definition 3.11. Let ρ and σ be two quantum states and let $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ be a measurement protocol with total number of samples $N = \sum_{i=1}^{L} N_i$.

(1) The **average total variation distance** between ρ and σ determined by $\mathfrak{M}$ is defined as

$$\|\rho - \sigma\|_{\mathfrak{M},\text{avg}} = \sum_{i=1}^{L} \frac{N_i}{N} \left\| p_\rho^{(i)} - p_\sigma^{(i)} \right\|_{\text{TV}}, \tag{3.23}$$

while the **maximum total variation distance** is defined as

$$\|\rho - \sigma\|_{\mathfrak{M},\max} = \max_{i\in[L]} \left\| p_\rho^{(i)} - p_\sigma^{(i)} \right\|_{\text{TV}}, \tag{3.24}$$

(2) The **average Hellinger distance** between ρ and σ determied by $\mathfrak{M}$ is defined as

$$\text{HD}_{\mathfrak{M},\text{avg}}(\rho,\rho) = \sum_{i=1}^{L} \frac{N_i}{N} \sqrt{1 - \text{BC}(p_\rho^{(i)}, p_\sigma^{(i)})}, \tag{3.25}$$

while the **maximum Hellinger distance** is defined as

$$\text{HD}_{\mathfrak{M},\max}(\rho,\rho) = \max_{i\in[L]} \sqrt{1 - \text{BC}(p_\rho^{(i)}, p_\sigma^{(i)})}, \tag{3.26}$$

(3) The **average classical sine distance** between ρ and σ determined by $\mathfrak{M}$ is defined as

$$\text{SDC}_{\mathfrak{M},\text{avg}}(\rho,\sigma) = \sum_{i=1}^{L} \frac{N_i}{N} \sqrt{1 - \text{FC}(p_\rho^{(i)}, p_\sigma^{(i)})}, \tag{3.27}$$

while the **maximum classical sine distance** is defined as

$$\text{SDC}_{\mathfrak{M},\max}(\rho,\sigma) = \max_{i\in[L]} \sqrt{1 - \text{FC}(p_\rho^{(i)}, p_\sigma^{(i)})}. \tag{3.28}$$

□

The average and maximum total variation distance between quantum states has been studied in the quantum information literature (see, for example, [70, 71]). While the distance measures defined above are not metrics on the set of quantum states in general, they are pseudometrics. A **pseudometric** on $\mathscr{X}$ is a function $\mathscr{d}\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ that is non-negative, satisfies $\mathscr{d}(x,x) = 0$ for all $x \in \mathscr{X}$, is symmetric, and satisfies the triangle inequality. Observe that $\mathscr{d}$ is a metric if for all $x, y \in \mathscr{X}$, $\mathscr{d}(x,y) = 0$ implies $x = y$.

Proposition 3.12. *The functions* $\mathscr{d}_1(\rho,\sigma) = \|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}}$, $\mathscr{d}_2(\rho,\sigma) = \|\rho-\sigma\|_{\mathfrak{M},\max}$, $\mathscr{d}_3 = \mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)$, $\mathscr{d}_4(\rho,\sigma) = \mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma)$, $\mathscr{d}_5 = \mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)$, *and* $\mathscr{d}_6(\rho,\sigma) = \mathrm{SDC}_{\mathfrak{M},\max}(\rho,\sigma)$ *are pseudometrics on the set of quantum states. Furthermore, if* $\mathfrak{M}$ *is informationally complete, then* $\mathscr{d}_1,\ldots,\mathscr{d}_6$ *are metrics.*

Proof. It can be directly verified from respective definitions that $\mathscr{d}_1,\ldots,\mathscr{d}_6$ are non-negative, symmetric, and vanish when both input arguments are equal. It remains to prove the triangle inequality.

Since TVD, Hellinger distance and the classical sine distance are metrics on the set of probability distributions, they satisfy the triangle inequality. It immediately follows that the average distance measures $\mathscr{d}_1, \mathscr{d}_3, \mathscr{d}_5$ satisfy the triangle inequality. To see that the maximum distance measures $\mathscr{d}_2, \mathscr{d}_4, \mathscr{d}_6$ also satisfy the triangle inequality, we use the fact that $\max_i(u_i + v_i) \leq \max_i u_i + \max_i v_i$ for any real vectors u, v. Therefore, $\mathscr{d}_1,\ldots,\mathscr{d}_6$ are pseudometrics.

If $\mathfrak{M}$ is informationally complete, then for any $\rho \neq \sigma$, there is some $i \in [L]$ and $k \in [M_i]$ such that $p_\rho^{(i)}(k) \neq p_\sigma^{(i)}(k)$. Thus, $\left\|p_\rho^{(i)} - p_\sigma^{(i)}\right\|_{\mathrm{TV}} > 0$, $\mathrm{HD}(p_\rho^{(i)}, p_\sigma^{(i)}) > 0$, and $\mathrm{SDC}(p_\rho^{(i)}, p_\sigma^{(i)}) > 0$, since TVD, Hellinger distance, and the sine distance are metrics on the set of probability distributions on $[M_i]$. It follows that $\mathscr{d}_1,\ldots,\mathscr{d}_6 > 0$, showing that they are metrics on $\mathscr{X}$. □

Finally, we define the quantum counterparts of these distance measures.

Definition 3.13. Let ρ and σ be two d-dimensional quantum states.

(1) The **trace distance** between ρ and σ is defined as

$$\|\rho-\sigma\|_{\mathrm{tr}}=\frac{1}{2}\|\rho-\sigma\|_1 . \tag{3.29}$$

(2) The **Bures distance** between ρ and σ is defined as

$$\mathrm{D}_{\mathrm{Bur}}(\rho,\sigma)=\sqrt{2-2\sqrt{F}(\rho,\sigma)}. \tag{3.30}$$

(3) The **sine distance** between ρ and σ is defined as

$$\mathrm{SD}(\rho,\sigma)=\sqrt{1-F(\rho,\sigma)}. \tag{3.31}$$

□

These quantum distance measures are metrics on the set of quantum states [68]. We show that these quantum distance measures can be obtained by optimizing the corresponding classical distance measures over all measurement protocols.

Proposition 3.14. *For any two states $\rho,\sigma\in\mathscr{X}$, we have*

$$\|\rho-\sigma\|_{\mathrm{tr}}=\max_{\mathfrak{M}}\|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}}=\max_{\mathfrak{M}}\|\rho-\sigma\|_{\mathfrak{M},\max}, \tag{3.32}$$

$$\mathrm{D}_{\mathrm{Bur}}(\rho,\sigma)=\sqrt{2}\max_{\mathfrak{M}}\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)=\sqrt{2}\max_{\mathfrak{M}}\mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma), \tag{3.33}$$

and

$$\mathrm{SD}(\rho,\sigma)=\max_{\mathfrak{M}}\mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)=\max_{\mathfrak{M}}\mathrm{SDC}_{\mathfrak{M},\max}(\rho,\sigma). \tag{3.34}$$

Proof. From [107, Lem. 9.1.1], [50, 36], we know that $\|\rho-\sigma\|_{\mathrm{tr}}=\max_{\boldsymbol{E}}\|p_{\boldsymbol{E},\rho}-p_{\boldsymbol{E},\sigma}\|_{\mathrm{TV}}$, where the maximization is over all POVMs. In particular, $\|p_{\boldsymbol{E},\rho}-p_{\boldsymbol{E},\sigma}\|_{\mathrm{TV}}\leq\|\rho-\sigma\|_{\mathrm{tr}}$ for every POVM $\boldsymbol{E}$. Then, from the definition of average and maximum TVD, we have that $\|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}}\leq\|\rho-\sigma\|_{\mathrm{tr}}$

and $\|\rho-\sigma\|_{\mathfrak{M},\max} \leq \|\rho-\sigma\|_{\mathrm{tr}}$. Since there is a two-outcome POVM $\boldsymbol{E}_*$ such that $\|\rho-\sigma\|_{\mathrm{tr}} = \|p_{\boldsymbol{E}_*,\rho} - p_{\boldsymbol{E}_*,\sigma}\|_{\mathrm{TV}}$ [107, Lem. 9.1.1], [50], the measurement protocol $\mathfrak{M}_* = \{(\boldsymbol{E}_*, 1)\}$ achieves the maximum in Eq. (3.32).

From Eq. (3.8), we know that for all $i \in [L]$, $\mathrm{BC}(p_\rho^{(i)}, p_\sigma^{(i)}) \geq \sqrt{F}(\rho,\sigma)$. Therefore, $\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma) \leq \mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma) \leq \sqrt{1-\sqrt{F}(\rho,\sigma)} = \mathrm{D}_{\mathrm{Bur}}(\rho,\sigma)/\sqrt{2}$. Since there is a measurement protocol $\mathfrak{M}_*$ for which $\mathrm{BC}_{\mathfrak{M}_*}(\rho,\sigma) = \sqrt{F}(\rho,\sigma)$ (Prop. 3.7), Eq. (3.33) holds. The same arguments show that Eq. (3.34) also holds. □

3.3 Relation of Bhattacharyya distance with other distance measures

In this section, we derive some inequalities between the geometric-average Bhattacharyya coefficient and the other classical distance measures introduced in the previous section. Most of these inequalities are straightforward generalizations of well-known inequalities in the literature. We first note down the well-known Fuchs-van de Graaf inequality.

Proposition 3.15 (Fuchs-van de Graaf inequality [36]). *For all quantum states ρ, σ, we have*

$$1-\sqrt{F}(\rho,\sigma) \leq \|\rho-\sigma\|_{\mathrm{tr}} \leq \sqrt{1-F(\rho,\sigma)}. \tag{3.35}$$

Specializing to classical distributions, for all $p, q \in \Delta_M$, we have

$$1-\mathrm{BC}(p,q) \leq \|p-q\|_{\mathrm{TV}} \leq \sqrt{1-\mathrm{FC}(p,q)}. \tag{3.36}$$

Now, we generalize the Fuchs-van de Graaf inequality to the classical distance measures defined in the previous sections. Since these classical distance measures are just distance measures on the probability distributions associated with a measurement protocol, we only need the classical version of Fuchs-van de Graaf inequality noted in Eq. (3.36) for proving the proposition below.

Proposition 3.16. *Let ρ and σ be any quantum states, and let $\mathfrak{M}$ be any measurement protocol.*

Then, we have

$$\begin{aligned}(\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma))^2 &\le 1-\mathrm{BC}_{\mathfrak{M}}(\rho,\sigma) \le (\mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma))^2 \\ &\le \|\rho-\sigma\|_{\mathfrak{M},\max} \le \mathrm{SDC}_{\mathfrak{M},\max}(\rho,\sigma) \le \sqrt{2}\ \mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma)\end{aligned} \tag{3.37}$$

and

$$\begin{aligned}(\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma))^2 &\le \|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}} \le \mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma) \\ &\le \sqrt{1-\mathrm{FC}_{\mathfrak{M}}(\rho,\sigma)} \le \sqrt{2}\sqrt{1-\mathrm{BC}_{\mathfrak{M}}(\rho,\sigma)}.\end{aligned} \tag{3.38}$$

Proof. Let $\mathfrak{M}=\{(\boldsymbol{E}^{(i)},N_i)\}_{i=1}^L$ be the given measurement protocol. For $i\in[L]$, denote $\lambda_i=N_i/N$, where $N=\sum_{i=1}^L N_i$ the total number of samples. We have $\lambda_1,\dots,\lambda_L\ge 0$ and $\sum_{i=1}^L\lambda_i=1$. Define $x_i=\mathrm{FC}(p_\rho^{(i)},p_\sigma^{(i)})$ and $y_i=\|p_\rho^{(i)}-p_\sigma^{(i)}\|_{\mathrm{TV}}$ for $i\in[L]$. Then, by Fuchs-van de Graaf inequality (Eq. (3.36)), $1-\sqrt{x_i}\le y_i\le\sqrt{1-x_i}$ for all $i\in[L]$. Observe that

$$\begin{aligned}
\mathrm{BC}_{\mathfrak{M}}(\rho,\sigma)&=\prod_{i=1}^L\sqrt{x_i}^{\lambda_i} & \mathrm{FC}_{\mathfrak{M}}(\rho,\sigma)&=\prod_{i=1}^L x_i^{\lambda_i}\\
\|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}}&=\sum_{i=1}^L\lambda_i y_i & \|\rho-\sigma\|_{\mathfrak{M},\max}&=\max_{i\in[L]} y_i\\
\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)&=\sum_{i=1}^L\lambda_i\sqrt{1-\sqrt{x_i}} & \mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma)&=\max_{i\in[L]}\sqrt{1-\sqrt{x_i}}\\
\mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)&=\sum_{i=1}^L\lambda_i\sqrt{1-x_i} & \mathrm{SDC}_{\mathfrak{M},\max}(\rho,\sigma)&=\max_{i\in[L]}\sqrt{1-x_i}.
\end{aligned} \tag{3.39}$$

We first prove each inequality in the chain of inequalities of Eq. (3.37). The proofs are in the following list, where each item is titled by the inequality to be proven.

(1) $(\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma))^2\le 1-\mathrm{BC}_{\mathfrak{M}}(\rho,\sigma)$:

$$\sum_i\lambda_i\sqrt{1-\sqrt{x_i}}\le\sqrt{\sum_i\lambda_i(1-\sqrt{x_i})}=\sqrt{1-\sum_i\lambda_i\sqrt{x_i}}\le\sqrt{1-\prod_i\sqrt{x_i}^{\lambda_i}}, \tag{3.40}$$

where we obtained the first inequality by concavity of the square-root function, and used the AM-GM inequality $\prod_{i=1}^L\sqrt{x_i}^{\lambda_i}\le\sum_{i=1}^L\lambda_i\sqrt{x_i}$ for the last inequality.

(2) $1-\mathrm{BC}_{\mathfrak{M}}(\rho,\sigma)\le(\mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma))^2$: Note that $\prod_i x_i^{\lambda_i}\ge\min_i x_i$. Then, since $x_i\ge 0$ for all

i, we have

$$\sqrt{1-\prod_i \sqrt{x_i}^{\lambda_i}} \leq \sqrt{1-\sqrt{\min_i x_i}} = \sqrt{1-\min_i \sqrt{x_i}} = \sqrt{\max_i (1-\sqrt{x_i})} = \max_i \sqrt{1-\sqrt{x_i}}. \tag{3.41}$$

(3) $(\mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma))^2 \leq \|\rho-\sigma\|_{\mathfrak{M},\max}$:

$$\left(\max_i \sqrt{1-\sqrt{x_i}}\right)^2 = \max_i \left(\sqrt{1-\sqrt{x_i}}\right)^2 \leq \max_i y_i, \tag{3.42}$$

where the last inequality follows from Fuchs-van de Graaf inequality.

(4) $\|\rho-\sigma\|_{\mathfrak{M},\max} \leq \mathrm{SDC}_{\mathfrak{M},\max}(\rho,\sigma)$:

$$\max_i y_i \leq \max_i \sqrt{1-x_i} \tag{3.43}$$

by Fuchs-van de Graaf inequality.

(5) $\mathrm{SDC}_{\mathfrak{M},\max}(\rho,\sigma) \leq \sqrt{2}\ \mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma)$ (as well as $\mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma) \leq \sqrt{2}\ \mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)$): For all $i \in [L]$, we have

$$\sqrt{1-x_i} = \sqrt{1+\sqrt{x_i}}\sqrt{1-\sqrt{x_i}} \leq \sqrt{2}\sqrt{1-\sqrt{x_i}}, \tag{3.44}$$

where we used the fact that $x_i \in [0,1]$. Taking the maximum (or average) gives the desired inequality.

We first prove each inequality in the chain of inequalities of Eq. (3.38). As before, the proofs are in the following list, where each item is titled by the inequality to be proven.

(1) $(\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma))^2 \leq \|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}}$: By convexity of the square function $a \mapsto a^2$ and Fuchs-van

de Graaf inequality, we have

$$\left(\sum_i \lambda_i \sqrt{1-\sqrt{x_i}}\right)^2 \leq \sum_i \lambda_i \left(\sqrt{1-\sqrt{x_i}}\right)^2 \leq \sum_i \lambda_i y_i. \tag{3.45}$$

(2) $\|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}} \leq \mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)$: By Fuchs-van de Graaf inequality, we have

$$\sum_i \lambda_i y_i \leq \sum_i \lambda_i \sqrt{1-x_i}. \tag{3.46}$$

(3) $\mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma) \leq \sqrt{1-\mathrm{FC}_{\mathfrak{M}}(\rho,\sigma)}$:

$$\sum_i \lambda_i \sqrt{1-x_i} \leq \sqrt{\sum_i \lambda_i (1-x_i)} = \sqrt{1-\sum_i \lambda_i x_i} \leq \sqrt{1-\prod_i x_i^{\lambda_i}}, \tag{3.47}$$

where we used AM-GM inequality to obtain the last inequality.

(4) $\sqrt{1-\mathrm{FC}_{\mathfrak{M}}(\rho,\sigma)} \leq \sqrt{2}\sqrt{1-\mathrm{BC}_{\mathfrak{M}}(\rho,\sigma)}$: Writing $x = \mathrm{FC}_{\mathfrak{M}}(\rho,\sigma)$ and noting that $x \in [0,1]$, we have $\sqrt{1-x} = \sqrt{1+\sqrt{x}}\sqrt{1-\sqrt{x}} \leq \sqrt{2}\sqrt{1-\sqrt{x}}$. □

We can use the definition $\mathrm{BD}_{\mathfrak{M}}(\rho,\sigma) = -\log \mathrm{BC}_{\mathfrak{M}}(\rho,\sigma)$ to derive inequalities for the average Bhattacharyya distance from the above relations.

We end this chapter by giving a continuity bound for quantum fidelity. It is known that the fidelity is a continuous function. A continuity bound quantifies how close the fidelity between two pairs of states (ρ,σ) and (ρ',σ') must be in terms of a distance between these states.

Proposition 3.17 (Continuity bound for fidelity). *For all quantum states $\rho,\sigma,\rho',\sigma'$, we have*

$$\begin{aligned} \left|F(\rho,\sigma) - F(\rho',\sigma')\right| &\leq 2\left(\mathrm{SD}(\rho,\rho') + \mathrm{SD}(\sigma,\sigma')\right) \\ &\leq 2\left(\mathrm{D}_{\mathrm{Bur}}(\rho,\rho') + \mathrm{D}_{\mathrm{Bur}}(\sigma,\sigma')\right) \\ &\leq 4\sqrt{\|\rho-\rho'\|_{\mathrm{tr}} + \|\sigma-\sigma'\|_{\mathrm{tr}}}. \end{aligned} \tag{3.48}$$

Proof. To obtain Eq. (3.48), we prove the following chain of inequalities.

(1) $|F(\rho,\sigma) - F(\rho',\sigma')| \leq 2(\mathrm{SD}(\rho,\rho') + \mathrm{SD}(\sigma,\sigma'))$: Note that $|F(\rho,\sigma) - F(\rho',\sigma')| = |(1 - F(\rho,\sigma)) - (1 - F(\rho',\sigma'))|$. Also note that for all $0 \leq x, y \leq 1$, we have

$$|x - y| \leq |\sqrt{x} - \sqrt{y}||\sqrt{x} + \sqrt{y}| \leq 2|\sqrt{x} - \sqrt{y}|. \tag{3.49}$$

Then, taking $x = 1 - F(\rho,\sigma)$ and $y = 1 - F(\rho',\sigma')$ in this equation, we obtain

$$|F(\rho,\sigma) - F(\rho',\sigma')| \leq 2|\mathrm{SD}(\rho,\sigma) - \mathrm{SD}(\rho',\sigma')|. \tag{3.50}$$

Since the sine distance is a metric, we can use the triangle inequality and the reverse triangle inequality to obtain

$$\begin{aligned} \big|\mathrm{SD}(\rho,\sigma) - \mathrm{SD}(\rho',\sigma')\big| &\leq \big|\mathrm{SD}(\rho,\sigma) - \mathrm{SD}(\rho',\sigma)\big| + \big|\mathrm{SD}(\rho',\sigma) - \mathrm{SD}(\rho',\sigma')\big| \\ &\leq \mathrm{SD}(\rho,\rho') + \mathrm{SD}(\sigma,\sigma'). \end{aligned} \tag{3.51}$$

(2) $\mathrm{SD}(\rho,\rho') \leq \mathrm{D}_{\mathrm{Bur}}(\rho,\rho')$ for all states ρ,ρ': Writing $x = F(\rho,\rho')$, we have $\mathrm{SD}(\rho,\rho') = \sqrt{1-x} = \sqrt{1+\sqrt{x}}\sqrt{1-\sqrt{x}} \leq \sqrt{2}\sqrt{1-\sqrt{x}} = \mathrm{D}_{\mathrm{Bur}}(\rho,\rho')$, where we used the fact that $x \in [0,1]$. We similarly obtain $\mathrm{SD}(\sigma,\sigma') \leq \mathrm{D}_{\mathrm{Bur}}(\sigma,\sigma')$, from which $\mathrm{SD}(\rho,\rho') + \mathrm{SD}(\sigma,\sigma') \leq \mathrm{D}_{\mathrm{Bur}}(\rho,\rho') + \mathrm{D}_{\mathrm{Bur}}(\sigma,\sigma')$ follows.

(3) $\mathrm{D}_{\mathrm{Bur}}(\rho,\rho') + \mathrm{D}_{\mathrm{Bur}}(\sigma,\sigma') \leq 2(\sqrt{\|\rho-\rho'\|_{\mathrm{tr}} + \|\sigma-\sigma'\|_{\mathrm{tr}}})$: First, we obtain

$$\mathrm{D}_{\mathrm{Bur}}(\rho,\rho') + \mathrm{D}_{\mathrm{Bur}}(\sigma,\sigma') \leq \sqrt{2}\left(\sqrt{\|\rho-\rho'\|_{\mathrm{tr}}} + \sqrt{\|\sigma-\sigma'\|_{\mathrm{tr}}}\right) \tag{3.52}$$

using Fuchs-van de Graaf inequality (Eq. (3.35)). Then, the desired inequality follows from concavity of the square-root. □

Note that similar ideas can also be used to give continuity bounds for square-root fidelity. These bounds can also be specialized to classical probability distributions to get continuity bounds for the Bhattacharyya coefficient. Note, however, that the Bhattacharyya distance is not continuous

on the set of classical probability distributions on a fixed alphabet.

Chapter 4

Statistical problem

We begin by describing the mathematical problem studied by Juditsky & Nemirovski [62] in Sec. 4.1. We then discuss the results of [62], including their estimation procedure and theoretical guarantees in Sec. 4.2. In Sec. 4.3, we discuss some drawbacks with the estimation procedure of [62], and subsequently, propose a simplified estimation procedure. We show that our estimation procedure satisfies all the theoretical guarantees of [62], and derive some additional results.

4.1 Mathematical formulation

Suppose that we have a set of "states" $\mathscr{X} \subseteq \mathbb{R}^D$, which is assumed to be a compact and convex set. We imagine that there is some state $x_{\text{true}} \in \mathscr{X}$ that is the "true state" of the system, but is unknown to us. We are given some vector $g \in \mathbb{R}^D$, and our goal is to estimate the linear form $\langle g, x_{\text{true}} \rangle = g^T x_{\text{true}}$. For intuition, one imagine the state x_{true} to be the quantum state ρ and the vector g to be the observable $\mathcal{O}$.

Now, the question arises as to what data we have available for estimating this linear form. We suppose that we have access to a *single* outcome of a random variable determined by x_{true}, chosen from a family of random variables described below. This random variable can be defined a over a joint space that contains all the data from the experiment, and therefore, a single random variable is sufficient to develop the general theory. The details of how data from many random variables can be incorporated into a single random variable is discussed at the end of this section.

Consider a family of random variables Z_μ, parameterized by $\mu \in \mathscr{M}$ for some subset $\mathscr{M} \subseteq \mathbb{R}^M$.

Each Z_μ take values in a Polish space $(\Omega, \mathscr{B}(\Omega))$, equipped with a σ-finite measure m that is not identically zero. Z_μ is assumed to have the probability density p_μ with respect to the reference measure m. Mathematically, p_μ is a non-negative, $\mathscr{B}(\Omega)$-measurable function satisfying $\int_\Omega p_\mu dm = 1$, so that

$$\mathbb{P}_\mu(B) = \int_B p_\mu dm \tag{4.1}$$

for $B \in \mathscr{B}(\Omega)$ defines a probability distribution on $(\Omega, \mathscr{B}(\Omega))$. [62] call the mapping $\mathscr{D}(\mu) = p_\mu$, from parameter $\mu \in \mathscr{M}$ to the density function p_μ, a **parametric density family**. The state $x_{\text{true}} \in \mathscr{X}$ determines the random variable $\mathrm{Z}_{\mathrm{A}(x_{\text{true}})}$ through an affine function $\mathrm{A}\colon \mathbb{R}^d \to \mathbb{R}^M$ satisfying $\mathrm{A}(\mathscr{X}) \subseteq \mathscr{M}$, and we are given one outcome of this random variable for the purpose of estimation. We will denote the affine map as $\mathrm{A}\colon \mathscr{X} \to \mathscr{M}$ to avoid writing $\mathscr{A}(\mathscr{X}) \subseteq \mathscr{M}$ repeatedly. The reason we use $\mathrm{A}(x_{\text{true}})$ instead of x_{true} as the parameter is to model situations where we don't have or need the full knowledge of x_{true}. For example, if we want to estimate the expectation value of an observable, it suffices to perform measurements that are informative enough to learn the observable but not perform full quantum tomography.

Our goal is to construct an estimator that uses an outcome of $\mathrm{Z}_{\mathrm{A}(x_{\text{true}})}$ to estimate $\langle g, x_{\text{true}}\rangle$. We define an estimator to be any real-valued Borel measurable function on $(\Omega, \mathscr{B}(\Omega))$. In practice, working with arbitrary measurable functions is challenging, from a theoretical as well as computational point of view. For this reason, [62] restrict their attention choosing an estimator from a set $\mathscr{F}$ that satisfies two properties: (1) it is a finite-dimensional vector space of Borel measurable functions on $(\Omega, \mathscr{B}(\Omega))$, and (2) it contains all the constant functions. Any estimator from the set $\mathscr{F}$ is called an **affine estimator**. We note at this point that the functions in $\mathscr{F}$ need not be affine functions, as Ω might not even have a linear structure. Nevertheless, this terminology is motivated by the later observation that for many problems of interest, the estimators in $\mathscr{F}$ turn out to be affine functions. In fact, we will see in the quantum case (Sec. 5.3.2), where Ω generally does not have a linear structure, that it is still possible to express our estimator as an affine function.

To be able to choose an appropriate estimator in $\mathscr{F}$ given outcomes from $p_{\mathrm{A}(x_{\text{true}})}$, we need to

make sure that the set of affine estimators $\mathscr{F}$ "interacts well" with the parametric density family $\mathscr{D}$. [62] formalize this idea by defining a *good pair* of parametric density family and affine estimators.

Definition 4.1 (Good pair). We call a pair $(\mathscr{D}, \mathscr{F})$ of parametric density family $\mathscr{D}$ and finite-dimensional space $\mathscr{F}$ of Borel functions on Ω a **good pair** if the following conditions hold.

(1) $\mathscr{M}$ is a relatively open convex set in $\mathbb{R}^m$.

(2) Whenever $\mu \in \mathscr{M}$, we have $p_\mu(\omega) > 0$ for all $\omega \in \Omega$.

(3) Whenever $\mu, \nu \in \mathscr{M}$, $\widehat{g}(\omega) = \log(p_\mu(\omega)/p_\nu(\omega)) \in \mathscr{F}$.

(4) Whenever $\widehat{g} \in \mathscr{F}$, the function

$$F_{\widehat{g}}(\mu) = \log\left(\int_\Omega \exp\left(\widehat{g}(\omega)\right) p_\mu(\omega) d\boldsymbol{m}\right) \tag{4.2}$$

is well-defined and concave in $\mu \in \mathscr{M}$. □

Note that the second condition that $p_\mu > 0$ is essential, for otherwise $\log(p_\mu)$ is ill-defined. We will discuss the implications of this assumption later, and also show that it does not restrict the power of our results in the quantum case.

We are now in a position to introduce the main objective of this section, which is to find an estimator that minimizes the estimation error. For this purpose, we need to formalize what we mean by estimation error of an estimator, since the error generally depends on the method used in the statistical analysis (for example, a specific concentration inequality). To circumvent such ambiguities, we focus on the smallest possible error of the estimator that one can achieve using any statistical method.

Definition 4.2 (δ-risk of an estimator). Given a confidence level $1 - \delta \in (0, 1)$, the **δ-risk** of an estimator $\widehat{g}$ is defined as

$$\mathscr{R}(\widehat{g}, \delta) = \inf\left\{\varepsilon \mid \inf_{x \in \mathscr{X}} \mathbb{P}_{\mathrm{A}(x)}\left(|\widehat{g} - \langle g, x\rangle| \leq \varepsilon\right) > 1 - \delta\right\}. \tag{4.3}$$

We refer to the δ-risk as risk when the confidence level is clear from context. □

Since the error $\mathscr{R}(\widehat{g}, \delta)$ does not depend on the state or the data, it is minimax in the sense we defined in the preliminaries. Since we want the smallest possible error that *any* estimator can achieve, we can minimize the δ-risk over all estimators, which leads us to the following definition.

Definition 4.3 (Minimax optimal risk). Given a confidence level $1 - \delta \in (0, 1)$, the **minimax optimal risk** is defined as

$$\mathscr{R}_*(\delta) = \inf_{\widehat{g}} \mathscr{R}(\widehat{g}, \delta), \tag{4.4}$$

where the infimum is over all measurable functions $\widehat{g}$ on $(\Omega, \mathscr{B}(\Omega))$. □

The term "minimax" alludes to the fact that we are looking for the best performance ("minimum over all estimators") in the worst case scenario ("no matter the state x").

Now, we study the situation where we estimate $\langle g, x_{\text{true}} \rangle$ using the outcomes of L independent random variables $\mathrm{Z}^{(1)}_{\mathrm{A}^{(1)}(x_{\text{true}})}, \ldots, \mathrm{Z}^{(L)}_{\mathrm{A}^{(L)}(x_{\text{true}})}$. As before, we suppose that for $i \in [L]$, we have a Polish space $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$, equipped with a σ-finite measure $\mathscr{m}^{(i)}$. For each $i \in [L]$, we also have a set of parameters $\mathscr{M}^{(i)}$, and we have a family of random variables $\{Z^{(i)}_{\mu_i} \mid \mu_i \in \mathscr{M}^{(i)}\}$ that takes values in $\Omega^{(i)}$. The random variable $\mathrm{Z}^{(i)}_{\mu_i}$ has probability density $p^{(i)}_{\mu_i}$ with respect to the reference measure $\mathscr{m}^{(i)}$. We call the mapping $\mathscr{D}^{(i)}(\mu_i) = p^{(i)}_{\mu_i}$ as the ith parametric density family. For each $i \in [L]$, we are given affine mappings $\mathrm{A}^{(i)} \colon \mathscr{X} \to \mathscr{M}^{(i)}$ that map the state x_{true} to the corresponding parameter in $\mathscr{M}^{(i)}$.

We suppose that we get one outcome each from the random variables $\mathrm{Z}^{(1)}_{\mathrm{A}^{(1)}(x_{\text{true}})}, \ldots, \mathrm{Z}^{(L)}_{\mathrm{A}^{(L)}(x_{\text{true}})}$ for estimation. We need an estimator to process the outcome of the random variable $\mathrm{Z}^{(i)}_{\mathrm{A}^{(i)}(x_{\text{true}})}$. For this purpose, we suppose that we have a set of estimators $\mathscr{F}^{(i)}$, which is a finite-dimensional linear vector space of measurable functions on $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$ that contains constant functions. As before, any function in $\mathscr{F}^{(i)}$ is called an affine estimator, and we use it to process the outcome obtained from $\mathrm{Z}^{(i)}_{\mathrm{A}^{(i)}(x_{\text{true}})}$. For this to work well, we assume that $(\mathscr{D}^{(i)}, \mathscr{F}^{(i)})$ is a good pair (see Def. 4.1).

At this point, we have L good pairs of parametric density families and set of affine estimators. For the theory developed for a single random variable to hold for many random variables, we need a

way to combine these L good pairs into one "large" good pair. This would enable us, in particular, to combine the L affine estimators into a single affine estimator for $\langle g, x_{\text{true}} \rangle$. [62] show how such a construction can be done, which leads us to our next definition.

Definition 4.4 (Direct product of good pairs). Considering the following quantities for $i \in [L]$. Let $(\Omega^{(i)}, \Sigma^{(i)})$ be a Polish space endowed with a Borel σ-finite measure $m^{(i)}$. Let $\mathscr{D}^{(i)}(\mu_i) = p^{(i)}_{\mu_i}$ be the parametric density family for $\mu_i \in \mathscr{M}^{(i)}$. Let $\mathscr{F}^{(i)}$ be a finite-dimensional linear space of Borel functions on $\Omega^{(i)}$ containing constants, such that the pair $(\mathscr{D}^{(i)}, \mathscr{F}^{(i)})$ is good. Then, the direct product of these good pairs, $(\mathscr{D}, \mathscr{F}) = \bigotimes_{i=1}^{L} (\mathscr{D}^{(i)}, \mathscr{F}^{(i)})$, is defined as follows.

(1) The large space is the Cartesian product $\Omega = \Omega^{(1)} \times \cdots \times \Omega^{(L)}$, endowed with the product Borel σ-algebra $\mathscr{B}(\Omega) = \mathscr{B}(\Omega^{(1)}) \otimes \cdots \otimes \mathscr{B}(\Omega^{(L)})$ and the product measure $m = m^{(1)} \times \cdots \times m^{(L)}$.

(2) The set of parameters is $\mathscr{M} = \mathscr{M}^{(1)} \times \cdots \times \mathscr{M}^{(L)}$, and the associated parametric density family is $\mathscr{D}(\mu) = p_\mu \equiv \prod_{i=1}^{L} p^{(i)}_{\mu_i}$ for $\mu = (\mu_1, \ldots, \mu_L) \in \mathscr{M}$.

(3) The linear space $\mathscr{F}$ comprises of all functions $\widehat{g}$ defined as $\widehat{g}(\omega_1, \omega_2, \ldots, \omega_L) = \sum_{i=1}^{L} \widehat{g}^{(i)}(\omega_i)$, where $\widehat{g}^{(i)} \in \mathscr{F}^{(i)}$ and $\omega_i \in \Omega^{(i)}$ for $i \in [L]$. □

In the above definition, we used the fact that the Borel σ-algebra on Ω is the product of Borel σ-algebras on $\Omega^{(i)}$, since each $\Omega^{(i)}$ is a Polish space [64, Lem. 1.2]. The first and the second conditions are chosen so that the random variables $\mathrm{Z}^{(1)}_{\mathrm{A}^{(1)}(x_{\text{true}})}, \ldots, \mathrm{Z}^{(L)}_{\mathrm{A}^{(i)}(x_{\text{true}})}$ are independent. We choose the affine mapping $\mathrm{A} \colon \mathscr{X} \to \mathscr{M}$ to be the direct sum $\mathrm{A} = \oplus_{i=1}^{L} \mathrm{A}^{(i)}$ of $\mathrm{A}^{(1)}, \ldots, \mathrm{A}^{(L)}$. The third condition is chosen such that the pair $(\mathscr{D}, \mathscr{F})$ satisfies the conditions of Def. 4.1, and is therefore itself a good pair. Thus, we have a good pair $(\Omega, \mathscr{F})$, and consequently, the theory developed for a single random variable also applies to this case. For this reason, we will focus on explaining constructing estimation procedures for the case of a single random variable.

Before we present the estimation procedure of [62], we present a general definition of the Bhattacharyya coefficient. The Bhattacharyya coefficient will be particularly important in the estimation procedure we develop in Sec. 4.3.

Definition 4.5 (Bhattacharyya coefficient). Given probability densities p_μ, p_ν with respect to the reference measure $\mathscr{m}$ on $(\Omega, \mathscr{B}(\Omega))$, the **Bhattacharyya coefficient** between the probability distributions $\mathbb{P}_\mu$ and $\mathbb{P}_\nu$ is defined as

$$\mathrm{BC}(\mu, \nu) = \int_\Omega \sqrt{p_\mu p_\nu}\, d\mathscr{m}. \tag{4.5}$$

The **Bhattacharyya distance** between $\mathbb{P}_\mu$ and $\mathbb{P}_\nu$ is defined as

$$\mathrm{BD}(\mu, \nu) = -\log(\mathrm{BC}(\mu, \nu)). \tag{4.6}$$

□

Since the Bhattacharyya coefficient and the Bhattacharyya distance are defined between probability distributions, we should technically write $\mathrm{BC}(\mathbb{P}_\mu, \mathbb{P}_\nu)$ and $\mathrm{BD}(\mathbb{P}_\mu, \mathbb{P}_\nu)$. We shorten this to $\mathrm{BC}(\mu, \nu)$ and $\mathrm{BD}(\mu, \nu)$ in Eq. (4.5) and Eq. (4.6) to avoid cumbersome notation in Sec. 4.3.

Observe that if Ω is a finite set with discrete σ-algebra, $\mathscr{m}$ is the counting measure, and $\mathscr{M}$ is the standard simplex, then $p_\mu = \mu$ is a discrete probability distribution for $\mu \in \mathscr{M}$, and we have $\mathrm{BC}(\mu, \nu) = \sum_i \sqrt{\mu_i \nu_i}$. This coincides with the definition for Bhattacharyya coefficient for discrete distributions that we saw in Def. 3.1. Moreover, we have the multiplicative (additive) property for Bhattacharyya coefficient (distance) for product distributions, as in the discrete case.

Lemma 4.6. *Suppose that for $i \in [L]$, we have N_i independent copies of a random variable Z_{μ_i} taking values in a Polish space $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$, having density $p^{(i)}_{\mu_i}$ with respect to a σ-finite reference measure $\mathscr{m}^{(i)}$, and $\mu_i \in \mathscr{M}^{(i)}$. Then, for $\Omega = \prod_{i=1}^L (\Omega^{(i)})^{N_i}$, $\mathscr{m} = \prod_{i=1}^L \mathscr{m}^{(i)}$, and $\mu, \nu \in \mathscr{M} = \prod_{i=1}^L (\mathscr{M}^{(i)})^{N_i}$, the Bhattacharyya coefficient between the distributions $\mathbb{P}_\mu$ and $\mathbb{P}_\nu$ on $(\Omega, \mathscr{B}(\Omega))$ with densities $p_\mu = \prod_{i=1}^L p_{\mu_i}$ and $p_\nu = \prod_{i=1}^L p_{\nu_i}$ with respect to $\mathscr{m}$ satisfies*

$$\mathrm{BC}(\mu, \nu) = \prod_{i=1}^{L} \left(\mathrm{BC}(\mu_i, \nu_i)\right)^{N_i}, \tag{4.7}$$

and the Bhattacharyya distance satisfies

$$\mathrm{BD}(\mu,\nu)=\sum_{i=1}^{L}N_i\mathrm{BD}(\mu_i,\nu_i). \tag{4.8}$$

Proof. Since $(\Omega^{(i)},\mathscr{B}(\Omega^{(i)}))$ is a Polish space for each $i\in[L]$, $\mathscr{B}(\Omega)=\otimes_{i=1}^{L}(\mathscr{B}(\Omega^{(i)}))^{\otimes N_i}$ [64, Lem. 1.2]. Then, by Fubini-Tonelli's theorem [64, Thm. 1.27], we have

$$\mathrm{BC}(\mu,\nu)=\int_{\Omega}p_{\mu}dm=\prod_{i=1}^{L}\left(\int_{\Omega^{(i)}}p_{\mu_i}^{(i)}dm^{(i)}\right)^{N_i}=\prod_{i=1}^{L}\left(\mathrm{BC}(\mu_i,\nu_i)\right)^{N_i}. \tag{4.9}$$

Eq. (4.8) follows from the definition of Bhattacharyya distance and Eq. (4.7). □

4.2 Juditsky and Nemirovski's estimation procedure

The main result of Juditsky & Nemirovski [62] is a procedure to construct an affine estimator whose error is within a small factor of the minimax optimal risk. As a result, their estimation procedure cannot be improved upon by any method by more than a small constant factor under the mathematical setting described in Sec. 4.1.

[62] propose the following procedure to construct an estimator for $\langle g,x_{\mathrm{true}}\rangle$ using an outcome of a single random variable.

Box 1: Juditsky & Nemirovski's estimation procedure [62]

(1) For $r\geq 0$, define the function $\Phi_r\colon(\mathscr{X}\times\mathscr{X})\times(\mathscr{F}\times(0,\infty))\to\mathbb{R}$ as

$$\Phi_r(x,y;\ \phi,\alpha)=\langle g,x\rangle-\langle g,y\rangle\ +\alpha\left[\log\left(\int_{\Omega}\exp(-\phi/\alpha)p_{A(x)}dm\right)\right.$$
$$\left.+\log\left(\int_{\Omega}\exp(\phi/\alpha)p_{A(y)}dm\right)\right]+2\alpha r. \tag{4.10}$$

(2) Denote the saddle-point value of Φ_r by $2\Phi_*(r)$:

$$\Phi_*(r) = \frac{1}{2} \sup_{x,y\in\mathscr{X}} \inf_{\phi\in\mathscr{F},\alpha>0} \Phi_r(x,y;\phi,\alpha) = \frac{1}{2} \inf_{\phi\in\mathscr{F},\alpha>0} \max_{x,y\in\mathscr{X}} \Phi_r(x,y;\phi,\alpha). \tag{4.11}$$

(3) Given a confidence level $1-\delta \in (0.75, 1)$ and a positive number $\epsilon_{\text{inf}} > 0$, find $\phi_* \in \mathscr{F}$ and $\alpha_* > 0$ such that

$$\max_{x,y\in\mathscr{X}} \Phi_{\log(2/\delta)}(x,y;\phi_*,\alpha_*) \leq 2\Phi_*(\log(2/\delta)) + \epsilon_{\text{inf}}. \tag{4.12}$$

This is achieved by minimizing the convex function

$$\overline{\Phi}_{\log(2/\delta)}(\phi,\alpha) = \max_{x,y\in\mathscr{X}} \Phi_{\log(2/\delta)}(x,y;\phi,\alpha). \tag{4.13}$$

(4) The estimator $\widehat{g}_* \in \mathscr{F}$ is then defined as

$$\widehat{g}_* = \phi_* + c \tag{4.14}$$

where the constant c is obtained by solving the optimization problem

$$\begin{aligned} c = &\frac{1}{2} \max_{x\in\mathscr{X}} \left[\langle g, x\rangle + \alpha_* \log\left(\int_\Omega \exp\left(-\phi_*/\alpha_*\right) p_{A(x)} d\mathscr{m} \right) \right] \\ &- \frac{1}{2} \max_{y\in\mathscr{X}} \left[-\langle g, y\rangle + \alpha_* \log\left(\int_\Omega \exp\left(\phi_*/\alpha_*\right) p_{A(y)} d\mathscr{m} \right) \right]. \end{aligned} \tag{4.15}$$

Given an observation $\omega \in \Omega$ of $\mathrm{Z}_{A(x_{\text{true}})}$, the estimate for $\langle g, x_{\text{true}}\rangle$ is given by $\widehat{g}_*(\omega)$ with an additive error of $\Phi_*(\log(2/\delta)) + 2\epsilon_{\text{inf}}$ for a confidence level of $1-\delta$. The number $\epsilon_{\text{inf}} > 0$ is introduced because there may not exist points $\alpha_* > 0$ and $\phi_* \in \mathscr{F}$ achieving the minimum in Eq. (4.11). Note that computing the estimator $\widehat{g}_*$ requires one to perform optimization and can be computationally costly. However, once the estimator has been computed, the estimates can be obtained using $\widehat{g}_*$ efficiently, assuming that $\phi_*(\omega)$ is easy to compute for all $\omega \in \Omega$.

[62] prove the following results concerning the estimation procedure in Box 1. For all the results discussed below, we assume that the mathematical premise of Sec. 4.1 holds. In particular, $(\mathscr{D}, \mathscr{M})$ is a good pair (Def. 4.1).

Proposition 4.7. *1. The function Φ_r defined in Eq. (4.10) is continuous and concave in $(x, y) \in \mathscr{X} \times \mathscr{X}$, and continuous and convex in $(\phi, \alpha) \in \mathscr{F} \times (0, \infty)$.*

2. Φ_r has a well-defined saddle-point value, $2\Phi_*(r)$, that satisfies $\Phi_*(r) \geq 0$.

3. The estimator $\widehat{g}_*$ constructed in Eq. (4.14) satisfies

$$\mathbb{P}_{\mathrm{A}(x_{\mathrm{true}})}\left(|\widehat{g}_* - \langle g, x_{\mathrm{true}}\rangle| \leq \Phi_*(\log(2/\delta)) + 2\epsilon_{\mathrm{inf}}\right) > 1 - \delta \tag{4.16}$$

for all $x_{\mathrm{true}} \in \mathscr{X}$ and $1 - \delta \in (0.75, 1)$.

The main result of [62] is that the estimator $\widehat{g}_*$ is minimax optimal up to a small constant factor in the sense noted below.

Theorem 4.8 (Lem. 3.2, [62]). *For $\delta \in (0, 0.25)$, the estimation error $\Phi_*(\log(2/\delta))$ satisfies*

$$\Phi_*(\log(2/\delta)) \leq \frac{2\log(2/\delta)}{\log(1/(4\delta))}\mathscr{R}_*(\delta), \tag{4.17}$$

where $\mathscr{R}_*(\delta)$ is the minimax optimal risk defined in Eq. (4.4).

Thm. 4.8 guarantees that $\Phi_*(\log(2/\delta))$, which is the estimation error of $\widehat{g}_*$, is within a multiplicative factor of $2\log(2/\delta)/\log(1/(4\delta))$ of the smallest possible error, given the mathematical premise of Sec. 4.1. Finally, we note a useful expression for $\Phi_*(r)$ given by [62].

Proposition 4.9 (Prop. 3.1, [62]). *The saddle-point value of the function Φ_r in Eq. (4.10) can be expressed as*

$$2\Phi_*(r) = \max_{x,y \in \mathscr{X}} \left\{\langle g, x\rangle - \langle g, y\rangle \;\middle|\; \mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y)) \geq \exp(-r)\right\}. \tag{4.18}$$

Moreover, the Bhattacharyya coefficient $\mathrm{BC}(\mu, \nu)$ *is a continuous and log-concave function of* $(\mu, \nu) \in \mathscr{M} \times \mathscr{M}$.

Since the Bhattacharyya coefficient is a log-concave function, the optimization defining Φ_* in Eq. (4.18) is convex, as we can take logarithm on both sides of the constraint. In the next section, we discuss some drawbacks of the estimation procedure given in Box 1, and subsequently, we propose a different estimation procedure with the same guarantees as [62].

4.3 Simplified estimation procedure

In this section, we present a slightly modified version of the estimation procedure developed in [89], and prove that it satisfies the same guarantees as the estimation procedure of [62]. Subsequently, we present some new results concerning this estimation procedure.

We begin by presenting the motivation for developing a different estimation procedure instead of using the procedure given in Box 1.

(1) The space of affine estimators $\mathscr{F}$ can be high-dimensional, especially when we have many outcomes from different random variables. This can make the minimization $\inf_{\alpha>0, \phi \in \mathscr{F}} \overline{\Phi}_{\log(2/\delta)}(\phi, \alpha)$ in Box 1 costly to implement.

(2) When $\mathscr{X}$ is high-dimensional, the computation of the function $\overline{\Phi}_{\log(2/\delta)}(\phi, \alpha) = \max_{x,y \in \mathscr{X}} \Phi_{\log(2/\delta)}(x, y; \phi, \alpha)$ can be costly, since for each ϕ, α, one needs to maximize $\Phi_{\log(2/\delta)}(x, y; \phi, \alpha)$ over $\mathscr{X} \times \mathscr{X}$.

(3) Since $\overline{\Phi}_{\log(2/\delta)}(\phi, \alpha) = \max_{x,y \in \mathscr{X}} \Phi_{\log(2/\delta)}(x, y; \phi, \alpha)$ is itself a maximum of the function $\Phi_{\log(2/\delta)}(x, y; \phi, \alpha)$, gradient based methods can be difficult to use for performing the minimization $\min_{\alpha>0, \phi \in \mathscr{F}} \overline{\Phi}_{\log(2/\delta)}(\phi, \alpha)$ over $\phi \in \mathscr{F}$ and $\alpha > 0$, even when the function $\Phi_{\log(2/\delta)}(x, y; \phi, \alpha)$ is smooth in ϕ and α. While subgradient methods can be used, they typically a take longer time to converge than gradient based methods.

(4) The estimator constructed in Eq. (4.14) is hard to study analytically because it depends

on ϕ_* which is defined implicitly through optimization, and the constant c is computed by solving a different optimization problem.

Our approach to constructing the estimation essentially amounts to solving the saddle point problem in Eq. (4.11) by first minimizing over $\phi \in \mathscr{F}$, then maximizing over $x, y \in \mathscr{X}$, and finally, minimizing over $\alpha > 0$. This is motivated by the observation that the minimization over ϕ can be calculated analytically. Thus, we circumvent the optimization over ϕ, which eliminates a costly part of computation compared to the estimation procedure of [62]. Moreover, the estimator we construct is more amenable to analytical treatment. We present our estimation procedure below, assuming the premise of Sec. 4.1.

Box 2: Estimation procedure proposed in [89]

(1) For $r \geq 0$, define the function $\Phi'_r \colon (\mathscr{X} \times \mathscr{X}) \times \mathbb{R}_+ \to \mathbb{R}$ as

$$\Phi'_r(x, y; \alpha) = 2\alpha r + \langle g, x\rangle - \langle g, y\rangle + 2\alpha \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))), \tag{4.19}$$

and denote

$$\Phi'_*(r) = \frac{1}{2} \min_{\alpha \geq 0} \max_{x,y \in \mathscr{X}} \Phi'_r(x, y; \alpha). \tag{4.20}$$

(2) Given a confidence level $1 - \delta \in (0, 1)$, find $\alpha_* \geq 0$ attaining the minimum in

$$2\Phi'_*(\log(2/\delta)) = \min_{\alpha \geq 0} \left[2\alpha \log(2/\delta) + \max_{x,y \in \mathscr{X}} \left(\langle g, x\rangle - \langle g, y\rangle + 2\alpha \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))) \right) \right], \tag{4.21}$$

and find points $x^*, y^* \in \mathscr{X}$ that attain the maximum in $\max_{x,y \in \mathscr{X}} \left(\langle g, x\rangle - \langle g, y\rangle + 2\alpha_* \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))) \right)$, so that

$$2\Phi'_*(\log(2/\delta)) = \Phi'_{\log(2/\delta)}(x^*, y^*; \alpha_*). \tag{4.22}$$

(3) Define

$$\phi_* = \frac{\alpha_*}{2} \log\left(\frac{p_{\mathrm{A}(x^*)}}{p_{\mathrm{A}(y^*)}}\right). \tag{4.23}$$

(4) The estimator $\widehat{g}_*$ is then obtained by setting

$$\widehat{g}_* = \phi_* + \frac{1}{2}\left(\langle g, x^*\rangle + \langle g, y^*\rangle\right). \tag{4.24}$$

Observe that our algorithm does not require one to compute the minimum over $\phi \in \mathscr{F}$, thus reducing the computational cost compared to the procedure of [62]. Furthermore, the estimator given in Eq. (4.24) is appealing from a theoretical standpoint because we have a closed-form expression in terms of the saddle-points (x^*, y^*) and α^*. The main difference between the procedure given in Box 2 and the procedure given in [89] is that we allow $\alpha \geq 0$ in Box 2, as opposed to $\alpha > 0$ in [89].

Since the estimation procedure in Box 2 is different from the estimation procedure of [62] given in Box 1, we need to prove that the estimator constructed in Box 2 satisfies all the guarantees of [62]. We begin by proving that $\Phi'_*(r)$ defined in Eq. (4.20) is equal to $\Phi_*(r)$ defined in Eq. (4.11).

Proposition 4.10. *The following results hold for all $r > 0$.*

1. $\log(\mathrm{BC}(\mu, \nu))$ *is well-defined for all $\mu, \nu \in \mathscr{M}$. It is continuous and concave in $(\mu, \nu) \in \mathscr{M} \times \mathscr{M}$.*

2. *The function $\Phi'_r(x, y; \alpha)$ is continuous and concave in $(x, y) \in \mathscr{X}$ for a fixed $\alpha \geq 0$, and is continuous and convex in $\alpha \geq 0$ for a fixed $(x, y) \in \mathscr{X}$. The inner maximization over $x, y \in \mathscr{X}$ in Eq. (4.21) is a convex optimization problem for each $\alpha \geq 0$, and the outer minimization over $\alpha \geq 0$ is convex.*

3. *The optimization problem in Eq. (4.20) is the dual problem of the optimization problem in Eq. (4.18) and strong duality holds, so that*

$$\Phi'_*(r) = \Phi_*(r). \tag{4.25}$$

Proof. 1. Since for all $\mu, \nu \in \mathscr{M}$, we have $\mathrm{BC}(\mu,\nu) = \int_\Omega \sqrt{p_\mu p_\nu} d\mathfrak{m}$ and $p_\mu, p_\nu > 0$ on Ω by definition of a good pair, we have $\mathrm{BC}(\mu,\nu) > 0$. Consequently, $\log(\mathrm{BC}(\mu,\nu))$ is well-defined for all $\mu, \nu \in \mathscr{M}$. The continuity and concavity of $\log(\mathrm{BC}(\mu,\nu))$ follows from Prop. 4.9.

2. The continuity and convexity properties of Φ_r' can be directly verified. The inner maximization in Eq. (4.21) over $(x,y) \in \mathscr{X} \times \mathscr{X}$ is convex for each $\alpha \geq 0$ because the objective function $\langle g, x\rangle - \langle g, y\rangle + 2\alpha \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y)))$ is concave in (x,y) and $\mathscr{X}$ is a convex set. Since the maximum of a family of convex functions is convex [8, Prop. 8.16], $2\alpha\log(2/\delta) + \max_{x,y\in\mathscr{X}} (\langle g, x\rangle - \langle g, y\rangle + 2\alpha \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))))$ is a convex function of α. It follows that the outer minimization over $\alpha \geq 0$ in Eq. (4.21) is convex.

3. First, rewrite the maximization problem in Eq. (4.18) as

$$2\Phi_*(r) = \max_{x,y\in\mathscr{X}} \{\langle g, x\rangle - \langle g, y\rangle \;\mid\; -2\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))) \leq 2r\}. \tag{4.26}$$

We add the factor of 2 to the constraint to ensure that the dual variable for this constraint coincides with the variable α in Eq. (4.19). The Lagrangian of the concave maximization problem in Eq. (4.26) is given by

$$\mathscr{L}(x,y;\alpha) = \langle g, x\rangle - \langle g, y\rangle + 2\alpha\,(\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))) + r) = \Phi_r'(x,y;\alpha), \tag{4.27}$$

where $x, y \in \mathscr{X}$ are the primal variables and $\alpha \geq 0$ is the dual variable. Since $\mathscr{X} \subseteq \mathbb{R}^D$ is a non-empty convex set, it has a non-empty relative interior [8, Fact. 6.14]. Taking any $x \in \operatorname{relint} \mathscr{X}$ and setting $y = x$, we have $\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y)) = 1$ and $0 = -2\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(x))) < 2r$, so that Slater's condition holds. Therefore, strong duality holds, and we have

$$2\Phi_*(r) = \inf_{\alpha\geq 0} \max_{x,y\in\mathscr{X}} \mathscr{L}(x,y;\alpha) = \inf_{\alpha\geq 0} \max_{x,y\in\mathscr{X}} \Phi_r'(x,y;\alpha) = 2\Phi_*'(r). \tag{4.28}$$

□

Next, we prove that Φ'_r in Eq. (4.19) has a saddle point in $(x, y) \in \mathscr{X}$ and $\alpha \geq 0$, and that the α-component of the saddle point is unique.

Proposition 4.11. *For $r > 0$, the following statements hold.*

1. *There is some $\alpha_* \geq 0$ that attains the minimum in Eq. (4.20).*

2. *The minimum over $\alpha \geq 0$ in Eq. (4.20) is attained at a unique $\alpha_* \geq 0$.*

3. *Φ'_r defined in Eq. (4.19) has a saddle point $(x^*, y^*; \alpha_*)$, where $x^*, y^* \in \mathscr{X}$ and $\alpha_* \geq 0$. Consequently, $2\Phi'_*(r) = \Phi'_r(x^*, y^*; \alpha_*)$.*

4. *$(x^*, y^*; \alpha_*)$ is a saddle point of Φ'_r if and only if $x^*, y^* \in \mathscr{X}$ attain the maximum in Eq. (4.18) and $\alpha_* \geq 0$ is a dual optimal of the optimization problem in Eq. (4.18). Consequently, we have*

$$\Phi'_*(r) = \frac{1}{2}\left(\langle g, x^*\rangle - \langle g, y^*\rangle\right). \tag{4.29}$$

Proof. 1. Eq. (4.28) shows that we can write $2\Phi'_*(r) = \inf_{\alpha\geq 0} \max_{x,y\in\mathscr{X}} \Phi'_r(x, y; \alpha)$. It remains to show that the infimum over $\alpha \geq 0$ can be replaced by a minimum. Denote $f_r(\alpha) = 2\alpha r + \max_{x,y\in\mathscr{X}}(\langle g, x\rangle - \langle g, y\rangle + 2\alpha \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))))$ and write $2\Phi'_*(r) = \inf_{\alpha\geq 0} f_r(\alpha)$.
Since $\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(x))) = 0$ for all $x \in \mathscr{X}$, we have the lower bound $f_r(\alpha) \geq 2\alpha r$. Therefore, $\lim_{\alpha\to\infty} f_r(\alpha) = \infty$, from which it follows that f_r is a coercive function. Since $\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))) \leq 0$ for all $x, y \in \mathscr{X}$ and $\alpha \geq 0$, we have $f_r(\alpha) \leq 2\alpha r + \max_{x,y\in\mathscr{X}}(\langle g, x\rangle - \langle g, y\rangle)$, so that f_r is a proper function. f_r is a convex function since the supremum of a family of convex functions is convex [8, Prop. 8.16]. Similarly, f_r is lsc because the supremum of a family of lsc functions is lsc [8, Lem. 1.26]. Then, by [8, Prop. 11.15], we can infer that f_r has a minimizer in $[0, \infty)$.

2. If the minimum of $f_r(\alpha)$ over $\alpha \geq 0$ occurs at $\alpha_* = 0$ and it is unique, then the statement holds. Thus, assume that there is at least one $\alpha_* > 0$ that attains the minimum. Note that the set $[0, \infty)$ is strictly convex in the sense that for all $\alpha, \beta \in [0, \infty)$ with $\alpha \neq \beta$, we have $(\alpha+\beta)/2 \in (0, \infty)$. Then, by [8, Prop. 11.8], $f_r(\alpha)$ has at most one minimizer in $[0, \infty)$, which implies that α_* must be unique.

3. By Prop. 4.10.1, $\Phi'_r(x, y; \alpha)$ is continuous and concave in $(x, y) \in \mathcal{X} \times \mathcal{X}$ for all $\alpha \geq 0$, and continuous and convex in $\alpha \geq 0$ for all $x, y \in \mathcal{X}$. Since $\mathcal{X}$ is compact and convex and $[0, \infty)$ is convex, by Sion-Kakutani minimax theorem [97], we have $2\Phi'_*(r) = \min_{\alpha \geq 0} \max_{x,y \in \mathcal{X}} \Phi'_r(x, y; \alpha) = \max_{x,y \in \mathcal{X}} \inf_{\alpha \geq 0} \Phi'_r(x, y; \alpha)$. By Prop. 4.10.3, we have $2\Phi'_*(r) = 2\Phi_*(r)$. It can be verified that $2\Phi'_*(r) = \max_{x,y \in \mathcal{X}} \inf_{\alpha \geq 0} \Phi'_r(x, y; \alpha)$ gives Eq. (4.18) after performing the minimization over $\alpha \geq 0$, since $\inf_{\alpha \geq 0} \alpha(r + \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y)))) = -\infty$ if $\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))) < -r$. Since the maximum of a continuous function on a compact set is attained at a point in the compact set, the set $\{(x, y) \in \mathcal{X} \times \mathcal{X} \mid -\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y)) \leq r\}$ is compact (since the intersection of a compact and a closed set is compact), and $\langle g, x \rangle - \langle g, y \rangle$ is a continuous function of (x, y), the maximum in Eq. (4.18) is always attained. Let $x^*, y^* \in \mathcal{X}$ be points that attain the maximum in Eq. (4.18), and let $\alpha_* \geq 0$ be the (unique) point that attains minimum in $\min_{\alpha \geq 0} \max_{x,y \in \mathcal{X}} \Phi'_r(x, y; \alpha)$. Then, we have $\min_{\alpha \geq 0} \max_{x,y \in \mathcal{X}} \Phi'_r(x, y; \alpha) = \max_{x,y \in \mathcal{X}} \Phi'_r(x, y; \alpha_*) \geq \Phi'_r(x^*, y^*; \alpha_*) \geq \inf_{\alpha \geq 0} \Phi'_r(x^*, y^*; \alpha) = \max_{x,y \in \mathcal{X}} \inf_{\alpha \geq 0} \Phi'_r(x, y; \alpha)$. Since $\min_{\alpha \geq 0} \max_{x,y \in \mathcal{X}} \Phi'_r(x, y; \alpha) = \max_{x,y \in \mathcal{X}} \inf_{\alpha \geq 0} \Phi'_r(x, y; \alpha) = 2\Phi'_*(r)$, we can conclude that $(x^*, y^*; \alpha_*)$ is a saddle point of Φ'_r with $2\Phi'_*(r) = \Phi'_r(x^*, y^*; \alpha_*)$.

4. If $(x^*, y^*; \alpha_*)$ is a saddle point of Φ'_r, then x^*, y^* attains the maximum in $2\Phi'_*(r) = \max_{x,y \in \mathcal{X}} \inf_{\alpha \geq 0} \Phi'_r(x, y; \alpha)$. Since performing the minimization over $\alpha \geq 0$ in $\max_{x,y \in \mathcal{X}} \inf_{\alpha \geq 0} \Phi'_r(x, y; \alpha)$ gives Eq. (4.18), we can conclude that $x^*, y^* \in \mathcal{X}$ attain the maximum in Eq. (4.18). Since $\alpha_* \geq 0$ attains the minimum in $\min_{\alpha \geq 0} \max_{x,y \in \mathcal{X}} \Phi'_r(x, y; \alpha)$, by Prop. 4.10.3, can infer that $\alpha_* \geq 0$ is the dual optimal of Eq. (4.18).

Now, suppose that x^*, y^* attains the maximum in Eq. (4.18). Then, x^*, y^* attains the maximum in $\max_{x,y \in \mathcal{X}} \inf_{\alpha \geq 0} \Phi'_r(x, y; \alpha)$, so that it is a valid (x, y)-component of the saddle point of Φ'_r. By Prop. 4.10.3, a dual optimal $\alpha_* \geq 0$ of the optimization in Eq. (4.18) attains the minimum in $\min_{\alpha \geq 0} \max_{x,y \in \mathcal{X}} \Phi'_r(x, y; \alpha)$, so that it is a valid α-component of the saddle point of Φ'_r. □

Owing to Prop. 4.10.3, we can use any primal optimal points $x^*, y^* \in \mathcal{X}$ that attain the maximum in Eq. (4.18), and the (unique) dual optimal point $\alpha_* \geq 0$ for the optimization in Eq. (4.18) to compute the estimator in Box 2. We have also shown that Φ'_r defined in Eq. (4.19)

has a well-defined saddle-point value that is equal to $2\Phi_*(r)$. Next, we need to find a $\phi \in \mathscr{F}$ so as to construct a nearly-optimal affine estimator for $\langle g, x_{\text{true}} \rangle$. For this purpose, we show that for a fixed $x, y \in \mathscr{X}$ and fixed $\alpha \geq 0$, the function Φ'_r is obtained as a minimization of Φ_r defined in Eq. (4.10) over all ϕ. Using this observation, we can find a suitable ϕ_* using the saddle point $(x^*, y^*; \alpha_*)$ of Φ'_r. Before constructing such a ϕ_*, we present the following useful characterization of coercivity for proper, lsc, convex functions on a finite-dimensional space. The version of this result for real-valued convex functions is stated in [43] without proof.

Proposition 4.12. *Let $\mathscr{V}$ be a finite-dimensional real vector space and let $f \colon \mathscr{V} \to \overline{\mathbb{R}}$ be a proper, lsc, convex function. Let $x_0 \in \mathscr{V}$ be a point where $f(x_0)$ is finite. Then, f is coercive if and only if for all non-zero $x \in \mathscr{V}$, we have $\lim_{t\to\infty} f(x_0 + tx) = \infty$.*

In particular, if $f \colon \mathscr{V} \to \mathbb{R}$ is a convex function, then f is coercive if and only if for all non-zero $x \in \mathscr{V}$, we have $\lim_{t\to\infty} f(tx) = \infty$

Proof. Given $\eta \in \mathbb{R}$, denote $\operatorname{lev}_{\leq\eta} f = \{x \in \mathscr{V} \mid f(x) \leq \eta\}$ to be the sublevel set of f at height η. It can be verified that $\operatorname{lev}_{\leq\eta} f$ is convex for all $\eta \in \mathbb{R}$. A set $K \subseteq \mathscr{V}$ is said to be a cone if for all $x \in K$ and all $\alpha > 0$, we have $\alpha x \in K$. Given a non-empty convex set $C \subseteq \mathscr{V}$, let $\operatorname{rec} C = \{x \in \mathscr{V} \mid x + C \subseteq C\}$ denote the recession cone of C. See [8, Prop. 6.49] for a proof that $\operatorname{rec} C$ is a convex cone.

It follows from the definition of coercivity that if f is coercive, then $\lim_{t\to\infty} f(x_0 + tx) = \infty$ for all non-zero $x \in \mathscr{V}$. Therefore, suppose that for all non-zero $x \in \mathscr{V}$, we have $\lim_{t\to\infty} f(x_0 + tx) = \infty$. Let $\xi = f(x_0) \in \mathbb{R}$, so that $x_0 \in \operatorname{lev}_{\leq\xi} f$. Assume, towards a contradiction, that $\operatorname{lev}_{\leq\xi} f$ is unbounded. Then, by [8, Cor. 6.52], there is some non-zero $y \in \operatorname{rec} \operatorname{lev}_{\leq\xi} f$. Since $\operatorname{rec} \operatorname{lev}_{\leq\xi} f$ is a cone, we have $ty \in \operatorname{rec} \operatorname{lev}_{\leq\xi} f$ for all $t > 0$. Because $x_0 \in \operatorname{lev}_{\leq\xi} f$, by the definition of a recession cone, we have $x_0 + ty \in \operatorname{lev}_{\leq\xi} f$ for all $t > 0$. But $\lim_{t\to\infty} f(x_0 + ty) = \infty$ by assumption, which contradicts $x_0 + ty \in \operatorname{lev}_{\leq\xi} f$ for all $t > 0$. Therefore, $\operatorname{lev}_{\leq\xi} f$ must be bounded. It follows from [8, Prop. 11.13] that f is coercive.

Now, if f is real-valued and convex, then it is proper and continuous [8, Cor. 8.40]. Then,

taking $x_0 = 0$ gives the desired result. □

We now show how to construct a ϕ_*.

Lemma 4.13. *Let $\alpha > 0$ be fixed. Then, for $r \geq 0$, the function*

$$\Phi_r^\alpha(x, y; \phi) = \langle g, x\rangle - \langle g, y\rangle \; + \alpha\Bigg[\log\left(\int_\Omega \exp(-\phi/\alpha)p_{\mathrm{A}(x)}d\mathfrak{m}\right) + \log\left(\int_\Omega \exp(\phi/\alpha)p_{\mathrm{A}(y)}d\mathfrak{m}\right)\Bigg] + 2\alpha r \tag{4.30}$$

defined on $(\mathscr{X}\times\mathscr{X})\times\mathscr{F}$ has a saddle point $(x^*, y^*; \phi_*)$ for $x^*, y^* \in \mathscr{X}$ and $\phi_* \in \mathscr{F}$. ϕ_* can be chosen as

$$\phi_* = \frac{\alpha}{2}\log\left(\frac{p_{\mathrm{A}(x^*)}}{p_{\mathrm{A}(y^*)}}\right). \tag{4.31}$$

Furthermore, (x^*, y^*) is the (x, y)-component of the saddle point of Φ_r^α if and only if it attains the maximum in $\max_{x,y\in\mathscr{X}}\left(2\alpha r + \langle g, x\rangle - \langle g, y\rangle + 2\alpha\log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y)))\right)$, and therefore, the saddle-point value of Φ_r^α is equal to

$$\Phi_r^\alpha(x^*, y^*; \phi_*) = 2\alpha r + \langle g, x^*\rangle - \langle g, y^*\rangle + 2\alpha\log(\mathrm{BC}(\mathrm{A}(x^*), \mathrm{A}(y^*))). \tag{4.32}$$

Proof. We adapt the proof of [43, Thm. 2.1] to show this result. From Prop. 4.7.1, we know that Φ_r^α is continuous and concave in $(x, y) \in \mathscr{X}\times\mathscr{X}$ and continuous and convex in $\phi \in \mathscr{F}$. Then, since $\mathscr{X}$ is compact and $\mathscr{F}$ is a finite-dimensional vector space, it follows from Sion-Kakutani minimax theorem [97] that Φ_r^α has a well-defined saddle-point value

$$\inf_{\phi\in\mathscr{F}}\max_{x,y\in\mathscr{X}}\Phi_r^\alpha(x, y; \phi) = \sup_{x,y\in\mathscr{X}}\inf_{\phi\in\mathscr{F}}\Phi_r^\alpha(x, y; \phi). \tag{4.33}$$

Since $\Phi_r^\alpha(x, y; \phi)$ is continuous in $x, y \in \mathscr{X}$ for each $\phi \in \mathscr{F}$, $\inf_{\phi\in\mathscr{F}}\Phi^\alpha(x, y; \phi)$ is upper semi-continuous in $(x, y) \in \mathscr{X}\times\mathscr{X}$ [8, Lem. 1.26]. Then, since $\mathscr{X}$ is compact, the maximum $\max_{x,y\in\mathscr{X}}\inf_{\phi\in\mathscr{F}}\Phi^\alpha(x, y; \phi)$ is attained in $\mathscr{X}\times\mathscr{X}$. Therefore, to show the existence of a saddle point, it suffices to show that

the minimum $\inf_{\phi\in\mathscr{F}} \max_{x,y\in\mathscr{X}} \Phi^\alpha(x,y;\phi)$ is attained in $\mathscr{F}$. To avoid technicalities concerning zero m-measure sets, in the remainder of the proof, we identify functions in $\mathscr{F}$ that are equal upto m-measure zero, and redefine $\mathscr{F}$ accordingly. This does not affect any calculations because ϕ appears in $\Phi^\alpha_r(x,y;\phi)$ only through integrals with respect to m.

Observe that for all $s\in\mathbb{R}$ and all $x,y\in\mathscr{X}$, we have $\Phi^\alpha_r(x,y;\phi+s)=\Phi^\alpha_r(x,y;\phi)$. Thus, we restrict our attention to the subspace $\mathscr{F}_0=\{\phi\in\mathscr{F}\mid\int_\Omega \phi p_\nu dm=0\}$ for a fixed $\nu\in\mathscr{M}$. Since, by the definition of a good pair, $\int_\Omega \phi p_\nu dm$ is well-defined for all $\phi\in\mathscr{F}$, the existence of a minimum of $\Phi^\alpha_r(x,y;\phi)$ over $\phi\in\mathscr{F}_0$ implies an existence of a minimum of $\Phi^\alpha_r(x,y;\phi)$ over $\phi\in\mathscr{F}$. Because $\max_{x,y\in\mathscr{X}}\Phi^\alpha(x,y;\phi)$ is a well-defined, convex, lsc function of $\phi\in\mathscr{F}_0$, to show that $\inf_{\phi\in\mathscr{F}_0}\max_{x,y\in\mathscr{X}}\Phi^\alpha(x,y;\phi)$ has a minimum in $\mathscr{F}_0$, it suffices to prove that $\max_{x,y\in\mathscr{X}}\Phi^\alpha_r(x,y;\phi)$ is coercive in $\phi\in\mathscr{F}_0$ [8, Prop. 11.15]. To that end, we show that for all $x,y\in\mathscr{X}$, $\Phi^\alpha(x,y;\phi)$ is coercive for $\phi\in\mathscr{F}_0$. Since $\Phi^\alpha_r(x,y;\phi)\le\max_{x,y\in\mathscr{X}}\Phi^\alpha_r(x,y;\phi)$, this also shows that $\max_{x,y\in\mathscr{X}}\Phi^\alpha_r(x,y;\phi)$ is coercive in $\phi\in\mathscr{F}_0$.

For $x,y\in\mathscr{X}$, write $\Phi^\alpha_r(x,y;\phi)=\langle g,x\rangle-\langle g,y\rangle+2\alpha r+\alpha\Theta^{x,y}(\phi/\alpha)$, where

$$\Theta^{x,y}(\phi)=\log\left(\int_\Omega \exp(-\phi)p_{\mathrm{A}(x)}dm\right)+\log\left(\int_\Omega\exp(\phi)p_{\mathrm{A}(y)}dm\right). \tag{4.34}$$

Since $\Theta^{x,y}(\phi)$ is a real-valued convex function on $\mathscr{F}_0$, to show that it is coercive in ϕ, it suffices to prove that $\Theta^{x,y}(t\phi)\to\infty$ as $t\to\infty$ for all non-zero $\phi\in\mathscr{F}_0$ (see Prop. 4.12). For all non-zero $\phi\in\mathscr{F}_0$, we have $\int_\Omega\max\{\phi,0\}p_\nu dm=\int_\Omega\max\{-\phi,0\}p_\nu dm>0$ since $p_\nu>0$ on Ω. Then, because $e^z>\max\{z,0\}$ for all $z\in\mathbb{R}$, we have $\Theta^{x,y}(\phi)>\log(\int_\Omega\max\{-\phi,0\}p_{\mathrm{A}(x)}dm)+\log(\int_\Omega\max\{\phi,0\}p_{\mathrm{A}(y)}dm)$. Since $p_{\mathrm{A}(x)},p_{\mathrm{A}(y)}>0$ on Ω for all $x,y\in\mathscr{X}$, we can conclude that $\Theta^{x,y}(t\phi)\to\infty$ as $t\to\infty$. It follows that $\Theta^{x,y}(\phi)$, and therefore, $\Phi^\alpha(x,y;\phi)$ is coercive in $\phi\in\mathscr{F}_0$ for all $x,y\in\mathscr{X}$. Therefore, Φ^α_r has a saddle point.

Now, suppose that (x^*,y^*) is the (x,y)-component of the saddle-point of Φ^α_r (i.e., attains the maximum of the function $\inf_\phi\Phi^\alpha_r(x,y;\phi)$). Then, if ϕ_* is the ϕ-component of the saddle point (i.e., attains the minimum of the function $\max_{x,y}\Phi^\alpha_r(x,y;\phi)$), it minimizes the function $\Phi^\alpha_r(x^*,y^*;\phi)$. This

is because $\Phi_r^\alpha(x^*, y^*; \phi_*) \geq \inf_\phi \Phi_r^\alpha(x^*, y^*; \phi) = \max_{x,y} \inf_\phi \Phi_r^\alpha(x, y; \phi) = \inf_\phi \max_{x,y} \Phi_r^\alpha(x, y; \phi) = \max_{x,y} \Phi_r^\alpha(x, y; \phi_*) \geq \Phi_r^\alpha(x^*, y^*; \phi_*)$. It follows that ϕ_*/α minimizes the function $\Theta^{x^*,y^*}(\phi/\alpha)$. Thus, we compute the minimum of $\Theta^{x,y}(\phi)$ over $\phi \in \mathscr{F}$ following [43, Thm. 2.1] and [62, Prop. 3.1].

For a given $x, y \in \mathscr{X}$, denote $\phi_{x,y} = (1/2)\log(p_{\mathrm{A}(x)}/p_{\mathrm{A}(y)}) \in \mathscr{F}$. Write any given $\phi \in \mathscr{F}$ as $\phi = \phi_{x,y} + \Delta$. Then, by Hölder's inequality, we have

$$\begin{aligned}
\exp\left(\frac{1}{2}\Theta^{x,y}(\phi_{x,y})\right) &= \int_\Omega \sqrt{p_{\mathrm{A}(x)} p_{\mathrm{A}(y)}}\, dm \\
&= \int_\Omega \left[(p_{\mathrm{A}(x)} p_{\mathrm{A}(y)})^{1/4} e^{-\Delta/2}\right] \left[(p_{\mathrm{A}(x)} p_{\mathrm{A}(y)})^{1/4} e^{\Delta/2}\right] dm \\
&\leq \sqrt{\int_\Omega (p_{\mathrm{A}(x)} p_{\mathrm{A}(y)})^{1/2} e^{-\Delta} dm} \sqrt{\int_\Omega (p_{\mathrm{A}(x)} p_{\mathrm{A}(y)})^{1/2} e^{\Delta} dm} \\
&= \exp\left(\frac{1}{2}\Theta^{x,y}(\phi)\right).
\end{aligned} \tag{4.35}$$

Since equality in Hölder's inequality holds if and only if $(p_{\mathrm{A}(x)} p_{\mathrm{A}(y)})^{1/2} e^{-2\Delta} = \zeta (p_{\mathrm{A}(x)} p_{\mathrm{A}(y)})^{1/2} e^{2\Delta}$ for some $\zeta \in \mathbb{R}$, Δ must be constant for equality. Therefore, every minimum of $\Theta^{x,y}$ is of the form $\phi^{x,y} + s$ for $s \in \mathbb{R}$. Since $\Phi_r^\alpha(x, y; \phi)$ is invariant under translations of the form $\phi \mapsto \phi + s$ for $s \in \mathbb{R}$, we can choose ϕ_* as

$$\frac{\phi_*}{\alpha} = \frac{1}{2}\log\left(\frac{p_{\mathrm{A}(x^*)}}{p_{\mathrm{A}(y^*)}}\right). \tag{4.36}$$

Since $\phi_{x,y} = (1/2)\log(p_{\mathrm{A}(x)}/p_{\mathrm{A}(y)})$ minimizes $\Theta^{x,y}(\phi)$ for all $x, y \in \mathscr{X}$, we can verify by direct substitution that $\max_{x,y\in\mathscr{X}} \inf_{\phi\in\mathscr{F}} \Phi_r^\alpha(x, y; \phi) = \max_{x,y\in\mathscr{X}} (2\alpha r + \langle g, x\rangle - \langle g, y\rangle + 2\alpha \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))))$. Since Φ_r^α has a well-defined saddle-point value, (x^*, y^*) is the (x, y)-component of the saddle point if and only if it attains the maximum in $\max_{x,y\in\mathscr{X}} \inf_{\phi\in\mathscr{F}} \Phi_r^\alpha(x, y; \phi)$, from which Eq. (4.32) follows. □

We are now in a position to show that the estimator constructed using Box 2 has the same estimation error as the estimator constructed using Box 1.

Theorem 4.14. *Given a confidence level $1 - \delta \in (0, 1)$, let $\alpha_* \geq 0$ attain the minimum in*

$$2\Phi_*(\log(2/\delta)) = \min_{\alpha\geq 0} \left[2\alpha \log(2/\delta) + \max_{x,y\in\mathscr{X}} (\langle g, x\rangle - \langle g, y\rangle + 2\alpha \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))))\right] \tag{4.37}$$

and $x^*, y^* \in \mathcal{X}$ *attain the maximum in* $\max_{x,y\in\mathcal{X}} (\langle g, x\rangle - \langle g, y\rangle + 2\alpha_* \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))))$. *Define*

$$\phi_* = \frac{\alpha_*}{2} \log\left(\frac{p_{\mathrm{A}(x^*)}}{p_{\mathrm{A}(y^*)}}\right). \tag{4.38}$$

Then, the estimator

$$\widehat{g}_* = \phi_* + \frac{1}{2}\left(\langle g, x^*\rangle + \langle g, y^*\rangle\right) \tag{4.39}$$

satisfies

$$\mathbb{P}_{\mathrm{A}(x_{\mathrm{true}})}\left(|\widehat{g}_* - \langle g, x_{\mathrm{true}}\rangle| \le \Phi_*(\log(2/\delta))\right) \ge 1 - \delta \tag{4.40}$$

for all $x_{\mathrm{true}} \in \mathcal{X}$.

Proof. For $\alpha_* = 0$, we have $\widehat{g}_* = (\langle g, x^*\rangle + \langle g, y^*\rangle)/2$ and $\Phi_*(r) = (\langle g, x^*\rangle - \langle g, y^*\rangle)/2$. Then, since x^*, y^* attain the maximum in $\max_{x,y\in\mathcal{X}} (\langle g, x\rangle - \langle g, y\rangle)$, for all $x_{\mathrm{true}} \in \mathcal{X}$, we have

$$\begin{aligned}
&\widehat{g}_* - \langle g, x_{\mathrm{true}}\rangle = \langle g, x^*\rangle - \langle g, x_{\mathrm{true}}\rangle - \Phi_*(r) \le (\langle g, x^*\rangle - \langle g, y^*\rangle) - \Phi_*(r) = \Phi_*(r) \\
&\langle g, x_{\mathrm{true}}\rangle - \widehat{g}_* = \langle g, x_{\mathrm{true}}\rangle - \langle g, y^*\rangle - \Phi_*(r) \le (\langle g, x^*\rangle - \langle g, y^*\rangle) - \Phi_*(r) = \Phi_*(r).
\end{aligned} \tag{4.41}$$

Therefore, $|\widehat{g}_* - \langle g, x_{\mathrm{true}}\rangle| \le \Phi_*(r)$ always holds. Thus, for the remainder of the proof, we take $\alpha_* > 0$.

Since (x^*, y^*) attains the maximum in $\max_{x,y\in\mathcal{X}} (\langle g, x\rangle - \langle g, y\rangle + 2\alpha_* \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y))))$, by Lem. 4.13, $(x^*, y^*; \phi_*)$ for ϕ_* defined in Eq. (4.38) is a saddle point of $\Phi^{\alpha_*}_{\log(2/\delta)}$ defined in Eq. (4.30). Consequently, the points x^*, y^* achieve the maximum in $\max_{x,y\in\mathcal{X}} \Phi^{\alpha_*}_{\log(2/\delta)}(x, y; \phi_*)$, so that

$$\begin{aligned}
&\Phi^{\alpha_*}_{\log(2/\delta)}(x, y^*; \phi_*) \le \Phi^{\alpha^*}_{\log(2/\delta)}(x^*, y^*; \phi_*) = 2\Phi_*(\log(2/\delta)) \quad (\forall x \in \mathcal{X}) \\
&\Phi^{\alpha_*}_{\log(2/\delta)}(x^*, y; \phi_*) \le \Phi^{\alpha^*}_{\log(2/\delta)}(x^*, y^*; \phi_*) = 2\Phi_*(\log(2/\delta)) \quad (\forall y \in \mathcal{X}).
\end{aligned} \tag{4.42}$$

Next, we rewrite the constant term in the estimator $\widehat{g}_*$ in Eq. (4.24) in a convenient form. Since $\int_\Omega \exp(-\phi_*/\alpha_*) p_{\mathrm{A}(x^*)} d\boldsymbol{m} = \int_\Omega \exp(\phi_*/\alpha_*) p_{\mathrm{A}(y^*)} d\boldsymbol{m}$ holds for ϕ_* given in Eq. (4.38), we have

$$c \equiv \frac{1}{2}(\langle g, x^*\rangle + \langle g, y^*\rangle) \tag{4.43}$$

$$
\begin{aligned}
&= \frac{1}{2}\left[\langle g, x^*\rangle + \alpha_* \log\left(\int_\Omega \exp(-\phi_*/\alpha_*) p_{\mathrm{A}(x^*)} d\mathscr{m}\right) + \alpha_* \log(2/\delta)\right] \\
&\qquad - \frac{1}{2}\left[-\langle g, y^*\rangle + \alpha_* \log\left(\int_\Omega \exp(\phi_*/\alpha_*) p_{\mathrm{A}(y^*)} d\mathscr{m}\right) + \alpha_* \log(2/\delta)\right] \\
&= \frac{1}{2}\Phi^{\alpha_*}_{\log(2/\delta)}(x^*, y^*; \phi_*) - \left[-\langle g, y^*\rangle + \alpha_* \log\left(\int_\Omega \exp(\phi_*/\alpha_*) p_{\mathrm{A}(y^*)} d\mathscr{m}\right) + \alpha_* \log(2/\delta)\right] \qquad (4.44) \\
&= \left[\langle g, x^*\rangle + \alpha_* \log\left(\int_\Omega \exp(-\phi_*/\alpha_*) p_{\mathrm{A}(x^*)} d\mathscr{m}\right) + \alpha_* \log(2/\delta)\right] - \frac{1}{2}\Phi^{\alpha_*}_{\log(2/\delta)}(x^*, y^*; \phi_*) \qquad (4.45)
\end{aligned}
$$

We have $\widehat{g}_* = \phi_* + c$.

We now prove a slightly more general statement than Eq. (4.40) following the ideas in [62, Lem. 3.1]. To that end, let $\epsilon' \geq 0$ be any non-negative number and define $\varepsilon = \Phi_*(\log(2/\delta)) + \epsilon'$. For all $x_{\mathrm{true}} \in \mathscr{X}$, using Eq. (4.44) and taking $x = x_{\mathrm{true}}$ in Eq. (4.42), we have

$$
\begin{aligned}
&\langle g, x_{\mathrm{true}}\rangle + \alpha_* \log\left(\int_\Omega \exp(-\widehat{g}_*/\alpha_*) p_{\mathrm{A}(x_{\mathrm{true}})} d\mathscr{m}\right) + \alpha_* \log(2/\delta) \\
&\quad = \langle g, x_{\mathrm{true}}\rangle + \alpha_* \log\left(\int_\Omega \exp(-\phi_*/\alpha_*) p_{\mathrm{A}(x_{\mathrm{true}})} d\mathscr{m}\right) + \alpha_* \log(2/\delta) - c \\
&\quad = \Phi^{\alpha_*}_{\log(2/\delta)}(x_{\mathrm{true}}, y^*; \phi_*) - \frac{1}{2}\Phi^{\alpha_*}_{\log(2/\delta)}(x^*, y^*; \phi_*) \qquad (4.46) \\
&\quad \leq \Phi_*(\log(2/\delta)) \\
&\quad = \varepsilon - \frac{\epsilon'}{2}.
\end{aligned}
$$

Similarly, using Eq. (4.45) and taking $y = x_{\mathrm{true}} \in \mathscr{X}$ in Eq. (4.42), we find that

$$
-\langle g, x_{\mathrm{true}}\rangle + \alpha_* \log\left(\int_\Omega \exp(\widehat{g}_*/\alpha_*) p_{\mathrm{A}(x_{\mathrm{true}})} d\mathscr{m}\right) + \alpha_* \log(2/\delta) \leq \varepsilon - \frac{\epsilon'}{2}. \qquad (4.47)
$$

Denoting $\delta' = \delta e^{-\epsilon'/2\alpha_*}$ and $\int_\Omega f p_{\mathrm{A}(x_{\mathrm{true}})} d\mathscr{m} = \mathbb{E}_{\mathrm{A}(x_{\mathrm{true}})}[f]$ for any $(\Omega, \mathscr{B}(\omega))$-measurable function f, we can divide Eq. (4.46) and Eq. (4.45) by $\alpha_* > 0$ and rearrange terms to obtain

$$
\begin{aligned}
\log\left(\mathbb{E}_{\mathrm{A}(x_{\mathrm{true}})}\left[\exp((\langle g, x_{\mathrm{true}}\rangle - \widehat{g}_* - \varepsilon)/\alpha_*)\right]\right) &\leq \log\left(\frac{\delta}{2}\right) - \frac{\epsilon_{\inf}}{2\alpha_*} \equiv \log\left(\frac{\delta'}{2}\right) \\
\log\left(\mathbb{E}_{\mathrm{A}(x_{\mathrm{true}})}\left[\exp((-\langle g, x_{\mathrm{true}}\rangle + \widehat{g}_* - \varepsilon)/\alpha_*)\right]\right) &\leq \log\left(\frac{\delta}{2}\right) - \frac{\epsilon_{\inf}}{2\alpha_*} \equiv \log\left(\frac{\delta'}{2}\right).
\end{aligned} \qquad (4.48)
$$

Then, from Markov's inequality, we have

$$\begin{aligned}\mathbb{P}_{\mathrm{A}(x_{\mathrm{true}})}\left(\langle g, x_{\mathrm{true}}\rangle - \widehat{g}_* - \varepsilon \geq 0\right) &\leq \mathbb{E}_{\mathrm{A}(x_{\mathrm{true}})}[\exp((\langle g, x_{\mathrm{true}}\rangle - \widehat{g}_* - \varepsilon)/\alpha_*)] \leq \frac{\delta'}{2} \\ \mathbb{P}_{\mathrm{A}(x_{\mathrm{true}})}\left(-\langle g, x_{\mathrm{true}}\rangle + \widehat{g}_* - \varepsilon \geq 0\right) &\leq \mathbb{E}_{\mathrm{A}(x_{\mathrm{true}})}[\exp((-\langle g, x_{\mathrm{true}}\rangle + \widehat{g}_* - \varepsilon)/\alpha_*)] \leq \frac{\delta'}{2}.\end{aligned} \tag{4.49}$$

Using the union bound and $\varepsilon = \Phi_*(\log(2/\delta)) + \epsilon'$, we obtain

$$\mathbb{P}_{\mathrm{A}(x_{\mathrm{true}})}\left(|\widehat{g}_* - \langle g, x_{\mathrm{true}}\rangle| \geq \Phi_*(\log(2/\delta)) + \epsilon'\right) \leq \delta' \tag{4.50}$$

for all $\epsilon' \geq 0$ and all $x_{\mathrm{true}} \in \mathcal{X}$. For all $\epsilon' > 0$, we have $\delta' < \delta$, and for $\epsilon' = 0$, we obtain Eq. (4.40). □

Therefore, the estimator $\widehat{g}_*$ constructed using Box 2 has the same estimation error as the estimator constructed using Box 1. Consequently, by Thm. 4.8, the estimator obtained using Box 2 is also minimax optimal under the premise of Sec. 4.1. Since $\Phi_*(\log(2/\delta)$ is within a constant factor of the minimax optimal risk, we can use it as a proxy to study the optimal estimation error. We give two additional results for the estimation procedure given in Box 2 and $\Phi_*(\log(2/\delta))$.

Proposition 4.15. *Let $r > 0$ and let $(x^*, y^*; \alpha_*)$ be computed according to Box 2. Then, the following statements hold.*

1.

$$0 \leq \Phi_*(r) \leq \Phi_*^{\max}(r) \equiv \frac{1}{2}\max_{x,y\in\mathcal{X}}\left(\langle g, x\rangle - \langle g, y\rangle\right) \tag{4.51}$$

2. $\Phi_*(r) = \Phi_*^{\max}(r)$ if and only if $\alpha_* = 0$.

Proof. 1. $\Phi_*(r) \geq 0$ was noted in Prop. 4.7.2. The upper bound is obtained from Eq. (4.18) by dropping the constraint.

2. By Prop. 4.10.3, we have $2\Phi_*(r) = 2\Phi'_*(r) = \min_{\alpha\geq 0}\max_{x,y\in\mathcal{X}}\Phi'_r(x, y; \alpha)$, where Φ'_r is defined in Eq. (4.19). Since α_* attains the minimum in $\min_{\alpha\geq 0}\max_{x,y\in\mathcal{X}}\Phi'_r(x, y; \alpha)$, for $\alpha_* = 0$, we obtain $2\Phi_*(r) = \max_{x,y\in\mathcal{X}}\Phi'_r(x, y; 0) = 2\Phi_*^{\max}(r)$. Conversely, if $\Phi_*(r) = \Phi_*^{\max}(r)$, then

$2\Phi_*^{\max}(r) = 2\Phi'_*(r) = \min_{\alpha\geq 0}\max_{x,y\in\mathscr{X}}\Phi'_r(x,y;\alpha) \leq \max_{x,y\in\mathscr{X}}\Phi'_r(x,y;0) = 2\Phi_*^{\max}(r)$. Thus, $\alpha_* = 0$ is a point that attains the minimum in $2\Phi'_*(r) = \min_{\alpha\geq 0}\max_{x,y\in\mathscr{X}}\Phi'_r(x,y;\alpha)$, and by uniqueness of α_* shown in Prop. 4.11.2, $\alpha_* = 0$ is the only such point. □

Finally, we give an alternate expression for the saddle-point value $2\Phi_*(r)$, obtained using Fenchel-Rockafellar duality. Specifically, we convert the maximization in Eq. (4.18) into a minimization problem, from which we also derive upper bounds on $\Phi_*(r)$.

Proposition 4.16. *Let $\mathscr{X} \subseteq \mathbb{R}^d$ and $\mathscr{M} \subseteq \mathbb{R}^M$. Suppose that* A *is linear. Then, for $r > 0$, the following statements hold.*

1. *For all $g \in \mathbb{R}^d$, we have*

$$\begin{aligned}&\max_{x,y\in\mathscr{X}}(\langle g,x\rangle - \langle g,y\rangle + \log(\mathrm{BC}(\mathrm{A}(x),\mathrm{A}(y))))\\ &\quad= \min_{u,v\in\mathbb{R}^M}\left(S_{\mathscr{X}}(g-\mathrm{A}^\dagger u) + S_{\mathscr{X}}(-g-\mathrm{A}^\dagger v) + \mathrm{BD}^*(u,v)\right),\end{aligned} \tag{4.52}$$

   where $S_{\mathscr{X}}$ is the support function of $\mathscr{X}$.

2. *The saddle-point value $\Phi_*(r)$ can be written as*

$$\Phi_*(r) = \inf_{\alpha>0}\min_{u,v\in\mathbb{R}^M}\alpha\left(S_{\mathscr{X}}\left(\frac{g}{2\alpha}-\mathrm{A}^\dagger u\right) + S_{\mathscr{X}}\left(-\frac{g}{2\alpha}-\mathrm{A}^\dagger v\right) + \mathrm{BD}^*(u,v) + r\right). \tag{4.53}$$

3. *The saddle-point value $\Phi_*(r)$ is bounded above as*

$$\begin{aligned}\Phi_*(r) \leq \inf_{\alpha>0}\min_{u\in\mathbb{R}^M}\quad &\alpha\left(\mathrm{BD}^*(u,-u)+r\right).\\ \text{s.t.}\quad &\mathrm{A}^\dagger u = \frac{g}{2\alpha}\end{aligned} \tag{4.54}$$

   Furthermore, if A *is injective, then we have*

$$\Phi_*(r) \leq \inf_{\alpha>0}\alpha\left(\mathrm{BD}^*\left((\mathrm{A}^+)^\dagger\left(\frac{g}{2\alpha}\right), -(\mathrm{A}^+)^\dagger\left(\frac{g}{2\alpha}\right)\right) + r\right), \tag{4.55}$$

where $\mathrm{A}^+ = (\mathrm{A}^\dagger\mathrm{A})^{-1}\mathrm{A}^\dagger$ *is the Moore-Penrose pseudoinverse of* A.

Proof. 1. First, we note that

$$\begin{aligned}
&\max_{x,y\in\mathscr{X}}(\langle g,x\rangle - \langle g,y\rangle + \log(\mathrm{BC}(\mathrm{A}(x),\mathrm{A}(y))))\\
&\quad= \max_{x,y\in\mathbb{R}^d}\left[\langle g,x\rangle - \langle g,y\rangle - (\mathrm{BD}(\mathrm{A}(x),\mathrm{A}(y)) + \chi_{\mathscr{X}\times\mathscr{X}}(x,y))\right]\\
&\quad= (\chi_{\mathscr{X}\times\mathscr{X}} + \mathrm{BD}\circ(\mathrm{A}\oplus\mathrm{A}))^*(g,-g).
\end{aligned} \tag{4.56}$$

Now, $\chi_{\mathscr{X}\times\mathscr{X}}$ is a proper, lsc, convex function since $\mathscr{X}$ is a compact and convex set. Moreover, from Prop. 4.10.1, we have that $-\log(\mathrm{BC})$ is well-defined, convex, and continuous on $\mathscr{M}\times\mathscr{M}$. From [8, Cor. 6.15], we have that $\operatorname{relint}\mathrm{A}(\mathscr{X}) = \mathrm{A}(\operatorname{relint}\mathscr{X})$ and $\operatorname{relint}(\mathscr{M}\times\mathscr{M} - (\mathrm{A}\oplus\mathrm{A})(\mathscr{X}\times\mathscr{X})) = \operatorname{relint}(\mathscr{M}\times\mathscr{M}) - \operatorname{relint}(\mathrm{A}\oplus\mathrm{A})(\mathscr{X}\times\mathscr{X})$. Since $\mathscr{M}$ is relatively open, the latter set is equal to $\mathscr{M}\times\mathscr{M} - \operatorname{relint}\mathrm{A}(\mathscr{X})\times\operatorname{relint}\mathrm{A}(\mathscr{X})$. Since $\mathscr{X}$ is a non-empty convex set, it has non-empty relatively interior [8, Fact 6.14]. Picking a point $x\in\operatorname{relint}\mathscr{X}$, we have $\mathrm{A}(x)\in\mathrm{A}(\operatorname{relint}\mathscr{X}) = \operatorname{relint}\mathrm{A}(\mathscr{X})$. But, by definition, $\mathrm{A}(x)\in\mathscr{M} = \operatorname{relint}\mathscr{M}$. It follows that $0\in\operatorname{relint}(\mathscr{M}\times\mathscr{M} - (\mathrm{A}\oplus\mathrm{A})(\mathscr{X}\times\mathscr{X}))$. Note that in finite dimensions, the notion of strong relative interior and relative interior coincide (see [8, Fact 6.14]). Thus, by [8, Thm. 15.27], we have

$$(\chi_{\mathscr{X}\times\mathscr{X}} + \mathrm{BD}\circ(\mathrm{A}\oplus\mathrm{A}))^*(g,-g) = \min_{u,v\in\mathbb{R}^M}\left(\chi^*_{\mathscr{X}\times\mathscr{X}}(g-\mathrm{A}^\dagger u,-g-\mathrm{A}^\dagger v) + \mathrm{BD}^*(u,v)\right). \tag{4.57}$$

Since the convex conjugate of characteristic function is the support function, and $\chi_{\mathscr{X}\times\mathscr{X}}(x,y) = \chi_{\mathscr{X}}(x) + \chi_{\mathscr{X}}(y)$ for $x,y\in\mathbb{R}^d$, we obtain $\chi^*_{\mathscr{X}\times\mathscr{X}}(a,b) = S_{\mathscr{X}}(a) + S_{\mathscr{X}}(b)$ for all $a,b\in\mathbb{R}^d$. Combining these observations gives Eq. (4.52).

2. From Prop. 4.10.3, we have

$$\begin{aligned}
2\Phi_*(r) &= \inf_{\alpha>0}\left[2\alpha r + \max_{x,y\in\mathscr{X}}(\langle g,x\rangle - \langle g,y\rangle + 2\alpha\log(\mathrm{BC}(\mathrm{A}(x),\mathrm{A}(y))))\right]\\
&= \inf_{\alpha>0} 2\alpha\left[r + \max_{x,y\in\mathscr{X}}\left(\left\langle\frac{g}{2\alpha},x\right\rangle - \left\langle\frac{g}{2\alpha},y\right\rangle + \log(\mathrm{BC}(\mathrm{A}(x),\mathrm{A}(y)))\right)\right].
\end{aligned} \tag{4.58}$$

Then, using Eq. (4.52), we obtain

$$2\Phi_*(r) = \inf_{\alpha>0} 2\alpha \left[r + \min_{u,v\in\mathbb{R}^M} \left(S_{\mathscr{X}}(g - \mathrm{A}^\dagger u) + S_{\mathscr{X}}(-g - \mathrm{A}^\dagger v) + \mathrm{BD}^*(u,v) \right) \right], \tag{4.59}$$

from which Eq. (4.53) follows.

3. Since $S_{\mathscr{X}}(0) = 0$, taking choosing $u, v \in \mathbb{R}^M$ satisfying $\mathrm{A}^\dagger u = g/2\alpha$ and $v = -u$, we obtain the upper bound in Eq. (4.54). If, in addition, $\mathrm{A}\colon \mathbb{R}^d \to \mathbb{R}^M$ is injective, then $\mathrm{A}^\dagger\mathrm{A}$ is invertible, and A has a Moore-Penrose pseudoinverse $\mathrm{A}^+\colon \mathbb{R}^M \to \mathbb{R}^d$ given by $\mathrm{A}^+ = (\mathrm{A}^\dagger\mathrm{A})^{-1}\mathrm{A}^\dagger$ [79, Sec. 3.6]. Moreover, $(\mathrm{A}^+)^\dagger = (\mathrm{A}^\dagger)^+$ for any linear map A [79, Sec. 3.6]. Thus, if we take $u = (\mathrm{A}^+)^\dagger g'$ for $g' = g/2\alpha$, then $\mathrm{A}^\dagger u = \mathrm{A}^\dagger(\mathrm{A}^+)^\dagger g' = (\mathrm{A}^+\mathrm{A})^\dagger g' = g'$, since $\mathrm{A}^+\mathrm{A} = \mathbb{I}$. We obtain Eq. (4.55) from Eq. (4.54) for this choice of u. □

In Eq. (4.53), the minimization over u, v is unconstrained, while the minimization over $\alpha > 0$, while constrained, is one-dimensional. Thus, if we have an expression for the convex conjugate BD^* of the Bhattacharyya distance, we can use Eq. (4.53) to compute the saddle-point value. If Ω is a finite set equipped with the discrete σ-algebra, the distributions are discrete, and therefore, we can use the closed-form expression for $\mathrm{BD}^*(u, -u)$ given in Eq. (3.11) to compute the upper bounds in Eq. (4.54) and Eq. (4.55).

Chapter 5

$\mathbb{TOOL}$: A minimax optimal procedure for learning expectation values

Our goal in this chapter is to develop an estimation method that can learn the expectation values of observables using a measurement protocol specified by the experimentalist. We will focus on the case of learning the expectation value of a single observable here, since we can use this procedure with the union bound to learn the expectation values of many observables simultaneously in the l_∞-norm. We formulate the quantum problem of learning the expectation value of an observable in Sec. 5.1, and show how it relates to general statistical problem studied in Ch. 4. In Sec. 5.2, we adapt the results of Sec. 4.3 to develop an estimation procedure for learning the expectation values. We call the estimation procedure so obtained $\mathbb{T}$he $\mathbb{O}$ptimal $\mathbb{O}$bservable expectation value $\mathbb{L}$earner, or $\mathbb{TOOL}$. We then discuss some properties of the estimator constructed by $\mathbb{TOOL}$ in Sec. 5.3. Finally, in Sec. 5.4, we present a convex optimization algorithm, along with convergence guarantees, that can be used to construct the estimator and estimation error for $\mathbb{TOOL}$.

5.1 Mathematical formulation

Suppose that $\mathscr{X}$ is the set of d-dimensional density matrices. The state $\rho \in \mathscr{X}$ is prepared by a quantum device, but is not known to us. We are given an observable $\mathscr{O}$, whose expectation value $\langle \mathscr{O} \rangle = \mathrm{Tr}(\mathscr{O}\rho)$ we wish to learn. Since we do not know the true state ρ, we perform measurements on it. Suppose that $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ is the measurement protocol that was, or will be, implemented in the experiment. For each $i \in [L]$, measuring the ith POVM gives an outcome $k \in [M_i]$ with probability $p_\rho^{(i)}(k) = \mathrm{Tr}(E_k^{(i)}\rho)$ according to Born's rule. Since we only have access to the

measurement outcomes, we need to construct an estimator that uses the outcomes observed in the experiment to give an estimate for $\langle \mathcal{O} \rangle$. An estimator in this context is a real-valued function that takes the observed measurement outcomes as input and outputs an estimate for $\langle \mathcal{O} \rangle$. Given a confidence level $1 - \delta \in (0, 1)$, our goal is to find an estimator that estimates the expectation value of $\mathcal{O}$ using the outcomes of $\mathfrak{M}$ with a confidence level of $1 - \delta$ no matter what state ρ is prepared by the device, such that the estimation error is as small as possible.

To proceed, we need to formalize what we mean by "smallest possible estimation error". For a given estimator, the error constructed can depend on the statistical method used, such as the specific concentration inequality used to derived confidence intervals. To avoid such ambiguities, we look at the smallest possible estimation error for a given estimator, as defined below.

Definition 5.1 (δ-risk of an estimator given a measurement protocol). Given an observable $\mathcal{O}$, a confidence level $1 - \delta \in (0, 1)$, and a measurement protocol $\mathfrak{M}$, the δ-risk of the estimator $\widehat{\mathcal{O}}$ for learning the expectation value of $\mathcal{O}$ with confidence level $1 - \delta$ is defined as

$$\mathscr{R}(\widehat{\mathcal{O}}, \mathcal{O}, \mathfrak{M}, \delta) = \inf \left\{ \varepsilon > 0 \mid \inf_{\sigma} \mathbb{P}_{\mathfrak{M},\sigma} \left(|\widehat{\mathcal{O}} - \mathrm{Tr}(\mathcal{O}\sigma)| \leq \varepsilon \right) > 1 - \delta \right\}, \tag{5.1}$$

where $\mathbb{P}_{\mathfrak{M},\sigma}$ is the joint probability distribution over the labels determined by $\mathfrak{M}$ and the state σ as per Born's rule. □

Since the risk $\mathscr{R}(\widehat{\mathcal{O}}, \mathcal{O}, \mathfrak{M}, \delta)$ does not depend on the underlying state or the data, we are working with minimax procedures in the sense defined in Sec. 2.4. Then, we can define the minimax optimal risk, which is obtained by minimizing the risk of an estimator over all estimation procedures.

Definition 5.2 (Minimax optimal risk given a measurement protocol). Given an observable $\mathcal{O}$, a confidence level $1 - \delta \in (0, 1)$ and a measurement protocol $\mathfrak{M}$, the minimax optimal risk for

learning the expectation value of $\mathscr{O}$ using outcomes of $\mathfrak{M}$ to a confidence level of $1-\delta$ is defined as

$$\mathscr{R}_*(\mathscr{O},\mathfrak{M},\delta)=\inf_{\widehat{\mathscr{O}}}\mathscr{R}(\widehat{\mathscr{O}},\mathscr{O},\mathfrak{M},\delta), \tag{5.2}$$

where the infimum is over all estimators. □

The minimax optimal risk $\mathscr{R}_*(\mathscr{O},\mathfrak{M},\delta)$ is the main quantity of interest in practice, as we usually know the measurement protocol $\mathfrak{M}$ that was/will be implemented in an experiment. $\mathscr{R}_*(\mathscr{O},\mathfrak{M},\delta)$ can also help us determine whether implementing the measurement protocol $\mathfrak{M}$ for learning $\langle\mathscr{O}\rangle$ is a scalable as we increase the system size. Note that when we talk about scalability with the system size, we look at measurement protocols $\mathfrak{M}$ and observables $\mathscr{O}$ that have a suitable definition as a function of the system size. For example, $\mathscr{O}$ can be the projector onto an n-qubit GHZ state, and $\mathfrak{M}$ can be the measurements of the stabilizer group of $\mathscr{O}$. Then, the number of qubits n gives a natural notion of system size, and we can study the scaling of the minimax optimal risk with respect to n.

In addition to such practical considerations, it is also useful to know from a theoretical standpoint what measurements are the best to implement for a given observable. For if such measurements happen to be implementable in an experiment, we can implement them to get optimal performance. Thus, we also define the minimax optimal risk that is obtained by minimizing $\mathscr{R}_*(\mathscr{O},\mathfrak{M},\delta)$ over all measurement protocols that use a fixed number of samples N.

Definition 5.3 (Minimax optimal risk over all measurement protocols). Given an observable $\mathscr{O}$ and a confidence level $1-\delta\in(0,1)$, the minimax optimal risk for learning the expectation value of $\mathscr{O}$ to a confidence level of $1-\delta$ using N samples is defined as

$$\mathscr{R}_*(\mathscr{O},N,\delta)=\inf_{\substack{\mathfrak{M}\\ N(\mathfrak{M})=N}}\mathscr{R}_*(\mathscr{O},\mathfrak{M},\delta), \tag{5.3}$$

where the infimum is over all measurement protocols that use N copies of the state. □

The main quantities of interest in this study are $\mathscr{R}_*(\mathscr{O},\mathfrak{M},\delta)$ and $\mathscr{R}_*(\mathscr{O},N,\delta)$. Our goal for this chapter is to construct an estimator that can achieve an estimation error to within a constant

factor of $\mathscr{R}_*(\mathscr{O}, \mathfrak{M}, \delta)$. The problem of finding a measurement protocol that achieves the smallest possible error for a given observable is postponed to Ch. 7.

Now, from Ch. 4, we know that the estimation procedure given in Sec. 4.3 achieves the minimax optimal risk to within a small constant factor, which follows from the results of [62]. Thus, we wish to use this statistical framework to construct an estimator and estimator error for learning the expectation value of $\mathscr{O}$. However, it is not possible to directly apply this framework because a key requirement of [62] is that all the probability densities must be strictly positive. To circumvent this problem, we suppose that for any given measurement protocol $\mathfrak{M}$, we instead implement the perturbed measurement protocol $\mathfrak{M}(\epsilon_\circ)$ defined below.

Definition 5.4 (Perturbed measurement protocol). Given a measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ and a positive number $\epsilon_\circ > 0$, we define the **perturbed measurement protocol** $\mathfrak{M}(\epsilon_\circ)$ as the measurement protocol where the POVM

$$\left(\frac{E_1^{(i)} + \epsilon_\circ \mathbb{I}/M_i}{1+\epsilon_\circ}, \ldots, \frac{E_{M_i}^{(i)} + \epsilon_\circ \mathbb{I}/M_i}{1+\epsilon_\circ} \right) \tag{5.4}$$

is measured N_i times, for $i \in [L]$. □

Observe that for all $\epsilon_\circ > 0$, the outcome probabilities obtained by implementing $\mathfrak{M}(\epsilon_\circ)$ is strictly positive for every state. Moreover, for $\epsilon_\circ \ll 1$, these probabilities are very close to the probabilities obtained by implementing $\mathfrak{M}$. Furthermore, the perturbed measurement protocol is trivial to implement in an experiment for any given $\epsilon_\circ > 0$, either by randomly sampling POVMs from $\mathfrak{M}$ and measuring them, or by post-processing the measurement outcomes of $\mathfrak{M}$. For the random sampling strategy, for each $i \in [L]$ and $r \in [N_i]$, at the rth repetition of the ith POVM, we sample the POVM $\boldsymbol{E}^{(i)}$ with probability $1/(1+\epsilon_\circ)$ and measure it, or with probability $\epsilon_\circ/(1+\epsilon_\circ)$, we (uniformly) randomly choose a number in $[M_i]$ and output it. This randomized strategy implements the measurement protocol $\mathfrak{M}(\epsilon_\circ)$. If the measurement protocol $\mathfrak{M}$ has already been implemented in an experiment, we can post-process the observed outcomes to obtain outcomes from $\mathfrak{M}(\epsilon_\circ)$ as follows. For each $i \in [L]$ and $r \in [N_i]$, if we observe the outcome $o_r^{(i)} \in [M_i]$ after the rth repetition of the

ith POVM $\boldsymbol{E}^{(i)}$, we output $o_r^{(i)}$ with probability $1/(1+\epsilon_\circ)$, or we output a (uniformly) randomly chosen number in $[M_i]$ with probability $\epsilon_\circ/(1+\epsilon_\circ)$. Therefore, from a practical point of view, we don't lose much by assuming that we measure $\mathfrak{M}(\epsilon_\circ)$. However, from a theoretical point of view, it is important to know that the theoretical guarantees for the estimator as well as the optimality results can be derived for $\mathfrak{M}$ instead of $\mathfrak{M}(\epsilon_\circ)$. We will show in Ch. 7 that all our guarantees and optimality results are valid for $\mathfrak{M}$ by choosing a sufficiently small $\epsilon_\circ > 0$. It is therefore sufficient to obtain results in this chapter for any given $\epsilon_\circ > 0$.

In the remainder of this section, we formally map the problem of learning the expectation value of an observable to the statistical problem studied in Sec. 4.1. For readers who wish to skip the details, we provide a quick summary of this mapping in Tab. 2. The first quantity of interest is the set of states $\mathscr{X}$. In the statistical problem, this is a compact and convex subset of $\mathbb{R}^D$. On the other hand, in the quantum case, the set of quantum states, while compact and convex, is a subset of $\mathbb{C}^{d\times d}$. This, however, is not a problem, since we can construct an *isometric isomorphism* from the set $\mathbb{S}_d$ of Hermitian matrices in $\mathbb{C}^{d\times d}$ to $\mathbb{R}^{d^2}$ (i.e., $D = d^2$ in the quantum case). The image of compact and convex sets under a linear map are compact and convex. It can be verified that $\mathscr{I}\colon \mathbb{S}_d \to \mathbb{R}^{d^2}$ defined as $\mathscr{I}(P) = ((P_{ii})_{1\le i\le d}, (\sqrt{2}\ \mathrm{Re}(P_{ij}))_{1\le i<j\le d}, (\sqrt{2}\ \mathrm{Im}(P_{ij}))_{1\le i<j\le d})$ for $P \in \mathbb{S}_d$ an isometric isomorphism, where the notation $(P_{ij})_{1\le i\le d}$ means $P_{11}, \cdots, P_{dd}$ and so forth. Since all the matrices in $\mathbb{S}_d$ appear in our construction only through inner products, we can directly work with $\mathbb{S}_d$ instead of $\mathbb{R}^{d^2}$. Thus, we omit the map $\mathscr{I}$ in the construction for brevity.

We take the vector g in Sec. 4.1 to be the observable $\mathscr{O} \in \mathbb{S}_d$. Thus, our goal is to estimate $\langle g, x_{\mathrm{true}}\rangle = \mathrm{Tr}(\mathscr{O}\rho)$ in the quantum scenario. For this purpose, we need to incorporate data obtained from an experiment. Recall that this is done in Sec. 4.1 by considering random variables $\mathrm{Z}^{(i)}$ for $i \in [L]$, taking values in a Polish space $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$. The random variable $\mathrm{Z}^{(i)}$ is assumed to have a density $p^{(i)}_{\mathrm{A}^{(i)}(x_{\mathrm{true}})}$ with respect to a σ-finite reference measure $\boldsymbol{m}^{(i)}$ on $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$. The probability densities $p^{(i)}_\mu$ are parameterized by $\mu \in \mathscr{M}^{(i)}$, where $\mathscr{M}^{(i)}$ is a relatively open convex set in a finite-dimensional space, and $\mathrm{A}^{(i)}\colon \mathscr{X} \to \mathscr{M}^{(i)}$ is an affine map that determines the parameter given the true state x_{true}.

In the quantum scenario, our data corresponds to measurement outcomes. Upon measuring the ith POVM $\boldsymbol{E}^{(i)}$, we observe an outcome in the set $[M_i]$. Thus, for each $i \in [L]$, we define $\Omega^{(i)} = [M_i]$, and equip $\Omega^{(i)}$ with the discrete topology $2^{\Omega^{(i)}}$. It can be verified that $\Omega^{(i)}$ is a separable complete metric space (and thus a Polish space), where the underlying metric is the discrete metric $\mathscr{d}(i,j) = 0$ iff $i = j$. The Borel σ-algebra generated by the discrete topology is also discrete. We equip $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$ with the counting measure $\mathscr{m}^{(i)}$, i.e., $\mathscr{m}^{(i)}(E) = |E|$ for any $E \subseteq \Omega^{(i)}$. The probability density functions with respect to $\mathscr{m}^{(i)}$ are just the discrete distributions on $\Omega^{(i)}$. Since the probabilities must be strictly positive, we take $\mathscr{M}^{(i)} = \{x \in \mathbb{R}^{M_i} \mid (\forall i) x_i > 0,\ \sum_i x_i = 1\}$ to be the relatively open simplex in M_i dimensions, and map $\mu \in \mathscr{M}^{(i)}$ to the density function $p^{(i)}_\mu = (\mu_1, \ldots, \mu_{M_i})$. Since the density function is just a discrete distributions over M_i symbols, we view it as a vector in the standard simplex Δ_{M_i}.

Next, we need to construct an affine map $\mathrm{A}^{(i)}$ that maps the state $\rho \in \mathscr{X}$ to a parameter $\mathrm{A}^{(i)}(\rho) \in \mathscr{M}^{(i)}$, which in turn determines the distribution $p^{(i)}_{\mathrm{A}^{(i)}(\rho)}$. Since this distribution determines the probability of measurement outcomes given a state, this is given by Born's rule. Therefore, given the parameter $\epsilon_\circ > 0$ (for the perturbed measurement protocol), we define the affine map $\mathrm{A}^{(i)}$ as

$$\mathrm{A}^{(i)}(\rho) = \left(\frac{\mathrm{Tr}(E^{(i)}_1 \rho) + \epsilon_\circ / M_i}{1 + \epsilon_\circ}, \cdots, \frac{\mathrm{Tr}(E^{(i)}_{M_i} \rho) + \epsilon_\circ / M_i}{1 + \epsilon_\circ} \right) \tag{5.5}$$

on $\mathscr{X}$.

Next, we need to choose a set $\mathscr{F}^{(i)}$ of affine estimators. In Sec. 4.1, this is a finite-dimensional space of real-valued measurable functions on $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$ that contains constant functions. Since $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$ is a discrete space, every function on $\Omega^{(i)}$ is measurable. Moreover, since $\Omega^{(i)}$ is a finite set, real-valued functions on $\Omega^{(i)}$ can be viewed as M_i-dimensional real vectors. This is because any function $\phi^{(i)}$ on $\Omega^{(i)}$ can be identified with the vector $(\phi^{(i)}(1), \ldots, \phi^{(i)}(M_i))$. Thus, we choose $\mathscr{F}^{(i)} \cong \mathbb{R}^{M_i}$ to be the set of all functions on $\Omega^{(i)}$, or equivalently, all M_i-dimensional vectors.

We verify that each $(\mathscr{D}^{(i)}, \mathscr{F}^{(i)})$ is a good pair, by checking that it satisfies the requirements of Def. 4.1. The first two conditions of Def. 4.1 hold by construction, and the third condition holds

because $\mathscr{F}^{(i)}$ is the set of all functions on $\Omega^{(i)}$. The last condition holds because for $\phi^{(i)} \in \mathscr{F}^{(i)}$, we have $F_{\phi^{(i)}}(\mu) = \log(\sum_{k=1}^{M_i} \exp(\phi_k^{(i)})\mu_k)$, which is a concave function of μ. We have thus verified that $(\mathscr{D}^{(i)}, \mathscr{F}^{(i)})$ is a good pair for $i \in [L]$. At this point, we have a set of estimators for the ith POVM for each $i \in [L]$. According to the measurement protocol $\mathfrak{M}$, for each $i \in [L]$, the ith POVM is measured N_i times. Thus, we need a way to construct an estimator that accounts for multiple measurement outcomes for every POVM measurement. For this purpose, we use the direct product of good pairs (Def. 4.4) to construct a large space that accounts for all the measurement outcomes.

The large space Ω is given by $\Omega = \prod_{i=1}^{L} (\Omega^{(i)})^{N_i}$, which we equip with the Borel σ-algebra $\mathscr{B}(\Omega)$. The reference measure on $(\Omega, \mathscr{B}(\Omega))$ is the product measure $m = \prod_{i=1}^{L} (m^{(i)})^{N_i}$. The set of parameters is given by $\mathscr{M} = (\mathscr{M}_1)^{N_1} \times \cdots \times (\mathscr{M}_L)^{N_L}$. The affine map $\mathrm{A} \colon \mathscr{X} \to \mathscr{M}$ is given by the direct sum $\mathrm{A} = \oplus_{i=1}^{L} \oplus_{r=1}^{N_i} \mathrm{A}^{(i)}$. Specifically, the ith map $\mathrm{A}^{(i)}$ is repeated N_i times to incorporate N_i outcomes for the ith POVM. The set of affine estimators $\mathscr{F}$ on the large space is defined as all the estimators of the form $\phi = \sum_{i=1}^{L} \sum_{r=1}^{N_i} \phi^{(i,r)}$, where $\phi^{(i,r)} \in \mathscr{F}^{(i)}$ for all $r \in [N_i]$ and acts on the rth outcome of the ith POVM. However, for a fixed i, since we receive many outcomes from the same distribution, by [62, Rem. 3.2], it suffices to take $\phi^{(r,i)} = \phi^{(i)}$ for all $r \in [N_i]$ and all $i \in [L]$. Therefore, given outcomes $\omega_1^{(i)}, \ldots, \omega_{N_i}^{(i)} \in \Omega^{(i)}$ sampled according to the distribution $p^{(i)}_{\mathrm{A}^{(i)}(\mathscr{I}(\rho))}$ for $i \in [L]$, the estimate computed by ϕ using these outcomes is $\phi(\omega_1^{(1)}, \ldots, \omega_{N_L}^{(L)}) = \sum_{i=1}^{L} \sum_{r=1}^{N_i} \phi^{(i)}(\omega_r^{(i)})$.

For convenience of the reader, we summarize the mappings defined in this section in Tab. 2.

5.2 Estimation procedure

In this section, we introduce $\mathbb{T}$he $\mathbb{O}$ptimal $\mathbb{O}$bservable expectation value $\mathbb{L}$earner ($\mathbb{TOOL}$) for estimating the expectation value of a given observable using the outcomes of the specified measurement protocol. Formally, $\mathbb{TOOL}$ consists of two parts: (I) a procedure for constructing the estimator and estimation error (Box 3), and (II) a procedure for estimating expectation value from measurement outcomes (Box 4). In part (I), $\mathbb{TOOL}$ takes the observable $\mathscr{O}$, the measurement protocol $\mathfrak{M}$, the confidence level $1 - \delta$, and the parameter $\epsilon_\circ > 0$ as input, and constructs an estimator $\widehat{\mathscr{O}}_*$ and the associated estimation error ε_*. This construction is done by solving a convex optimization

$\mathscr{X}$	Set of density matrices
$\Omega^{(i)}$	Measurement labels $\{1, \ldots, N_i\}$
$m^{(i)}$	Counting measure on $(\Omega^{(i)}, \mathscr{B}(\Omega^{(i)}))$ with $\mathscr{B}(\Omega^{(i)}) = 2^{\Omega^{(i)}}$
$\mathscr{M}^{(i)}$	Relatively open simplex $\{x \in \mathbb{R}^{M_i} \mid (\forall i)\ x_i > 0,\ \sum_i x_i = 1\}$
p_μ	$p_\mu = (\mu_1, \ldots, \mu_{M_i})$, $\mu \in \mathscr{M}^{(i)}$
$\mathrm{A}^{(i)}$	$(\mathrm{A}^{(i)}(\rho))_k = \frac{\mathrm{Tr}(E_k^{(i)}\rho)+\epsilon_\circ/M_i}{1+\epsilon_\circ}$, $k \in [M_i]$, $\epsilon_\circ > 0$
$\mathscr{F}^{(i)}$	Real-valued functions on $\Omega^{(i)}$, identified with M_i-dimensional real vectors
g	Observable $\mathscr{O}$

Table 2: A dictionary mapping the quantities for the statistical problem given in Sec. 4.1 to the corresponding quantities for the quantum problem of estimating expectation values. The index i varies from 1 to L, where L denotes the number of measurement settings.

problem (see Sec. 5.4), and is usually the most computationally intensive part of the estimation procedure. The estimator $\widehat{\mathscr{O}}_*$ and the estimation error ε_* can be stored classically, and reused as many times as necessary for the same input configuration (observable $\mathscr{O}$, measurement protocol $\mathfrak{M}$, confidence level $1-\delta$, and parameter $\epsilon_\circ$). Importantly, the construction of the estimator and estimation error in part (I) does not depend on the measurement outcomes, and therefore, it can be done either before or after the measurements are performed in an experiment. In part (II), we discuss how to use the estimator $\widehat{\mathscr{O}}_*$ to estimate the expectation value of $\mathscr{O}$ given the measurement outcomes obtained after implementing $\mathfrak{M}$ in an experiment. The estimator $\widehat{\mathscr{O}}_*$ can compute the estimate efficiently from the data, and runs very fast in practice. Since part (II) is fairly straightforward, we will use $\mathbb{TOOL}$ synonymously with part (I) in practice.

We now discuss part (I), which is construction of the estimator and estimation error. This estimator construction is performed by adapting the results of Sec. 4.3 to the problem of learning expectation value of an observable. Note that the construction we describe below differs from the construction given in [92], in that we allow $\alpha_* = 0$ in Box 3.

Box 3: TOOL estimator construction

Input: Observable $\mathcal{O}$, measurement protocol $\mathfrak{M}$, confidence level $1-\delta \in (0,1)$, parameter $0 < \epsilon_\circ \ll 1$

Estimator construction:

(1) Find $\alpha_* \geq 0$ that achieves the minimum in

$$\varepsilon_* = \min_{\alpha \geq 0} \left[\frac{\alpha}{N} \log\left(\frac{2}{\delta}\right) + \max_{\chi_1,\chi_2 \in \mathscr{X}} \left(\frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) - \alpha \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) \right) \right], \tag{5.6}$$

and $\chi_1^*, \chi_2^* \in \mathscr{X}$ that achieve the maximum in

$$\max_{\chi_1,\chi_2 \in \mathscr{X}} \left(\frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) - \alpha_* \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) \right). \tag{5.7}$$

Here, $N = \sum_{i=1}^{L} N_i$ is the total number of samples used by $\mathfrak{M}$, $\mathfrak{M}(\epsilon_\circ)$ is the perturbed measurement protocol defined in Def. 5.4, and $\mathrm{BC}_{\mathfrak{M}(\epsilon_\circ)}$ is the average Bhattacharyya distance defined in Def. 3.4. See Eq. (5.10) for an explicit expression for $\mathrm{BC}_{\mathfrak{M}(\epsilon_\circ)}$.

(2) For $i \in [L]$, set

$$\phi_*^{(i)}(k) = \frac{\alpha_*}{2} \log\left(\frac{\mathrm{Tr}(E_k^{(i)}\chi_1^*) + \epsilon_\circ/M_i}{\mathrm{Tr}(E_k^{(i)}\chi_2^*) + \epsilon_\circ/M_i} \right), \tag{5.8}$$

where $k \in [M_i]$.

(3) Define the estimator $\widehat{\mathcal{O}}_*$ to be the function

$$\widehat{\mathcal{O}}_* = \frac{1}{N} \sum_{i=1}^{L} \sum_{r=1}^{N_i} \phi_*^{(i)} + \frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1^*) + \mathrm{Tr}(\mathcal{O}\chi_2^*)\right), \tag{5.9}$$

where the rth copy of $\phi_*^{(i)}$ accepts the outcome observed in the rth repetition of the ith POVM as input, for $r \in [N_i]$ and $i \in [L]$.

Output: estimator $\widehat{\mathcal{O}}_*$, estimation error ε_*

In practice, we can store the elements $\phi^{(i)}(k)$ for $k \in [M_i]$ and $i \in [L]$, the constant $(\mathrm{Tr}(\mathcal{O}\chi_1^*) +$

$\mathrm{Tr}(\mathcal{O}\chi_2^*))/2$, and the estimation error ε_* computed in Box 3 in memory for future use. This takes at most $O(M)$ memory, where $M = \sum_{i=1}^{L}$ is the total number of POVM elements. While it may not be obvious from Eq. (5.9), $\widehat{\mathcal{O}}_*$ is actually an affine function of the observed frequencies, which we show in Prop. 5.12. Importantly, the estimator $\widehat{\mathcal{O}}_*$ and the estimation error ε_* satisfy the rigorous guarantee that for all states ρ, the true expectation value $\mathrm{Tr}(\mathcal{O}\rho)$ lies within an error of ε_* to the estimate with confidence level of $1-\delta$. To show this, we need the following result connecting the average Bhattacharyya distance between two states determined by the measurement protocol $\mathfrak{M}(\epsilon_\circ)$ (defined in Def. 3.4), and the Bhattacharyya distance (defined in Eq. (4.6)) between the parameters defined by the mapping in Tab. 2.

Lemma 5.5. *For the mapping given in Tab. 2, the Bhattacharyya distance between the parameters* $\mathrm{A}(\chi_1)$ *and* $\mathrm{A}(\chi_2)$ *for states* $\chi_1, \chi_2 \in \mathscr{X}$ *is given by*

$$\begin{aligned}\mathrm{BD}(\mathrm{A}(\chi_1), \mathrm{A}(\chi_2)) &= N\ \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1, \chi_2) \\ &= -\sum_{i=1}^{L} N_i \log\left(\sum_{k=1}^{M_i} \sqrt{\left(\frac{\mathrm{Tr}(E_k^{(i)}\chi_1) + \epsilon_\circ/M_i}{1+\epsilon_\circ}\right)\left(\frac{\mathrm{Tr}(E_k^{(i)}\chi_2) + \epsilon_\circ/M_i}{1+\epsilon_\circ}\right)}\right).\end{aligned} \tag{5.10}$$

Furthermore, $\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1, \chi_2)$ *is well-defined, continuous, and jointly convex for all* $\chi_1, \chi_2 \in \mathscr{X}$.

Proof. We use the definitions in Tab. 2 in this proof. For $i \in [L]$, we have $\mathrm{BC}(\mu^{(i)}, \nu^{(i)}) = \sum_{k=1}^{M_i} \sqrt{\mu_k^{(i)}\nu_k^{(i)}}$, where $\mu^{(i)}, \nu^{(i)}$ are elements of the relatively open standard simplex $\mathscr{M}^{(i)}$. The map A is given by $\mathrm{A} = \oplus_{i=1}^{L} \oplus_{r=1}^{N_i} \mathrm{A}^{(i)}$, where $\mathrm{A}^{(i)} \colon \mathscr{X} \to \mathscr{M}^{(i)}$. Then, by multiplicativity of the Bhattacharyya coefficient (Lem. 4.6), for all states $\chi_1, \chi_2 \in \mathscr{X}$, we have

$$\mathrm{BC}(\mathrm{A}(\chi_1), \mathrm{A}(\chi_2)) = \prod_{i=1}^{L}\left[\sum_{k=1}^{M_i} \sqrt{\left(\frac{\mathrm{Tr}(E_k^{(i)}\chi_1) + \epsilon_\circ/M_i}{1+\epsilon_\circ}\right)\left(\frac{\mathrm{Tr}(E_k^{(i)}\chi_2) + \epsilon_\circ/M_i}{1+\epsilon_\circ}\right)}\right]^{N_i}. \tag{5.11}$$

Since

$$\frac{\mathrm{Tr}(E_k^{(i)}\chi_1) + \epsilon_\circ/M_i}{1+\epsilon_\circ} = \mathrm{Tr}\left(\frac{E_k^{(i)} + \epsilon_\circ/M_i\mathbb{I}}{1+\epsilon_\circ}\chi_1\right), \tag{5.12}$$

and $\mathrm{BD} = -\log(\mathrm{BC})$ by definition, we obtain Eq. (5.10) from the above equations and Def. 3.4. From

Prop. 4.10.1, we have that $-\log(\mathrm{BC}(\mathrm{A}(\chi_1), \mathrm{A}(\chi_2)))$ is well-defined, continuous and jointly convex, from which it follows that $\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1, \chi_2)$ is well-defined, continuous, and jointly convex. □

The estimator and estimation error constructed by 𝕋𝕆𝕆𝕃 satisfies the rigorous guarantee noted below.

Proposition 5.6. *The estimator $\widehat{\mathcal{O}}_*$ and error ε_* constructed by 𝕋𝕆𝕆𝕃 for learning the expectation value of $\mathcal{O}$ using outcomes of $\mathfrak{M}$ satisfy*

$$\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}\left(|\widehat{\mathcal{O}}_* - \mathrm{Tr}(\mathcal{O}\rho)| \leq \varepsilon_*\right) \geq 1 - \delta \tag{5.13}$$

for all $\rho \in \mathscr{X}$, where $\epsilon_\circ > 0$ is the parameter used in the construction.

Proof. From Thm. 4.14, the optimal points $\alpha'_* \geq 0$ and $x^*, y^* \in \mathscr{X}$ of

$$2\Phi_*(\log(2/\delta)) = \min_{\alpha' \geq 0}\left[2\alpha' \log(2/\delta) + \max_{x,y \in \mathscr{X}}\left(\langle g, x\rangle - \langle g, y\rangle + 2\alpha' \log(\mathrm{BC}(\mathrm{A}(x), \mathrm{A}(y)))\right)\right] \tag{5.14}$$

can be used to construct the estimator $\widehat{g}_* = \phi'_* + (\langle g, x^*\rangle + \langle g, y^*\rangle)/2$, where $\phi'_* = (\alpha'_*/2)\log(p_{\mathrm{A}(x^*)}/p_{\mathrm{A}(y^*)})$. This estimator satisfies the guarantee $\mathbb{P}_{\mathrm{A}(x_{\mathrm{true}})}(|\widehat{g}_* - \langle g, x_{\mathrm{true}}\rangle| \leq \Phi_*(\log(2/\delta))) \geq 1 - \delta$ for all $x_{\mathrm{true}} \in \mathscr{X}$. To derive Eq. (5.13) from this guarantee, we use the mapping given in Tab. 2, and the fact that $\mathrm{BD}(\mathrm{A}(\chi_1), \mathrm{A}(\chi_2)) = N\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1, \chi_2)$ (Lem. 5.5). Additionally, we identify $\Phi_*(\log(2/\delta))$ with ε_*, map α' to α/N in Eq. (5.6), and map ϕ'_* to ϕ_*/N in Eq. (5.8). Then, the estimator $\widehat{g}_*$ becomes $\widehat{\mathcal{O}}_*$ in Eq. (5.9) under these mappings. □

Now, we move on to part (II), where we show how to use the estimator constructed in Box 3 on experimental data.

Box 4: 𝕋𝕆𝕆𝕃 estimation procedure

Input: Observable $\mathcal{O}$, measurement protocol $\mathfrak{M}$, confidence level $1 - \delta \in (0.75, 1)$, parameter $0 < \epsilon_\circ \ll 1$, outcomes $\boldsymbol{o} = (o_1^{(1)}, \ldots, o_{N_1}^{(1)}, \ldots, o_1^{(L)}, \ldots, o_{N_L}^{(L)})$ of $\mathfrak{M}(\epsilon_\circ)$

Estimation procedure:

(1) If the estimator $\widehat{\mathcal{O}}_*$ and the error ε_* in Box 3 have been pre-computed for the input configuration $(\mathcal{O}, \mathfrak{M}, 1-\delta, \epsilon_\circ)$, then proceed to (2). Else, compute $\widehat{\mathcal{O}}_*$ and ε_* according to Box 3.

(2) Compute the estimate for $\langle \mathcal{O} \rangle$ using the outcomes $\boldsymbol{o}$ observed in the experiment as

$$\widehat{\mathcal{O}}_*(\boldsymbol{o}) = \frac{1}{N} \sum_{i=1}^{L} \sum_{r=1}^{N_i} \phi_*^{(i)}(o_r^{(i)}) + \frac{1}{2} \left(\mathrm{Tr}(\mathcal{O}\chi_1^*) + \mathrm{Tr}(\mathcal{O}\chi_2^*) \right). \tag{5.15}$$

Here, $o_r^{(i)} \in [M_i]$ denotes the outcome observed in the rth repetition of the ith POVM, for $r \in [N_i]$ and $i \in [L]$.

Output: estimate $\widehat{\mathcal{O}}_*(\boldsymbol{o})$, estimation error ε_*

Step (2) of Box 4 requires $O(N)$ time to implement on a computer, where $N = \sum_{i=1}^{L} N_i$ is the total number of samples, since the entries of $\phi_*^{(i)}$ and the constant term in $\widehat{\mathcal{O}}_*$ are computed before step (2) in Box 4. In practice, we see that computing an estimate from the observed outcomes is very fast.

5.3 Properties of the estimator

In this section, we prove some properties for the estimator $\widehat{\mathcal{O}}_*$ and the estimation error ε_* constructed by $\mathbb{TOOL}$. In Sec. 5.3.1, we use the result of [62] to show that the estimator constructed by $\mathbb{TOOL}$ is minimax optimal to a constant factor under the premise of Sec. 5.1. This shows that the estimation error ε_* is an important quantity to study in its own right. We therefore derive alternate expressions for ε_* and bounds on it. In particular, we show that under certain conditions, ε_* is related to an f-divergence. In Sec. 5.3.2, we show that the estimator $\widehat{\mathcal{O}}_*$ can be expressed as an affine estimator of the observed frequencies. Subsequently, we compute the bias of this estimator.

5.3.1 Minimax optimality

We begin by showing that the estimator $\widehat{\mathcal{O}}_*$ constructed by $\mathbb{TOOL}$ is minimax optimal to a constant factor in the following sense.

Proposition 5.7. *The estimation error ε_* of the estimator $\widehat{\mathcal{O}}_*$ constructed by $\mathbb{TOOL}$ with parameter $\epsilon_\circ > 0$ to learn the expectation value an observable $\mathcal{O}$ using the outcomes of $\mathfrak{M}$ to a confidence level of $1-\delta \in (0.75, 1)$ satisfies*

$$\mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta) \leq \varepsilon_* \leq \frac{2\log(2/\delta)}{\log(1/(4\delta))} \mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta), \tag{5.16}$$

where $\mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta)$ is the minimax optimal risk defined in Eq. (5.2).

Proof. This follows from Thm. 4.8 under the mappings defined in Tab. 2 and in the proof of Prop. 5.6. □

As in the case of the theoretical guarantee derived in Prop. 5.6, the above optimality result applies when the measurement outcomes are obtained from the perturbed measurement protocol $\mathfrak{M}(\epsilon_\circ)$. We prove in Thm. 7.13 that we get optimality guarantees for $\mathbb{TOOL}$ even for the case of $\epsilon_\circ = 0$. Observe that

$$2 < \frac{2\log(2/\delta)}{\log(1/(4\delta))} < 6.54 \tag{5.17}$$

for confidence levels greater than 90%. Thus, the multiplicative factor in Eq. (5.16) is small for large enough confidence levels.

The estimation error ε_* is therefore an important quantity that needs to be studied in its own right, as it can help understand the optimal performance for learning the expectation values of observables. While a detailed study of ε_* is postponed to Ch. 7, where we relate ε_* to the minimax norm, we note down some additional expressions for ε_* in this section. We begin with the following expression for ε_*, which is helpful in computing it analytically as well as numerically. Additionally, we show that solving the optimization problem in Eq. (5.18) is sufficient to compute the estimator $\widehat{\mathcal{O}}_*$ in Box 3.

Proposition 5.8. *The estimation error ε_* of the estimator constructed by* TOOL *with parameter $\epsilon_\circ > 0$, for learning the expectation value of $\mathcal{O}$ using outcomes of $\mathfrak{M}$, can be expressed as*

$$\begin{aligned} \varepsilon_* = \max_{\chi_1,\chi_2\in\mathscr{X}} \quad & \frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) \\ \text{s.t.} \quad & \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) \leq \frac{1}{N}\log\left(\frac{2}{\delta}\right). \end{aligned} \tag{5.18}$$

Furthermore, if $\chi_1^*, \chi_2^* \in \mathscr{X}$ are the points achieving the maximum in Eq. (5.18), and $\alpha_* \geq 0$ denotes the optimal value of the dual variable for the constraint $\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) \leq (1/N)\log(2/\delta)$ in Eq. (5.18), then $(\chi_1^*, \chi_2^*; \alpha_*)$ can be used to construct the estimator $\widehat{\mathcal{O}}_*$ in Box 3.

Proof. It follows from Prop. 4.9 that

$$\begin{aligned} \varepsilon_* = \max_{\chi_1,\chi_2\in\mathscr{X}} \quad & \frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) \\ \text{s.t.} \quad & \mathrm{BC}(\mathrm{A}(\chi_1),\mathrm{A}(\chi_2)) \geq \left(\frac{\delta}{2}\right), \end{aligned} \tag{5.19}$$

where we use the fact that ε_* is equal to $\Phi_*(\log(2/\delta))$. Since $\mathrm{BD}(\mathrm{A}(\chi_1),\mathrm{A}(\chi_2)) = -\log(\mathrm{BC}(\mathrm{A}(\chi_1),\mathrm{A}(\chi_2)))$, Eq. (5.18) follows from Lem. 5.5.

Next, from Prop. 4.11.4, it follows that $(\chi_1^*, \chi_2^*; \alpha_*)$ as defined in the statement of Prop. 5.8 is a saddle point (maximum in $\chi_1, \chi_2 \in \mathscr{X}$ and minimum in $\alpha \geq 0$) of the function

$$\Phi_{\log(2/\delta)}(\chi_1,\chi_2;\alpha) = \frac{\alpha}{N}\log\left(\frac{2}{\delta}\right) + \max_{\chi_1,\chi_2\in\mathscr{X}}\left(\frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) - \alpha\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2)\right). \tag{5.20}$$

Note that by Prop. 4.11.3, a saddle point for the function $\Phi_{\log(2/\delta)}(\chi_1,\chi_2;\alpha)$ always exists. Moreover, $\alpha_* \geq 0$ is unique (Prop. 4.11.2). It is shown in Prop. 4.10.3 that Eq. (5.6) is the dual optimization problem of Eq. (5.18), which shows that α_* is the dual optimal. Since $(\chi_1^*, \chi_2^*; \alpha_*)$ is a saddle point of $\Phi_{\log(2/\delta)}(\chi_1,\chi_2;\alpha)$, α_* achieves the minimum in Eq. (5.18), and $\chi_1^*, \chi_2^* \in \mathscr{X}$ achieve the maximum in $\max_{\chi_1,\chi_2\in\mathscr{X}} \Phi_{\log(2/\delta)}(\chi_1,\chi_2;\alpha_*)$. It follows that $(\chi_1^*, \chi_2^*; \alpha_*)$ can be used for constructing the estimator in Box 3. □

We now give another expression for the estimation error in terms of the saddle points of Eq. (5.6). This expression shows that, under some conditions, the error ε_* is proportional to an f-divergence. We first define this f-divergence below.

Definition 5.9. Given probability distributions p, q on M symbols, define

$$\mathfrak{d}(p,q) = \sum_{i \in \operatorname{supp} p} \frac{(p_i - q_i)^2}{2\sqrt{p_i q_i}}. \tag{5.21}$$

if p, q have the same support, and ∞ otherwise. □

We can, in fact, extend the definition of $\mathfrak{d}$ to an f-divergence between arbitrary probability distributions, which we show below.

Proposition 5.10. *The function $f\colon (0,\infty) \to \mathbb{R}$ given by*

$$f(x) = \frac{(x-1)^2}{2\sqrt{x}} \tag{5.22}$$

defines an f-divergence $\mathfrak{d}(\mathbb{P},\mathbb{Q}) = D_f(\mathbb{P},\mathbb{Q})$ between probability distributions $\mathbb{P},\mathbb{Q}$ on a measurable space according to Eq. (2.3). $\mathfrak{d}$ is symmetric in its arguments. For discrete probability distributions p, q on M symbols, $\mathfrak{d}(p,q)$ is equal to Eq. (5.21).

Proof. We can write

$$f(x) = \frac{1}{2\sqrt{x}} - \sqrt{x} + \frac{1}{2}x^{3/2}. \tag{5.23}$$

Since $1/\sqrt{x}$, $-\sqrt{x}$, and $x^{3/2}$ are convex on $(0,\infty)$, f is convex. We also have $f(1) = 0$. Thus, $\mathfrak{d} \equiv D_f$ as defined in Eq. (2.3) is an f-divergence. $\mathfrak{d}$ is symmetric because $xf(1/x) = f(x)$ for all $x \in (0,\infty)$ [82, Rem. (7.3)].

It remains to verify that we obtain Eq. (5.21) for discrete distributions. For that purpose, observe that $f'(\infty) = \lim_{x\to 0} xf(1/x) = \infty$. Thus, if p is not absolutely continuous with respect to q, then $\mathfrak{d}(p,q) = \infty$. Since p, q are discrete, p is absolute continuous with respect to q iff the support of p is contained in the support of q. Since $\mathfrak{d}(q,p) = \mathfrak{d}(p,q)$, it follows that $\mathfrak{d}(p,q) < \infty$ only when

the supports of p and q are equal. □

The f-divergence $\mathfrak{d}$ is closely related to the χ^2-divergence. The χ^2-divergence (or Pearson divergence) is an f-divergence defined by $f_P(x) = (x-1)^2/2$ according to Def. 2.5, while the reverse χ^2-divergence (or the Neyman divergence) is an f-divergence defined by $f_N = (x-1)^2/2x$ (see [94, Sec. (2.3)]). The χ^2-divergence is an important quantity in statistics, as it bounds the performance of statistical methods for hypothesis testing and estimation (see, for example, [21], where the χ^2-divergence is relevant for estimating the expectation values of Pauli observables). $\mathfrak{d}$ defined in Eq. (5.21) is a symmetrized version of the χ^2-divergence, obtained by taking the geometric mean of f_P and f_N, since $f(x) = \sqrt{f_P(x) f_N(x)}$ for all $x \in (0, \infty)$.

Below, we give bounds on the estimation error ε_* in terms of the optimal points χ_1^*, χ_2^* and α_* of Eq. (5.6). When these optimal points satisfy some conditions, we can show that the error can be expressed in terms of $\mathfrak{d}$.

Proposition 5.11. *Let $\chi_1^*, \chi_2^* \in \mathscr{X}$ and $\alpha_* \geq 0$ be the primal and dual optimal points respectively for the optimization in Eq. (5.18). For $i \in [L]$, let $\mathrm{A}^{(i)}$ be the linear map defined in Tab. 2. Then, the following statements hold.*

1. The error ε_* can be expressed as

$$\varepsilon_* = \frac{1}{2} \left(\mathrm{Tr}(\mathscr{O} \chi_1^*) - \mathrm{Tr}(\mathscr{O} \chi_2^*) \right). \tag{5.24}$$

2. The error ε_* satisfies the bound

$$0 \leq \alpha_* \sum_{i=1}^{L} \frac{N_i}{N} \frac{\mathfrak{d}(\mathrm{A}^{(i)}(\chi_1^*), \mathrm{A}^{(i)}(\chi_2^*))}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1^*), \mathrm{A}^{(i)}(\chi_2^*))} \leq \varepsilon_*. \tag{5.25}$$

3. If the points χ_1^*, χ_2^* are full rank and $\alpha_* > 0$, then we have

$$\varepsilon_* = \alpha_* \sum_{i=1}^{L} \frac{N_i}{N} \frac{\mathfrak{d}(\mathrm{A}^{(i)}(\chi_1^*), \mathrm{A}^{(i)}(\chi_2^*))}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1^*), \mathrm{A}^{(i)}(\chi_2^*))}. \tag{5.26}$$

Proof. 1. The estimation error ε_* corresponds to the saddle-point value $\Phi_*(\log(2/\delta))$ in Eq. (4.20). From Prop. 4.11.3, we know that Eq. (4.20) has a saddle point. Moreover, from the proof of Prop. 4.11.3, we know that this saddle point attains the maximum in Eq. (4.18). From these observations, we can infer that Eq. (5.6) has a saddle point $(\chi_1^*, \chi_2^*; \alpha_*)$, with $\chi_1^*, \chi_2^* \in \mathscr{X}$ and $\alpha_* \geq 0$. The equation $\varepsilon_* = (\mathrm{Tr}(\mathscr{O}\chi_1^*) - \mathrm{Tr}(\mathscr{O}\chi_2^*))/2$ then follows from Eq. (5.18).

2. From Eq. (5.18), we can infer that $\varepsilon_* \geq 0$ since any $\chi_1 = \chi_2$ satisfies the constraint of the optimization. Furthermore, $\mathfrak{d}(\mathrm{A}^{(i)}(\chi_1), \mathrm{A}^{(i)}(\chi_2)) < \infty$ and $\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1), \mathrm{A}^{(i)}(\chi_2)) > 0$ for all $i \in [L]$ and all $\chi_1, \chi_2 \in \mathscr{X}$, since $\mathrm{A}^{(i)}(\chi_1)$ lies inside the relatively open simplex. Thus, if $\alpha_* = 0$, then Eq. (5.25) holds trivially. We therefore take $\alpha_* > 0$ for the remainder of the proof.

From Eq. (5.6), we can write

$$\varepsilon_* = \left[\frac{\alpha_*}{N}\log\left(\frac{2}{\delta}\right) + \max_{\chi_1,\chi_2\in\mathscr{X}}\left(\frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2)\right) - \alpha_*\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2)\right)\right], \tag{5.27}$$

where χ_1^*, χ_2^* attain the maximum. Slater's condition holds for this optimization problem since it is convex and $\mathrm{relint}\,\mathscr{X} \neq \varnothing$. Thus, KKT conditions are necessary and sufficient for optimality (see Sec. 2.5 for details). In particular, the gradient of the Lagrangian with respect to the primal variables vanishes at the optimum. The Lagrangian for the optimization problem, after converting it to a minimization problem by changing signs of the objective function, can be written as

$$\begin{aligned}\mathscr{L}(\chi_1,\chi_2;\kappa_1,\kappa_2,\nu_1,\nu_2) = -\frac{1}{2}\mathrm{Tr}(\mathscr{O}\chi_1) + \frac{1}{2}\mathrm{Tr}(\mathscr{O}\chi_2) - 2\frac{\alpha_*}{N}\left(\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) + \log\left(\frac{2}{\delta}\right)\right)\\ + \nu_1(\mathrm{Tr}(\chi_1) - 1) + \nu_2(\mathrm{Tr}(\chi_2) - 1) - \mathrm{Tr}(\kappa_1\chi_1) - \mathrm{Tr}(\kappa_2\chi_2),\end{aligned} \tag{5.28}$$

where $\nu_1, \nu_2 \in \mathbb{R}$ are the dual variables for the constraints $\mathrm{Tr}(\chi_1) = 1$ and $\mathrm{Tr}(\chi_2) = 1$ respectively, and $\kappa_1, \kappa_2 \in \mathbb{S}_d$ are positive semi-definite matrices that are dual variables for the constraints $\chi_1 \geq 0$ and $\chi_2 \geq 0$ respectively. The gradient of $\mathscr{L}$ with respect to χ_1 and χ_2 is given by

$$\begin{aligned}\nabla_{\chi_1}\mathscr{L}(\chi_1,\chi_2;\kappa_1,\kappa_2,\nu_1,\nu_2) &= -\frac{1}{2}\mathscr{O} - 2\frac{\alpha_*}{N}\nabla_{\chi_1}\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) + \nu_1 - \kappa_1,\\ \nabla_{\chi_2}\mathscr{L}(\chi_1,\chi_2;\kappa_1,\kappa_2,\nu_1,\nu_2) &= \frac{1}{2}\mathscr{O} - 2\frac{\alpha_*}{N}\nabla_{\chi_2}\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) + \nu_2 - \kappa_2.\end{aligned} \tag{5.29}$$

To proceed, we evaluate the gradient of $\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}$ using Eq. (5.10). Denoting $\widetilde{E}_k^{(i)} = (E_k^{(i)} + \epsilon_\circ \mathbb{I}/M_i)/(1+\epsilon_\circ)$, we obtain

$$\begin{aligned}
\nabla_{\chi_1} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) &= \frac{1}{2}\sum_{i=1}^{L} N_i \sum_{k=1}^{M_i} \frac{\widetilde{E}_k^{(i)}}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1),\mathrm{A}^{(i)}(\chi_2))} \sqrt{\frac{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2)}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1)}}, \\
\nabla_{\chi_2} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2) &= \frac{1}{2}\sum_{i=1}^{L} N_i \sum_{k=1}^{M_i} \frac{\widetilde{E}_k^{(i)}}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1),\mathrm{A}^{(i)}(\chi_2))} \sqrt{\frac{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1)}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2)}}.
\end{aligned} \tag{5.30}$$

Using the definition of $\mathrm{A}^{(i)}$ in Tab. 2 and $N = \sum_{i=1}^{L} N_i$, we obtain

$$\begin{aligned}
\mathrm{Tr}(\chi_1 \nabla_{\chi_1} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2)) &= \frac{N}{2}, \\
\mathrm{Tr}(\chi_2 \nabla_{\chi_2} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2)) &= \frac{N}{2}.
\end{aligned} \tag{5.31}$$

Furthermore, we have

$$\begin{aligned}
&\frac{1}{N}\left(\mathrm{Tr}(\chi_2 \nabla_{\chi_1} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2)) + \mathrm{Tr}(\chi_1 \nabla_{\chi_2} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1,\chi_2))\right) - 1 \\
&= \frac{1}{2N}\sum_{i=1}^{L} \frac{N_i}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1),\mathrm{A}^{(i)}(\chi_2))} \sum_{k=1}^{M_i} \sqrt{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1)\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2)} \left(\frac{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2)}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1)} + \frac{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1)}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2)} - 2\right) \\
&= \frac{1}{2N}\sum_{i=1}^{L} \frac{N_i}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1),\mathrm{A}^{(i)}(\chi_2))} \sum_{k=1}^{M_i} \sqrt{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1)\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2)} \frac{(\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1) - \mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2))^2}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1)\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2)} \\
&= \sum_{i=1}^{L} \frac{N_i}{N} \frac{\mathfrak{d}(\mathrm{A}^{(i)}(\chi_1),\mathrm{A}^{(i)}(\chi_2))}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1),\mathrm{A}^{(i)}(\chi_2))}.
\end{aligned} \tag{5.32}$$

Denote the dual optimal points as $\kappa_1^*, \kappa_2^*, \nu_1^*, \nu_2^*$. Then, using the above gradient calculations, the fact that $\nabla_{\chi_1} \mathscr{L} = 0$ and $\nabla_{\chi_2} \mathscr{L} = 0$ at optimality, and $\mathrm{Tr}(\kappa_1^* \chi_1^*) = \mathrm{Tr}(\kappa_2^* \chi_2^*) = 0$ by complementary

slackness, we obtain

$$\begin{aligned}
0 &= \mathrm{Tr}(\chi_1^* \nabla_{\chi_1} \mathscr{L}(\chi_1^*, \chi_2^*)) = -\frac{1}{2}\mathrm{Tr}(\mathscr{O}\chi_1^*) - \alpha_* + \nu_1^* \\
0 &= \mathrm{Tr}(\chi_2^* \nabla_{\chi_2} \mathscr{L}(\chi_1^*, \chi_2^*)) = \frac{1}{2}\mathrm{Tr}(\mathscr{O}\chi_2^*) - \alpha_* + \nu_2^* \\
0 &= \mathrm{Tr}(\chi_2^* \nabla_{\chi_1} \mathscr{L}(\chi_1^*, \chi_2^*)) = -\frac{1}{2}\mathrm{Tr}(\mathscr{O}\chi_2^*) - 2\frac{\alpha_*}{N}\mathrm{Tr}(\chi_2^* \nabla_{\chi_1} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1^*, \chi_2^*)) + \nu_1^* - \mathrm{Tr}(\kappa_1^*\chi_2^*) \\
0 &= \mathrm{Tr}(\chi_1^* \nabla_{\chi_1} \mathscr{L}(\chi_1^*, \chi_2^*)) = \frac{1}{2}\mathrm{Tr}(\mathscr{O}\chi_1^*) - 2\frac{\alpha_*}{N}\mathrm{Tr}(\chi_1^* \nabla_{\chi_2} \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1^*, \chi_2^*)) + \nu_2^* - \mathrm{Tr}(\kappa_1^*\chi_1^*).
\end{aligned} \tag{5.33}$$

Solving for ν_1^*, ν_2^* from the first two equations, and rearranging the last two equations, we obtain

$$\alpha_* \sum_{i=1}^{L} \frac{N_i}{N} \frac{\mathfrak{d}(\mathrm{A}^{(i)}(\chi_1), \mathrm{A}^{(i)}(\chi_2))}{\mathrm{BC}(\mathrm{A}^{(i)}(\chi_1), \mathrm{A}^{(i)}(\chi_2))} = \varepsilon_* - \frac{1}{2}\left(\mathrm{Tr}(\kappa_1^*\chi_2^*) + \mathrm{Tr}(\kappa_2^*\chi_1^*)\right), \tag{5.34}$$

where we used the fact that $\varepsilon_* = (\mathrm{Tr}(\mathscr{O}\chi_1^*) - \mathrm{Tr}(\mathscr{O}\chi_2^*))/2$. Since $\mathrm{Tr}(\kappa_1^*\chi_2^*), \mathrm{Tr}(\kappa_2^*\chi_1^*) \geq 0$, we obtain Eq. (5.25).

3. By complementary slackness, we have $\mathrm{Tr}(\kappa_1^*\chi_1^*) = 0$ and $\mathrm{Tr}(\kappa_2^*\chi_2^*) = 0$. If χ_1^* and χ_2^* are full rank, then we must have $\kappa_1^* = \kappa_2^* = 0$ for complementary slackness to hold. When $\alpha_* > 0$, we also have Eq. (5.34). From these observations, we obtain Eq. (5.26). □

Numerically, we observe that the optimal points χ_1^*, χ_2^* are full rank when we measure all the Pauli operators, but not necessarily full rank when we measure only some Pauli operators. Motivated by this observation, we hypothesize that χ_1^*, χ_2^* are full rank when the implemented measurement protocol is informationally complete. Further investigation is necessary to determine whether or not this claim holds, and more generally, to characterize the conditions under which χ_1^*, χ_2^* are full rank.

5.3.2 Bias

We show in this section that the estimator $\widehat{\mathscr{O}}_*$ constructed by $\mathbb{TOOL}$ is affine in the observed frequencies. We begin by defining what we mean by observed frequencies. Suppose that upon implementing the measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^L$ in an experiment, we observed the

outcomes $o_r^{(i)} \in [M_i]$ for $r \in [N_i]$ and $i \in [L]$. Then, for each $i \in [L]$, the fraction of N_i outcomes that are equal to a particular label $k \in [M_i]$ is called the **observed frequency** $f_k^{(i)}$ of label k of the ith POVM. Mathematically, this can be expressed as

$$f_k^{(i)} = \frac{|\{r \in [N_i] \mid o_r^{(i)} = k\}|}{N_i}. \tag{5.35}$$

We denote the vector of observed frequencies for the ith POVM as

$$\boldsymbol{f}^{(i)} = (f_1^{(i)}, \ldots, f_{M_i}^{(i)}), \tag{5.36}$$

and the large vector containing the observed frequencies of all the POVMs is denoted

$$\boldsymbol{f} = (\boldsymbol{f}^{(1)}, \ldots, \boldsymbol{f}^{(L)}). \tag{5.37}$$

We show below that we can think of $\widehat{\mathscr{O}}_*$ as a function of $\boldsymbol{f}$ instead of the measurement outcomes, and show that it is affine in $\boldsymbol{f}$.

Proposition 5.12. *Let $\widehat{\mathscr{O}}_*$ be the estimator constructed by $\mathbb{TOOL}$ to learn the expectation value of $\mathscr{O}$ using outcomes of $\mathfrak{M}$. If $\boldsymbol{o} = (o_1^{(1)}, \ldots, o_{N_1}^{(1)}, \ldots, o_1^{(L)}, \ldots, o_{N_L}^{(L)})$ denotes the outcomes observed upon implementing $\mathfrak{M}$, and $\boldsymbol{f}$ denotes the corresponding vector of observed frequencies, then we have*

$$\widehat{\mathscr{O}}_*(\boldsymbol{o}) = \widehat{\mathscr{O}}_*(\boldsymbol{f}) = \sum_{i=1}^{L} \frac{N_i}{N} \left\langle \boldsymbol{\phi}_*^{(i)}, \boldsymbol{f}^{(i)} \right\rangle + \frac{1}{2} \left(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*) \right), \tag{5.38}$$

where we denote

$$\boldsymbol{\phi}_*^{(i)} = (\phi_*^{(i)}(1), \ldots, \phi_*^{(i)}(M_i)), \tag{5.39}$$

for $i \in [L]$.

Proof. For $i \in [L]$ and $k \in [M_i]$, denote $\boldsymbol{e}^{(i,k)}$ to be the M_i-dimensional vector with the kth component equal to 1 and zero elsewhere. Then, we have $\phi_*^{(i)}(k) = \left\langle \boldsymbol{\phi}_*^{(i)}, \boldsymbol{e}^{(i,k)} \right\rangle$. Consequently, the

estimator $\widehat{\mathscr{O}}_*$ in Eq. (5.9) can be written as

$$\widehat{\mathscr{O}}_*(\boldsymbol{o}) = \frac{1}{N}\sum_{i=1}^{L}\sum_{r=1}^{N_i}\left\langle \boldsymbol{\phi}_*^{(i)}, \boldsymbol{e}^{(i,o_r^{(i)})}\right\rangle + \frac{1}{2}(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*)). \tag{5.40}$$

Now, observe that we can write

$$\boldsymbol{f}^{(i)} = \frac{1}{N_i}\sum_{r=1}^{N_i}\boldsymbol{e}^{(i,o_r^{(i)})} \tag{5.41}$$

for all $i \in [L]$. From the above equations, we obtain Eq. (5.38). □

Using the above result, we can compute the bias of the estimator constructed by $\mathbb{TOOL}$.

Corollary 5.13. *If the true quantum state is ρ, then the bias of the estimator $\widehat{\mathscr{O}}_*$ constructed by $\mathbb{TOOL}$ with parameter $\epsilon_\circ > 0$ for learning the expectation value of $\mathscr{O}$ using outcomes of $\mathfrak{M}$ is equal to*

$$\mathrm{Tr}(\mathscr{O}\rho) - \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*)\right) + \frac{\alpha_*}{2N}\left(\mathrm{KL}(\mathbb{P}_{\mathfrak{M},\rho}\|\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\chi_1^*}) - \mathrm{KL}(\mathbb{P}_{\mathfrak{M},\rho}\|\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\chi_2^*})\right), \tag{5.42}$$

where $\mathfrak{M}(\epsilon_\circ)$ is the perturbed measurement protocol defined in Def. 5.4.

Proof. Note that for all $i \in [L]$, we have $\mathbb{E}[\boldsymbol{f}^{(i)}] = p_\rho^{(i)}$, where $(p_\rho^{(i)})_k = \mathrm{Tr}(E_k^{(i)}\rho)$ for $k \in [M_i]$. Then, denoting $\widetilde{E}_k^{(i)} = (E_k^{(i)} + \epsilon_\circ \mathbb{I}/M_i)/(1+\epsilon_\circ)$ for $i \in [L]$, from Eq. (5.38), we obtain

$$\begin{aligned}
\mathbb{E}[\widehat{\mathscr{O}}_*] &= \sum_{i=1}^{L}\frac{N_i}{N}\left\langle \boldsymbol{\phi}_*^{(i)}, p_\rho^{(i)}\right\rangle + \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*)\right) \\
&= \frac{\alpha_*}{2N}\sum_{i=1}^{L}N_i\sum_{k=1}^{M_i}\mathrm{Tr}(E_k^{(i)}\rho)\log\left(\frac{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1^*)}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2^*)}\right) + \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*)\right) \\
&= \frac{\alpha_*}{2N}\sum_{i=1}^{L}N_i\left[\sum_{k=1}^{M_i}\mathrm{Tr}(E_k^{(i)}\rho)\log\left(\frac{\mathrm{Tr}(E_k^{(i)}\rho)}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_2^*)}\right) - \sum_{k=1}^{M_i}\mathrm{Tr}(E_k^{(i)}\rho)\log\left(\frac{\mathrm{Tr}(E_k^{(i)}\rho)}{\mathrm{Tr}(\widetilde{E}_k^{(i)}\chi_1^*)}\right)\right] \\
&\quad + \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*)\right) \\
&= \frac{\alpha_*}{2N}\sum_{i=1}^{L}N_i\left[\mathrm{KL}(p_\rho^{(i)}\|\mathrm{A}^{(i)}(\chi_2^*)) - \mathrm{KL}(p_\rho^{(i)}\|\mathrm{A}^{(i)}(\chi_1^*))\right] + \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*)\right) \\
&= \frac{\alpha_*}{2N}\left[\mathrm{KL}(\mathbb{P}_{\mathfrak{M},\rho}\|\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\chi_2^*}) - \mathrm{KL}(\mathbb{P}_{\mathfrak{M},\rho}\|\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\chi_1^*})\right] + \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1^*) + \mathrm{Tr}(\mathscr{O}\chi_2^*)\right),
\end{aligned} \tag{5.43}$$

where to obtain the last equality, we used the fact that $\mathrm{KL}(p \otimes q \| p' \otimes q') = \mathrm{KL}(p\|q) + \mathrm{KL}(p'\|q')$ for probability distributions p, p', q, q'. Since the bias of $\widehat{\mathcal{O}}_*$ is defined as $\mathrm{Tr}(\mathcal{O}\rho) - \mathbb{E}[\widehat{\mathcal{O}}_*]$, we obtain Eq. (5.42). □

Thus, the estimator constructed by $\mathbb{TOOL}$ is biased in general.

5.4 Optimization algorithm

To construct the estimator given in Box 3, we need to perform the optimization given in Eq. (5.6). For this, we present the approach described in [92, App. B] for performing this optimization.

Eq. (5.6) has two optimizations – an inner optimization over density matrices χ_1, χ_2, and an outer optimization over $\alpha \geq 0$. Both these optimization problems are convex, and therefore, can be solved with rigorous convergence guarantees.

Box 5: Optimization algorithm

Input: Observable $\mathcal{O}$, measurement protocol $\mathfrak{M}$, confidence level $1 - \delta \in (0.75, 1)$, parameter $0 < \epsilon_\circ \ll 1$

Optimization algorithm:

(1) For a given $\alpha \geq 0$, define the function $f_\alpha \colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ as

$$f_\alpha(\chi_1, \chi_2) = -\frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) + \alpha \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1, \chi_2). \tag{5.44}$$

Then, solve the inner convex optimization problem in Eq. (5.6)

$$\begin{aligned} &\max_{\chi_1, \chi_2 \in \mathscr{X}} \left[\frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) - \alpha \mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1, \chi_2)\right] \\ &= - \min_{(\chi_1, \chi_2) \in \mathscr{X} \times \mathscr{X}} f_\alpha(\chi_1, \chi_2) \end{aligned} \tag{5.45}$$

using the version of Nesterov's second method [76] given in Ref. [101].

(2) Perform the outer convex optimization over α in Eq. (5.6) using any algorithm that can find a local minimum of a real-valued function on $[0, \infty)$.

Nesterov's second method [76, 101], used for optimizing the function f_α defined in Eq. (5.44) over $\mathscr{X} \times \mathscr{X}$, is an accelerated version of proximal gradient descent with the property that each iterate lies in $\mathscr{X} \times \mathscr{X}$. Nesterov's second method is, therefore, suitable for optimizing convex functions of the density matrix. Nesterov's second method requires the Lipschitz constant of the gradient of the objective function. When this Lipschitz constant is not known explicitly, one can use a backtracking scheme to estimate the Lipschitz constant [101]. For implementing Nesterov's second method for the inner optimization in Eq. (5.45), one needs to be able to project a given Hermitian matrix onto the set of density matrices. This can be done, for example, by diagonalizing the Hermitian matrix, and projecting the eigenvalues onto the standard simplex using [104]. This is one of the computationally costliest parts of the algorithm since diagonalizing a $d \times d$ matrix requires $O(d^3)$ time in the worst case. Additionally, because we need to compute $\mathrm{Tr}(E_k^{(i)} \chi_1)$ and $\mathrm{Tr}(E_k^{(i)} \chi_2)$ for $k \in [M_i]$ and $i \in [L]$ to evaluate $\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}(\chi_1, \chi_2)$, we have an extra time cost of $O(Md^2)$, where $M = \sum_{i=1}^{L} M_i$ is the total number of POVM elements. Similarly, since we need to store all the POVM elements in memory, we also use $O(Md^2)$ memory. The outer optimization over α in Eq. (5.6) is convex and one-dimensional. Therefore, the optimization algorithm given in Box 5 takes a total of $O(d^3 + Md^2)$ time and $O(Md^2)$ memory. Finally, we remark that while theoretically $\epsilon_\circ$ can be an arbitrarily small positive number, from a practical perspective, a very small value of $\epsilon_\circ$ can lead to numerical issues since the gradient of f_α can be very large. We leave the problem of devising methods to circumvent such numerical issues for future work.

Since the outer optimization over $\alpha \geq 0$ is one-dimensional and convex, off-the-shelf routines work well because it suffices to find a local minimum (recall that every local minimum of a convex function is a global minimum). The inner optimization over $\mathscr{X} \times \mathscr{X}$, however, can be high-dimensional, and one needs a convergence guarantee. Below, we show that the inner optimization problem satisfies the requirements that guarantee the convergence of Nesterov's second method to within a specified precision.

Proposition 5.14. *1. The objective function f_α defined in Eq. (5.44) is convex and smooth on an open set containing $\mathscr{X} \times \mathscr{X}$.*

2. The gradient of f_α is a Lipschitz continuous function on $\mathscr{X} \times \mathscr{X}$.

3. Nesterov's second method for optimizing f over $\mathscr{X} \times \mathscr{X}$ is guaranteed to converge to the optimum to within the specified precision.

Proof. 1. To define the derivatives of f_α on $\mathscr{X} \times \mathscr{X}$, f needs to be well-defined on an open set $\mathscr{D}$ containing $\mathscr{X} \times \mathscr{X}$. For that purpose, denote $r = \epsilon_\circ/(2 \max_{i\in[L]} M_i)$ and define $\mathscr{D} = \mathrm{conv}(\cup_{\chi_1,\chi_2\in\mathscr{X}} B_1(\chi_1, r)\times B_1(\chi_2, r))$, where $B_1(\chi, r) = \{\chi' \in \mathscr{X} \mid \|\chi' - \chi\|_1 < r\}$. Since $\cup_{\chi_1,\chi_2\in\mathscr{X}} B_1(\chi_1, r)\times B_1(\chi_2, r)$ is open, it is contained in $\mathrm{int}\,\mathscr{D}$. Since $\mathscr{D}$ is convex, so is $\mathrm{int}\,\mathscr{D}$ [8, Prop. (3.45)]. But $\mathscr{D}$ is the smallest convex set containing $\cup_{\chi_1,\chi_2\in\mathscr{X}} B_1(\chi_1, r) \times B_1(\chi_2, r)$, implying that $\mathscr{D} = \mathrm{int}\,\mathscr{D}$, so that $\mathscr{D}$ is open. Furthermore, $\mathscr{X} \times \mathscr{X} \subseteq \mathscr{D}$ by construction. Now, for all $\chi_1 \in \mathscr{X}$ and all $\chi \in B_1(\chi_1, r)$, we have $\mathrm{Tr}(E_k^{(i)}\chi) = \mathrm{Tr}(E_k^{(i)}\chi_1) + \mathrm{Tr}(E_k^{(i)}(\chi - \chi_1))$. We have $0 \leq \mathrm{Tr}(E_k^{(i)}\chi_1) \leq 1$, and by matrix Hölder's inequality, we have $|\mathrm{Tr}(E_k^{(i)}(\chi-\chi_1))| \leq \left\|E_k^{(i)}\right\|_\infty \|\chi - \chi_1\|_1 \leq r$, since $\left\|E_k^{(i)}\right\|_\infty \leq 1$ for each POVM element $E_k^{(i)}$. A similar statement holds for all $\chi_2 \in \mathscr{X}$ and all $\chi \in B_1(\chi_2, r)$. It follows from Eq. (5.10) that $\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}$, and therefore, f_α, is well-defined on $\mathscr{D}$. Furthermore, it can be verified from Lem. 5.5 that $\mathrm{BD}_{\mathfrak{M}(\epsilon_\circ)}$ is convex and smooth on $\mathscr{D}$, and thus, f_α is convex and smooth on $\mathscr{D}$.

2. Since f_α is smooth on $\mathscr{D}$, its derivatives are continuous. In particular, its Hessian is continuous on $\mathscr{X} \times \mathscr{X}$, and since $\mathscr{X}$ is compact, the Hessian is bounded on $\mathscr{X} \times \mathscr{X}$. Then, from the mean value theorem, it follows that ∇f_α is Lipschitz continuous.

3. Since f_α is convex with a Lipschitz continuous gradient, and $\mathscr{X} \times \mathscr{X}$ is a compact and convex set, Nesterov's second method for minimizing f_α over $\mathscr{X} \times \mathscr{X}$ is guaranteed to converge (see [100, Sec. 3] and [101, Thm. 1(c)]). □

An open source implementation of the algorithm in Box 5 can be found in [90].

Chapter 6

Application to fidelity estimation

In this chapter, we apply $\mathbb{TOOL}$ to the problem of estimating the fidelity with a pure state. The results presented in this chapter are borrowed from the papers [91, 92], albeit with some modifications to fit the presentation of the previous chapters. We begin by presenting the motivation for studying the problem of fidelity estimation.

A typical goal in experiments and applications is to prepare a pure quantum state ρ_{target}, which we call the target state, as a resource for other quantum information tasks like computation and communication. However, due to noise and experimental imperfections in the current quantum devices, one usually ends up preparing a mixed state ρ, which we hope is close to the target state. A commonly used measure in experiments to check whether two states are close to each other is the quantum fidelity (Def. 3.5). When the target state ρ_{target} is pure, the quantum fidelity between the experimentally prepared state ρ and the target state ρ_{target} is equal to $F(\rho_{\text{target}}, \rho) = \text{Tr}(\rho_{\text{target}}\rho)$ (Prop. 3.6.5). Thus, estimating fidelity with ρ_{target} is equivalent to learning the expectation value of ρ_{target}, which allows us to use $\mathbb{TOOL}$ for this problem.

In Sec. 6.1, we apply $\mathbb{TOOL}$ on experimental data to learn the fidelity with a desired target state. We also compare the fidelity estimates given by $\mathbb{TOOL}$ with those obtained from quantum tomography performed using maximum likelihood estimation (MLE). In Sec. 6.2, inspired by direct fidelity estimation (DFE) [33, 26], we study a randomized Pauli measurement scheme, where the probability of sampling a Pauli operator is determined by ρ_{target}. We show that $\mathbb{TOOL}$ can give better guarantees than DFE for this sampling scheme, and it is also possible to construct exact

confidence intervals for fidelity.

6.1 Comparison with maximum likelihood estimation

In this section, we test the estimator constructed by 𝕋𝕆𝕆𝕃 on experimental data from a trapped-ion quantum processor [84] for the task of estimating fidelity with a quantum state. We consider data for three different 4-qubit target states: a GHZ state, a W-state, and a locally-rotated linear cluster state. Each dataset consists of 81 Pauli measurements, with 100 repetitions of each Pauli measurement.

𝕋𝕆𝕆𝕃 is used to construct an estimator for each of the target states we consider (GHZ state, W state, locally-rotated linear cluster state) for a confidence level of 95% following Box 3. We fix the parameter $\epsilon_\circ = 10^{-5}$ in Box 3. The estimates and the errors are computed using the experimental data following Box 4.

For comparison, we also perform Maximum Likelihood Estimation (MLE) [53, 59] to reconstruct the quantum state, as it is a popular tool that is used in many experimental studies. The fidelity is estimated from the reconstructed quantum state $\widehat{\rho}$ by computing $\mathrm{Tr}(\rho_{\mathrm{target}}\widehat{\rho})$. Since MLE only provides a point estimate, we need some method to obtain uncertainty bound for the computed estimate. For this purpose, we use a variant of the bootstrap method [27] for computing confidence intervals. To construct a bootstrap confidence interval, we "artifically" generate outcomes according to the observed frequencies in the experiment (these artificial outcomes are generated using a classical computer). The state is reconstructed using the artificially generated outcomes, which then gives an "artificial" estimate for fidelity. This process is repeated many times, and we construct a (possibly asymmetric) confidence interval at the specified confidence level around the median of the artificial estimates. The confidence interval is then shifted from the median to the original MLE estimate computed from the experimental data. Note that the confidence intervals so constructed are heuristic, and generally does not satisfy the guarantee that the true value lies inside the computed confidence interval at the specified confidence level. For comparison purposes, we define the error for bootstrap as half the size of the bootstrap confidence interval. We call this error

the "bootstrap error".

In Tab. 3, we list the fidelity estimates obtained from TOOL and MLE, along with the respective errors. When we compare the error given by TOOL with the bootstrap error, it should be understood that we are comparing the size of the respective confidence intervals. We can see from Tab. 3 that the estimates obtained from TOOL and MLE agree well with each other. The error ε_* computed from TOOL, however, is about 2.5 times the size of bootstrap error. There are two main reasons for this discrepancy. One, the bootstrap error depends on the state that is prepared, unlike TOOL which gives worst-case (minimax) errors by construction. Two, the bootstrap error is heuristic, and not always guaranteed to be correct. This is especially true when the fidelity is high, as we get close to the boundary of the set of quantum states. In addition, MLE is prone to several problems, which have been well-documented in the literature [67, 85, 87, 32, 15]. In contrast, TOOL comes equipped with rigorous guarantees, and is minimax optimal (see Thm. 7.13).

	TOOL		MLE	
	Estimate	ε_*	Estimate	Bootstrap error
GHZ	0.84	0.053	0.84	0.023
W	0.89	0.049	0.88	0.019
Cluster	0.79	0.048	0.79	0.021

Table 3: Fidelity estimates and estimation error for a 4-qubit GHZ state, W state, and a cluster state obtained from experimental data for a confidence level of 95%. Estimates are calculated using TOOL and MLE. The error for MLE is obtained from Monte-Carlo (MC) resampling.

To study the claim that MLE with bootstrap can give incorrect results, we consider a numerical example. Our goal is to estimate the fidelity with a 3-qubit W-state. For this, we measure the eigenvalues of all the non-identity Pauli operators that have non-zero overlap with the W state. Each of these Pauli operators is measured 100 times. In order to perform these measurements numerically, we choose a true state ρ that has a fidelity of 0.991 with the target state. Such a high fidelity means that the true state is close to the boundary of the set of density matrices. We perform

a total of 150 simulations, where the state ρ is prepared, the Pauli measurements are performed, the fidelity is estimated using MLE, and confidence intervals are constructed using bootstrap. We find that MLE with bootstrap gives an empirical coverage probability of 72%, which is much smaller than the chosen confidence level of 95%. This means that the error reported by bootstrap is too small. Therefore, for a realistic situation that one may encounter in practice, we find that MLE with bootstrap gives incorrect results.

In addition to the statistical correctness of TOOL compared to MLE and bootstrap, TOOL has the advantage that it is less computationally costly to implement compared to MLE and bootstrap. This is because the costly optimization to compute the estimator only needs to be performed once for TOOL, whereas one needs to repeatedly reconstruct the state for constructing bootstrap confidence interval. Moreover, TOOL can construct the estimator independently of the experiments, so that the estimator construction does not lead to a bottleneck. MLE, on the other hand, can only be performed after the experiment is completed, which can lead to a computational bottleneck in the characterization process. Finally, since the estimator constructed by TOOL can be reused and can also efficiently compute estimates from data, this estimator may be used for uncertainty quantification in place of the MLE estimator. This might be useful, for example, in situations where we wish to model and understand the effects of noise on the computed estimates.

6.2 Comparison with direct fidelity estimation

Direct fidelity estimation (DFE) [33, 26] is a technique that estimates the fidelity with a given target state without reconstructing the state (hence "direct"). This is achieved by judiciously sampling Pauli operators based on the specified target state and measuring them. In this section, we introduce a slightly different importance sampling scheme for Pauli operators for estimating the fidelity with a given target state, and show that TOOL gives improves upon the performance of DFE for this sampling scheme. We show that TOOL can use outcomes of the modified importance sampling scheme to get a significant improvement in the dependence of the sample complexity on δ over a biased version of DFE, in the worst case over all target states. Note that we compare with a

biased version of DFE because the sample complexity of the commonly used unbiased version of DFE in the worst case over all target states is infinity. For well-conditioned states, the performance of $\mathbb{TOOL}$ for the modified importance sampling scheme matches with the performance of DFE.

We begin by describing the DFE measurement protocol. Let ρ_{target} be an n-qubit pure target state and denote $d = 2^n$. For $i \in \{0, \ldots, d^2 - 1\}$, the Pauli P_i is sampled with probability $(\text{Tr}(P_i \rho_{\text{target}}))^2/d$, and subsequently we measure the POVM $\{(\mathbb{I} + P_i)/2, (\mathbb{I} - P_i)/2\}$. This procedure is repeated many times, and the outcomes are used to learn the fidelity with ρ_{target} according to the estimation procedure given by [33, 26]. This estimation procedure constructs an unbiased estimator for fidelity in terms of the expectation values of Pauli operators that are sampled according to the DFE sampling scheme (we refer the reader to [33, 26] for details). The sample complexity of DFE in the worst case over all target states is equal to *infinity*. This is because there are target states for which there is an arbitrarily small (but non-zero) probability that we sample Pauli operators which have an arbtirarily small overlap with the target state, which in turn leads to an arbitrarily large sample complexity for DFE. That said, since the probability of sampling such Pauli operators is very small, it is perhaps unfair to study the performance of DFE in the worst case over all target states. For this reason, we focus on comparing with a biased version of DFE that was noted in [33, 26], which avoids sampling Pauli operators that have too small an overlap with the target states. [33, 26] show that the number of samples required to learn the fidelity to an error $\varepsilon > 0$ and confidence level $1 - \delta > 0$ using biased DFE is bounded above by

$$O\left(\frac{d}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) + \frac{1}{\varepsilon^2 \delta}\right) \tag{6.1}$$

in the worst case over all target states.

If instead of looking at the worst-case performance over all target states, we look at "well-conditioned" target states, the performance of DFE can be significantly improved. ρ_{target} is said to be **well-conditioned** with parameter $\beta > 0$ if for all $k \in \{0, \ldots, d^2 - 1\}$, we have either $\text{Tr}(P_k \rho_{\text{target}}) = 0$ or $|\text{Tr}(P_k \rho_{\text{target}})| \geq \beta$ [33]. Many interesting states, such as stabilizer states and

Dicke states, are well-conditioned [33, 26]. Observe that well-conditioned target states avoid the problem of having arbitrarily small (but non-zero) overlap of Pauli observables with the target state, and therefore, there is no need to truncate the sampling probabilities and introduce a bias. For target states that are well-conditioned with parameter β, the sample complexity of DFE is [33]

$$O\left(\frac{1}{\beta^2\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right). \tag{6.2}$$

If β scales polynomially with $1/n$, then DFE can efficiently estimate the fidelity with ρ_{target}. Importantly, for all stabilizer states, we have $\beta = 1$, which means that we can estimate the fidelity with a pure stabilizer state with a number of samples that does not scale with the dimension of the system.

We work with a slightly different sampling scheme for estimating the fidelity with ρ_{target} that gives the same or better guarantees than DFE. This importance sampling scheme was studied in [98, 92].

Box 6: Importance sampling-based Pauli measurements

Input: Pure state ρ_{target}, total number of samples N.

Procedure:

(1) Sample a non-identity Pauli operator P_i for $i \in [d^2-1]$ with probability

$$p_i = \frac{|\text{Tr}(P_i\rho_{\text{target}})|}{\sum_{i=1}^{d^2-1}|\text{Tr}(P_i\rho_{\text{target}})|}. \tag{6.3}$$

(2) Measure the eigenvalue of the sampled Pauli, and record the measurement outcome.

(3) Flip the measurement outcome $\pm 1 \to \mp 1$ if $\text{Tr}(P_i\rho_{\text{target}}) < 0$.

(4) Repeat the procedure N times.

Output: Post-processed measurement outcomes

Because we flip the measurement outcome in step (3) of Box 6 depending on the sign of $\mathrm{Tr}(P_i\rho_{\mathrm{target}})$, we are effectively measuring $S_i = \mathrm{sign}(\mathrm{Tr}(P_i\rho))P_i$. We seek to measure S_i because we can write the state ρ_{target} as

$$\begin{aligned}\rho_{\mathrm{target}} &= \frac{\mathbb{I}}{d} + \sum_{i=1}^{d^2-1} \frac{\mathrm{Tr}(P_i\rho_{\mathrm{target}})}{d} P_i \\ &= \frac{\mathbb{I}}{d} + \frac{\mathscr{N}}{d} \sum_{i=1}^{d^2-1} p_i S_i,\end{aligned} \tag{6.4}$$

where

$$\mathscr{N} = \sum_{i=1}^{d^2-1} |\mathrm{Tr}(P_i\rho_{\mathrm{target}})| \tag{6.5}$$

is the normalization factor in Eq. (6.3). Moreover, if the true state is ρ, then the probability of obtaining $+1$ outcome in the measurement protocol defined in Box 6 is equal to

$$\sum_{i=1}^{d^2-1} p_i \mathrm{Tr}\left(\frac{\mathbb{I}+S_i}{2}\rho\right). \tag{6.6}$$

Thus, the effective POVM describing the measurement protocol in Box 6 is $\{\Theta_0, \mathbb{I}-\Theta_0\}$, where

$$\Theta_0 = \sum_{i=1}^{d^2-1} p_i \left(\frac{\mathbb{I}+S_i}{2}\right) = \left(\frac{d+(\mathscr{N}-1)}{2\mathscr{N}}\right)\rho_{\mathrm{target}} + \left(\frac{\mathscr{N}-1}{2\mathscr{N}}\right)(\mathbb{I}-\rho_{\mathrm{target}}). \tag{6.7}$$

Motivated by this observation, we compute the sample complexity of $\mathbb{TOOL}$ for estimating the fidelity with ρ_{target} by measuring the POVM $\{\omega_1\rho_{\mathrm{target}} + \omega_2(\mathbb{I}-\rho_{\mathrm{target}}), (1-\omega_1)\rho_{\mathrm{target}} + (1-\omega_2)(\mathbb{I}-\rho_{\mathrm{target}})\}$, where $\omega_1, \omega_2 \in [0,1]$. The case of $\omega_1 = \omega_2$ gives a trivial POVM, and can therefore be discarded. Consequently, there is no loss of generality in taking $\omega_1 > \omega_2$. The following result is a modified version of [92, Thm. C.1].

Proposition 6.1. *Given a pure state ρ_{target}, consider the measurement protocol $\mathfrak{M} = \{(\{\Theta, \mathbb{I}-\Theta\}, N)\}$, where*

$$\Theta = \omega_1\rho_{\mathrm{target}} + \omega_2(\mathbb{I}-\rho_{\mathrm{target}}) \tag{6.8}$$

for $\omega_1, \omega_2 \in [0,1]$ with $\omega_1 > \omega_2$. Given $\epsilon_\circ > 0$, denote $\omega_i' = (\omega_1 + \epsilon_\circ/2)/(1+\epsilon_\circ)$ for $i = 1, 2$. Then, the following results hold.

1. *The error ε_* of the estimator constructed by $\mathbb{TOOL}$ with parameter $\epsilon_\circ > 0$ using the outcomes of $\mathfrak{M}$ to a confidence level of $1-\delta \in (0,1)$ can be written as*

$$\begin{aligned} \varepsilon_* = \max_{\lambda_1,\lambda_2\in[0,1]} \quad & \frac{1}{2}(\lambda_1 - \lambda_2) \\ \text{s.t.} \quad & -\log\left(\sqrt{(\omega_2' + (\omega_1' - \omega_2')\lambda_1)(\omega_2' + (\omega_1' - \omega_2')\lambda_2)}\right. \\ & \left. + \sqrt{((1-\omega_2') + (\omega_2' - \omega_1')\lambda_1)((1-\omega_2)' + (\omega_2' - \omega_1')\lambda_2)}\right) \geq \frac{1}{N}\log\left(\frac{2}{\delta}\right). \end{aligned} \tag{6.9}$$

   Furthermore, the estimator in Box 3 can be constructed in $O(1)$ time and memory irrespective of the system dimension.

2. *The number of samples needed by $\mathbb{TOOL}$ to estimate $\mathrm{Tr}(\rho_{\text{target}}\rho)$ to within an error of $\varepsilon \in (0, 0.5)$ and a confidence level of $1-\delta \in (0,1)$ using the outcomes of $\mathfrak{M}$ and parameter $\epsilon_\circ > 0$ is at most*

$$\frac{(1+\epsilon_\circ)^2}{2(\omega_1 - \omega_2)^2}\frac{\log(2/\delta)}{\varepsilon^2}. \tag{6.10}$$

Proof. 1. For every state χ, there is a number $\lambda \in [0,1]$ and an observable $\mathcal{O}^\perp$ such that $\mathrm{Tr}(\mathcal{O}^\perp) = 1$, $\mathrm{Tr}(\rho_{\text{target}}\mathcal{O}^\perp) = 0$, and

$$\chi = \lambda\rho_{\text{target}} + (1-\lambda)\mathcal{O}^\perp. \tag{6.11}$$

The number λ is uniquely determined by χ. Using this fact along with Eq. (3.8), it can be verified that

$$\begin{aligned} \mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2) = & \sqrt{(\omega_2' + (\omega_1' - \omega_2')\lambda_1)(\omega_2' + (\omega_1' - \omega_2')\lambda_2')} \\ & + \sqrt{((1-\omega_2') + (\omega_2' - \omega_1')\lambda_1)((1-\omega_2') + (\omega_2' - \omega_1')\lambda_2)} \end{aligned} \tag{6.12}$$

for all $\chi_1, \chi_2 \in \mathscr{X}$, where $\lambda_1, \lambda_2 \in [0,1]$ are determined by χ_1, χ_2 according to Eq. (6.11). Then, Eq. (5.18) gives Eq. (6.9). Eq. (5.6) is just the dual problem of the optimization in Eq. (6.9) (see Prop. 4.10.3), and therefore, the primal optimal $(\lambda_1^*, \lambda_2^*)$ and dual optimal (α_*) points can be

computed in $O(1)$ time and memory. It can be verified that ϕ_* in Eq. (5.8) depends only on λ_1^*, λ_2^* and α_*, and therefore, the estimator in Eq. (5.9) can be computed in $O(1)$ time and memory.

2. From Eq. (5.18) and Prop. 4.15.2, for $\varepsilon_* < 0.5$, we must have $\alpha_* > 0$. By complementary slackness, this implies $\mathrm{BC}_{\mathfrak{M}}(\chi_1^*, \chi_2^*) = (\delta/2)^{1/N}$, where χ_1^*, χ_2^* denote the states attaining the maximum in Eq. (5.18). Let $\lambda_1^*, \lambda_2^* \in [0,1]$ be determined by χ_1^*, χ_2^* as per Eq. (6.11). Therefore, denoting $\gamma = (\delta/2)^{2/N}$, there is some $a^* \in [0,1]$ such that

$$\begin{aligned}\sqrt{(\omega_2' + (\omega_1' - \omega_2')\lambda_1^*)(\omega_2' + (\omega_1' - \omega_2')\lambda_2^*)} &= a^*\sqrt{\gamma} \\ \sqrt{((1-\omega_2') + (\omega_2' - \omega_1')\lambda_1^*)((1-\omega_2') + (\omega_2' - \omega_1')\lambda_2^*)} &= (1-a^*)\sqrt{\gamma}\end{aligned} \tag{6.13}$$

and $\varepsilon_* = (\lambda_1^* - \lambda_2^*)/2$. Solving for λ_1^*, λ_2^* in terms of a^*, we obtain

$$\begin{aligned}\lambda_1^* &= \frac{(2a^*-1)\gamma + (1-2\omega_2')}{2(\omega_1' - \omega_2')} + \frac{\sqrt{1-\gamma}}{2(\omega_1' - \omega_2')}\sqrt{1-(2a^*-1)^2\gamma} \\ \lambda_2^* &= \frac{(2a^*-1)\gamma + (1-2\omega_2')}{2(\omega_1' - \omega_2')} - \frac{\sqrt{1-\gamma}}{2(\omega_1' - \omega_2')}\sqrt{1-(2a^*-1)^2\gamma}.\end{aligned} \tag{6.14}$$

Thus, we have the bound

$$\varepsilon_* \leq \frac{\sqrt{1-\gamma}}{2(\omega_1' - \omega_2')}\sqrt{1-(2a^*-1)^2\gamma} \leq \frac{\sqrt{1-\gamma}}{2(\omega_1' - \omega_2')}. \tag{6.15}$$

Using $\gamma = (\delta/2)^{2/N}$ and Eq. (8.17), we obtain

$$\varepsilon_* \leq \frac{1}{2(\omega_1' - \omega_2')}\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{6.16}$$

Setting $\varepsilon_* = \varepsilon$ and solving for N gives Eq. (6.10). □

We now compute a bound on the sample complexity of $\mathbb{TOOL}$ for the measurement protocol described in Box 6. The following result is a modified version of [92, Thm. II.2].

Corollary 6.2. *$\mathbb{TOOL}$ can estimate the fidelity with a pure target state ρ_{target} to an error of*

$\varepsilon \in (0, 0.5)$ and a confidence level of $1 - \delta \in (0, 1)$ using

$$O\left(\left(\frac{\mathcal{N}}{d}\right)^2 \frac{1}{\varepsilon^2} \log\left(\frac{2}{\delta}\right)\right) \tag{6.17}$$

outcomes of the measurement protocol in Box 6, where $\mathcal{N}$ is given in Eq. (6.5). Furthermore, for all ρ_{target}, we have $\mathcal{N} \leq \sqrt{d+1}(d-1)$, and if ρ_{target} is well-conditioned with parameter $\beta > 0$, then the sample complexity of $\mathbb{TOOL}$ is

$$O\left(\frac{1}{\beta^2 \varepsilon^2} \log\left(\frac{2}{\delta}\right)\right). \tag{6.18}$$

Proof. From Eq. (6.7), the effective POVM of Box 6 can be written as $\{\omega_1 \rho_{\text{target}} + \omega_2(\mathbb{I} - \rho_{\text{target}}), (1 - \omega_1)\rho_{\text{target}} + (1 - \omega_2)(\mathbb{I} - \rho_{\text{target}})\}$ with

$$\begin{aligned} \omega_1 &= \left(\frac{d + (\mathcal{N} - 1)}{2\mathcal{N}}\right), \\ \omega_2 &= \left(\frac{\mathcal{N} - 1}{2\mathcal{N}}\right). \end{aligned} \tag{6.19}$$

Then, Eq. (6.17) follows from Prop. 6.1.

The bound on $\mathcal{N}$ can be obtained by solving the convex optimization $\sum_{i=1}^{d^2-1} x_i$ subject to the constraints $x_i \geq 0$ for all $i \in [d^2 - 1]$, and $\sum_{i=1}^{d^2-1} x_i^2 \leq d - 1$. See [92, Thm. II.2] for details.

Now, suppose that ρ_{target} is well-conditioned with parameter $\beta > 0$. Since ρ_{target} is pure, we must have $\sum_{i=1}^{d^2-1} (\text{Tr}(P_i \rho_{\text{target}}))^2 = d - 1$. Then, if K denotes the number of non-identity Paulis with non-zero overlap with ρ_{target}, we have $K \leq (d-1)/\beta$ by definition of well-conditioned state, and therefore, $\mathcal{N} \leq K \leq (d-1)/\beta$. Then, Eq. (6.18) follows from Eq. (6.17). □

The sample complexity of $\mathbb{TOOL}$ for the measurement procedure given in Box 6 is bounded above by $O(d \log(2/\delta)/\varepsilon^2)$ in worst case over all true states and all target states. Observe that the dependence on δ is significantly better for $\mathbb{TOOL}$ compared to biased DFE in the worst case over all target states ($O(\log(1/\delta))$ for $\mathbb{TOOL}$ versus $O(1/\delta)$ in DFE). On the other hand, if the target states are well-conditioned, then the sample complexity of $\mathbb{TOOL}$ is the same as the sample complexity of DFE.

Chapter 7

Lower bounds on learning expectation values

In this chapter, we first define the minimax norm and studying its properties in Sec. 7.1. Then, in Sec. 7.2, we show that for a fixed measurement protocol, the minimax norm gives a tight lower bound on the estimation error. Subsequently, we show in that $\mathbb{TOOL}$ can achieve this lower bound to within a small factor. Often the measurement protocol is fixed based on experimental constraints and the observable whose expectation value we wish to learn. If we had the ability to implement any measurement protocol of our choice, then we show in Sec. 7.3 that measuring in the eigenbasis of the observable is optimal. Finally, in Sec. 7.4, we discuss extension of our lower bound to learning the expectation values of many observables simultaneously in l_∞-norm.

7.1 Minimax norm

In this section, we define the minimax norm and study its properties. We also look at a few different interpretations of the minimax norm.

Since we only have access to finitely many outcomes in any experiment implemented in practice, we cannot have perfect knowledge of the quantum state or the expectation value. The minimax norm seeks to quantify how large the uncertainty in estimating the expectation value must be for any estimation procedure in the worst case.

Definition 7.1 (Minimax norm). Given a measurement protocol $\mathfrak{M}$ and a confidence level

$1-\delta \in (0,1)$, the minimax norm of an observable $\mathcal{O}$ determined by $\mathfrak{M}$ and $1-\delta$ is defined as

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} = \max_{\chi_1,\chi_2\in\mathscr{X}} \quad & \frac{1}{2}\left(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)\right) \\ \text{s.t.} \quad & \mathrm{BD}_{\mathfrak{M}}(\chi_1,\chi_2) \leq \frac{1}{N}\log\left(\frac{2}{\delta}\right), \end{aligned} \tag{7.1}$$

where N is the total number of samples used by $\mathfrak{M}$ and $\mathrm{BD}_{\mathfrak{M}}$ is the average Bhattacharyya distance defined in Eq. (3.5). □

In the following two paragraphs, we motivate the definition of the minimax norm by relating it to the estimation error. We begin by motivating the need to study the worst-case estimation error. Suppose that the state ρ was prepared in an experiment, but we don't know ρ. Our goal is to learn the expectation value of the observable $\mathcal{O}$. So we choose a measurement protocol $\mathfrak{M}$ for this purpose, and perform measurements specified by $\mathfrak{M}$ on ρ. The outcomes observed in the experiment will give us some confidence region $\mathscr{C}$ within which the expectation value $\mathrm{Tr}(\mathcal{O}\rho)$ must lie with probability greater than $1-\delta$. We can define the estimation error as half size of the confidence region, given as $(1/2)\max_{o_1,o_2\in\mathscr{C}}(o_1-o_2)$. Since the measurement outcomes are obtained probabilistically, they can be different every time the experiment is performed. Therefore, it is helpful to know what is the worst-case error in learning the expectation value of $\mathcal{O}$ using the measurement protocol $\mathfrak{M}$, no matter what state ρ was/will be prepared by the device, or what measurements outcomes were/will be observed when implementing $\mathfrak{M}$.

We can, therefore, ask for the following constraint: if ρ is the true state, then the "distance" between ρ and another state χ_1 consistent with the measurements is bounded above by some number u. Since we want the error to be dependent on the measurement protocol but independent of the measurement outcomes, we use a distance measure defined on the probability distributions determined by the measurement protocol, which leads us to classical distance measures we studied in Ch. 3. Furthermore, we posit that u should depend on the chosen confidence level $1-\delta$ and the number of samples N as follows: u increases as δ decreases, and u decreases when N increases. The reason is that if δ is very small, then we need to allow for a larger error since we want to estimate to

a very high confidence level. On the other hand, if N is large, then we have a large amount of data, using which we can reduce the estimation error. When defining the minimax norm, we claim that the "correct" distance measure to look at is the average Bhattacharyya distance determined by $\mathfrak{M}$ and the "correct" upper bound is $u = \log(2/\delta)/N$. This gives us the constraint $\mathrm{BD}_{\mathfrak{M}}(\chi_1, \rho) \leq \log(2/\delta)/N$. Since we want our uncertainty bound to be valid no matter what state ρ was prepared by the device, we look at all the states χ_1 and χ_2 that satisfy the constraint $\mathrm{BD}_{\mathfrak{M}}(\chi_1, \chi_2) \leq \log(2/\delta)/N$. Then, the minimax norm defined in Eq. (7.1) is just (half the) maximum difference in expectation value of $\mathscr{O}$ between states satisfying this constraint.

Since the average Bhattacharyya distance $\mathrm{BD}_{\mathfrak{M}}$ is the negative logarithm of the geometric-average Bhattacharyya coefficient $\mathrm{BC}_{\mathfrak{M}}$ (Eq. (3.6)), and the geometric-average classical fidelity satisfies $\mathrm{FC}_{\mathfrak{M}} = \mathrm{BC}_{\mathfrak{M}}^2$, we can rewrite the minimax norm in terms of these as follows.

$$\begin{aligned} \|\mathscr{O}\|_{\mathfrak{M},\delta} &= \max_{\chi_1,\chi_2 \in \mathscr{X}} \quad \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2)\right) \\ &\qquad \text{s.t.} \quad \mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2) \geq \left(\frac{\delta}{2}\right)^{1/N}, \\ &= \max_{\chi_1,\chi_2 \in \mathscr{X}} \quad \frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2)\right) \\ &\qquad \text{s.t.} \quad \mathrm{FC}_{\mathfrak{M}}(\chi_1, \chi_2) \geq \left(\frac{\delta}{2}\right)^{2/N}. \end{aligned} \tag{7.2}$$

Thus, we can interpret the minimax norm in terms of $\mathrm{BC}_{\mathfrak{M}}$ or $\mathrm{FC}_{\mathfrak{M}}$ instead of $\mathrm{BD}_{\mathfrak{M}}$. The advantage of working with $\mathrm{BD}_{\mathfrak{M}}$ is that it is a proper convex function (see Prop. 3.8).

We now prove that $\|\cdot\|_{\mathfrak{M},\delta}$ is a seminorm on the set $\mathbb{S}_d$ of $d \times d$ Hermitian matrices. To understand why $\|\cdot\|_{\mathfrak{M},\delta}$ is only a seminorm and not a norm on $\mathbb{S}_d$, we return to the connection between the minimax norm and the error for estimating expectation values. If the observable whose expectation value we want to learn is $\mathscr{O} = c\mathbb{1}$ for some $c \in \mathbb{R}$, then its expectation value is equal to c, no matter what the state is. Thus, the error for learning the expectation value of such an observable should be zero. This notion is captured by the minimax norm, where we have $\|c\mathbb{1}\|_{\mathfrak{M},\delta} = 0$ for all $c \in \mathbb{R}$. This is what leads to $\|\cdot\|_{\mathfrak{M},\delta}$ being a seminorm instead of a norm. Indeed, we show that when

we "mod out" the constant matrices from $\mathbb{S}_d$, $\|\cdot\|_{\mathfrak{M},\delta}$ becomes a norm.

Proposition 7.2. *Fix the measurement protocol $\mathfrak{M}$ and the confidence level $1-\delta \in (0,1)$. Then, the following statements hold.*

1. $\|\cdot\|_{\mathfrak{M},\delta}$ is a seminorm on $\mathbb{S}_d$.

2. Denoting $\mathscr{I} = \{c\mathbb{1} \mid c \in \mathbb{R}\}$, $\|\cdot\|_{\mathfrak{M},\delta}$ is a norm on $\mathbb{S}_d/\mathscr{I}$.

Proof. 1. First, we prove that $\|\cdot\|_{\mathfrak{M},\delta}$ is non-negative. Since $\chi_1 = \chi_2$ satisfies $\mathrm{BD}_{\mathfrak{M}}(\chi_1,\chi_2) = 0$ and $\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2) = 0$, we have $\|\mathscr{O}\|_{\mathfrak{M},\delta} \geq 0$ for all observables $\mathscr{O}$.

Next, we prove that $\|c\mathscr{O}\|_{\mathfrak{M},\delta} = |c|\,\|\mathscr{O}\|_{\mathfrak{M},\delta}$ for all $c \in \mathbb{R}$ and observables $\mathscr{O}$. Given any $c \geq 0$, we have $\|c\mathscr{O}\|_{\mathfrak{M},\delta} = c\,\|\mathscr{O}\|_{\mathfrak{M},\delta}$ by linearity and the fact that multiplication with a positive constant commutes with maximization. To handle the case of $c < 0$, we note that $\mathrm{BD}_{\mathfrak{M}}(\chi_1,\chi_2) = \mathrm{BD}_{\mathfrak{M}}(\chi_2,\chi_1)$, and thus the optimization in Eq. (7.1) is invariant under the exchange of χ_1, χ_2. Thus, for $c = -|c| < 0$, we obtain

$$\|c\mathscr{O}\|_{\mathfrak{M},\delta} = |c| \max_{\chi_1,\chi_2 \in \mathscr{X}} \left\{ -\frac{1}{2}\left(\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2)\right) \;\middle|\; \mathrm{BD}_{\mathfrak{M}}(\chi_1,\chi_2) \leq \frac{1}{N}\log\left(\frac{2}{\delta}\right) \right\} = |c|\,\|\mathscr{O}\|_{\mathfrak{M},\delta}\,. \tag{7.3}$$

Finally, we prove the triangle inequality. Given $\mathscr{O}_1, \mathscr{O}_2 \in \mathbb{S}_d$, we have $\mathrm{Tr}((\mathscr{O}_1+\mathscr{O}_2)(\chi_1-\chi_2)) = \mathrm{Tr}(\mathscr{O}_1(\chi_1-\chi_2)) + \mathrm{Tr}(\mathscr{O}_2(\chi_1-\chi_2))$. Then, because $\max_{z\in Z}(f(z)+g(z)) \leq \max_{z\in Z} f(z) + \max_{z\in Z} g(z)$ for all real-valued functions f, g and all sets Z, we can infer that $\|\mathscr{O}_1+\mathscr{O}_2\|_{\mathfrak{M},\delta} \leq \|\mathscr{O}_1\|_{\mathfrak{M},\delta} + \|\mathscr{O}_2\|_{\mathfrak{M},\delta}$.

2. The elements of $\mathbb{S}_d/\mathscr{I}$ are cosets $[\mathscr{O}] = \{\mathscr{O} + c\mathbb{1} \mid c \in \mathbb{R}\}$ for $\mathscr{O} \in \mathbb{S}_d$. Given a coset $[\mathscr{O}] \in \mathbb{S}_d/\mathscr{I}$, we show that $\|\mathscr{A}\|_{\mathfrak{M},\delta} = \|\mathscr{O}\|_{\mathfrak{M},\delta}$ for all $\mathscr{A} \in [\mathscr{O}]$. Since $\mathscr{A} \in [\mathscr{O}]$, we have $\mathscr{A} = \mathscr{O} + c\mathbb{1}$ for some real number c, and therefore, $\mathrm{Tr}(\mathscr{A}\chi_1) - \mathrm{Tr}(\mathscr{A}\chi_2) = \mathrm{Tr}(\mathscr{O}(\chi_1-\chi_2)) + c\mathrm{Tr}(\chi_1) - c\mathrm{Tr}(\chi_2) = \mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2)$ for all $\chi_1, \chi_2 \in \mathscr{X}$. It follows that $\|\mathscr{O}\|_{\mathfrak{M},\delta} = \|\mathscr{A}\|_{\mathfrak{M},\delta}$. Thus, we can unambiguously define $\|[\mathscr{O}]\|_{\mathfrak{M},\delta} = \|\mathscr{O}\|_{\mathfrak{M},\delta}$, so that $\|[\mathscr{O}]\|_{\mathfrak{M},\delta}$ is a seminorm on $\mathbb{S}_d/\mathscr{I}$.

Thus, to show that $\|\cdot\|_{\mathfrak{M},\delta}$ is a norm on $\mathbb{S}_d/\mathscr{I}$, it suffices to show that $[\mathscr{O}] \neq [0]$ implies $\|[\mathscr{O}]\|_{\mathfrak{M},\delta} > 0$. Given $[\mathscr{O}] \neq [0]$, choose any representative $\mathscr{O}$ of $[\mathscr{O}]$, so that $\mathscr{O} \neq c\mathbb{1}$ for any $c \in \mathbb{R}$. Then, we must have $\lambda_{\max}(\mathscr{O}) - \lambda_{\min}(\mathscr{O}) > 0$. Denoting $N = N(\mathfrak{M})$, $\sqrt{\gamma} = (\delta/2)^{1/N} \in (0,1)$, and

$|\lambda_{\max}\rangle, |\lambda_{\min}\rangle$ to be eigenstates of $\mathcal{O}$ corresponding to the maximum and minimum eigenvalue, we have for all $\chi \in \mathscr{X}$ that the states $\chi_1' = (1-\sqrt{\gamma})\,|\lambda_{\max}\rangle\langle\lambda_{\max}| + \sqrt{\gamma}\chi$ and $\chi_2' = (1-\sqrt{\gamma})\,|\lambda_{\min}\rangle\langle\lambda_{\min}| + \sqrt{\gamma}\chi$ satisfy $\mathrm{BC}_{\mathfrak{M}}(\chi_1', \chi_2') \geq \sqrt{F}(\chi_1', \chi_2') \geq 1 - \|\chi_1' - \chi_2'\|_{\mathrm{tr}} \geq \sqrt{\gamma}$. The first inequality is a consequence of Prop. 3.7, and the second inequality is a consequence of Fuchs-van de Graaf inequality (Eq. (3.35)). Therefore, from Eq. (7.2), we have $2\,\|\mathcal{O}\|_{\mathfrak{M},\delta} \geq (\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O}))(1-\sqrt{\gamma})$. Thus, for any finite N, we have $\|\mathcal{O}\|_{\mathfrak{M},\delta} > 0$, proving the claim. □

Due to the above result, we refer to $\|\cdot\|_{\mathfrak{M},\delta}$ as the minimax *norm*, instead of the minimax seminorm, even though it is a seminorm on $\mathbb{S}_d$. Since the minimax norm is closely related to the set of density matrices satisfying the constraint in Eq. (7.1), we define the constraint set below.

Definition 7.3. Given a measurement protocol $\mathfrak{M}$ using N samples and a confidence level $1-\delta \in (0,1)$, we define the constraint set

$$\mathscr{C}(\mathfrak{M}, \delta) = \left\{ (\chi_1, \chi_2) \in \mathscr{X} \times \mathscr{X} \mid \mathrm{BD}_{\mathfrak{M}}(\chi_1, \chi_2) \leq \frac{1}{N} \log\left(\frac{2}{\delta}\right) \right\}, \tag{7.4}$$

and the constraint-difference set as

$$\Delta\mathscr{C}(\mathfrak{M}, \delta) = \left\{ \frac{1}{2}(\chi_1 - \chi_2) \mid (\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta) \right\}. \tag{7.5}$$

□

There is a duality between the constraint-difference set and the minimax norm, which we show in Prop. 7.6. Thus, it is useful to study some properties of the constraint set and the constraint-difference set.

Proposition 7.4. *Fix the measurement protocol $\mathfrak{M}$ and the confidence level $1-\delta \in (0,1)$. Then, the following statements hold.*

1. The constraint set $\mathscr{C}(\mathfrak{M}, \delta)$ satisfies the following properties:

i. $\mathscr{C}(\mathfrak{M}, \delta)$ is compact and convex.

ii. For all $\chi \in \mathscr{X}$, we have $(\chi, \chi) \in \mathscr{C}(\mathfrak{M}, \delta)$.

iii. $(\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)$ iff $(\chi_2, \chi_1) \in \mathscr{C}(\mathfrak{M}, \delta)$.

2. The constraint-difference set $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ satisfies the following properties:

i. $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ is compact and convex.

ii. $0 \in \Delta\mathscr{C}(\mathfrak{M}, \delta)$.

iii. $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ is symmetric: $-\Delta\mathscr{C}(\mathfrak{M}, \delta) = \Delta\mathscr{C}(\mathfrak{M}, \delta)$

iv. $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ is balanced: for all $a \in \mathbb{R}$ with $|a| \leq 1$, we have $a\Delta\mathscr{C}(\mathfrak{M}, \delta) \subseteq \Delta\mathscr{C}(\mathfrak{M}, \delta)$.

v. $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ absorbs all traceless observables: for all $\mathcal{O} \in \mathbb{S}_d$ with $\mathrm{Tr}(\mathcal{O}) = 0$, there is some $r > 0$ such that $a\mathcal{O} \in \Delta\mathscr{C}(\mathfrak{M}, \delta)$ for all $|a| \leq r$.

vi. $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ spans the subspace of traceless observables: $\operatorname{span} \Delta\mathscr{C}(\mathfrak{M}, \delta) = \{\mathcal{O} \in \mathbb{S}_d \mid \mathrm{Tr}(\mathcal{O}) = 0\}$.

Proof. 1. i. Since $\mathrm{BD}_{\mathfrak{M}}$ is a proper convex function (Prop. 3.8), the constraint set $\mathscr{C}(\mathfrak{M}, \delta)$ is convex (see [8, Cor. (8.5)]). To show that $\mathscr{C}(\mathfrak{M}, \delta)$ is compact, we write it as

$$\mathscr{C}(\mathfrak{M}, \delta) = \left\{ (\chi_1, \chi_2) \in \mathscr{X} \times \mathscr{X} \mid \mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2) \geq \left(\frac{\delta}{2}\right)^{1/N} \right\}. \tag{7.6}$$

Since $\mathrm{BC}_{\mathfrak{M}}$ is continuous, the constraint $\mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2) \geq (\delta/2)^{1/N}$ defines a closed set. Since $\mathscr{X} \times \mathscr{X}$ is compact, and the intersection of a closed set and a compact set is compact, $\mathscr{C}(\mathfrak{M}, \delta)$ is compact.

ii. Since for all $\chi \in \mathscr{X}$, $\mathrm{BD}_{\mathfrak{M}}(\chi, \chi) = 0$, we have $(\chi, \chi) \in \mathscr{C}(\mathfrak{M}, \delta)$.

iii. Since $\mathrm{BD}_{\mathfrak{M}}(\chi_1, \chi_2) = \mathrm{BC}_{\mathfrak{M}}(\chi_2, \chi_1)$, we have $(\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)$ iff $(\chi_2, \chi_1) \in \mathscr{C}(\mathfrak{M}, \delta)$.

2. i. Since $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ is the linear image of $\mathscr{C}(\mathfrak{M}, \delta)$, it is a compact and convex set.

ii. Since $(\chi, \chi) \in \mathscr{C}(\mathfrak{M}, \delta)$ for all $\chi \in \mathscr{X}$, we have $0 \in \Delta\mathscr{C}(\mathfrak{M}, \delta)$.

iii. Since $(\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)$ iff $(\chi_2, \chi_1) \in \mathscr{C}(\mathfrak{M}, \delta)$, we have $-\Delta\mathscr{C}(\mathfrak{M}, \delta) = \Delta\mathscr{C}(\mathfrak{M}, \delta)$.

iv. Since $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ is symmetric, it suffices to show that $a\Delta\mathscr{C}(\mathfrak{M}, \delta) \subseteq \Delta\mathscr{C}(\mathfrak{M}, \delta)$ for $a \in [0, 1]$. Because $\Delta\mathscr{C}(\mathfrak{M}, \delta)$ is convex and $0 \in \Delta\mathscr{C}(\mathfrak{M}, \delta)$, $a\Delta\mathscr{C}(\mathfrak{M}, \delta) \subseteq \Delta\mathscr{C}(\mathfrak{M}, \delta)$ for $a \in [0, 1]$ holds.

v. Since $\Delta\mathscr{C}(\mathfrak{M},\delta)$ is balanced, it suffices to prove that there is some $r>0$ such that $r\mathcal{O}\in\Delta\mathscr{C}(\mathfrak{M},\delta)$. First take $\mathcal{O}'=\mathcal{O}/\left\|\mathcal{O}\right\|_1$. Since $\mathrm{Tr}(\mathcal{O}')=0$ and $\left\|\mathcal{O}'\right\|_1=1$, there are states χ_1,χ_2 such that $\mathcal{O}'=(\chi_1-\chi_2)/2$. Furthermore, for any $a\in[0,1)$ and any $\chi\in\mathscr{X}$, taking $\chi_1'=a\chi_1+(1-a)\chi$ and $\chi_2'=a\chi_2+(1-a)\chi$, we have $a\mathcal{O}'=(\chi_1'-\chi_2')/2$. From Eq. (3.37) and Eq. (3.32), we have $1-\mathrm{BC}_{\mathfrak{M}}(\chi_1',\chi_2')\leq\left\|\chi_1'-\chi_2'\right\|_{\mathfrak{M},\max}\leq\left\|\chi_1'-\chi_2'\right\|_{\mathrm{tr}}=a\left\|\chi_1-\chi_2\right\|_{\mathrm{tr}}\leq a$. Thus, for $a=1-(\delta/2)^{1/N}$, we have $\mathrm{BC}_{\mathfrak{M}}(\chi_1',\chi_2')\geq(\delta/2)^{1/N}$, so that $(\chi_1'-\chi_2')/2\in\Delta\mathscr{C}(\mathfrak{M},\delta)$. It follows that taking $r=a/\left\|\mathcal{O}\right\|_1$ gives $r\mathcal{O}\in\Delta\mathscr{C}(\mathfrak{M},\delta)$.

vi. Since $\Delta\mathscr{C}(\mathfrak{M},\delta)$ is absorbing for traceless operators, any traceless $\mathcal{O}\in\mathbb{S}_d$ can be written as $\mathcal{O}=c(\chi_1-\chi_2)/2$ for $(\chi_1-\chi_2)/2\in\Delta\mathscr{C}(\mathfrak{M},\delta)$ and some $c\in\mathbb{R}$. □

A set that is both convex and balanced is called a disc [75, Def. (4.2.7)]. Parts ii and iii of Prop. 7.4.2 show that $\Delta\mathscr{C}(\mathfrak{M},\delta)$ is a disc. The fact that $\Delta\mathscr{C}(\mathfrak{M},\delta)$ is absorbing for traceless matrices means that it can be expanded to fully cover the subspace of traceless matrices.

We now prove some simple but useful properties of the minimax norm.

Proposition 7.5. *Let $\mathcal{O}$ be an observable, $\mathfrak{M}$ be a measurement protocol, and $1-\delta\in(0,1)$ be the confidence level. Then, the following statements hold.*

1. *For any $c\in\mathbb{R}$, we have $\left\|\mathcal{O}+c\mathbb{1}\right\|_{\mathfrak{M},\delta}=\left\|\mathcal{O}\right\|_{\mathfrak{M},\delta}$.*

2. *Denoting $\lambda_{\max}(\mathcal{O})$ and $\lambda_{\min}(\mathcal{O})$ as the maximum and minimum eigenvalues of $\mathcal{O}$, we have*

$$\left\|\mathcal{O}\right\|_{\mathfrak{M},\delta}\leq\frac{\lambda_{\max}(\mathcal{O})-\lambda_{\min}(\mathcal{O})}{2}. \tag{7.7}$$

3. *The optimization in Eq. (7.1) defining the minimax norm is convex. Moreover, there is an extreme point of the constraint set $\mathscr{C}(\mathfrak{M},\delta)$ that attains the maximum in Eq. (7.1).*

4. *$\left\|\mathcal{O}\right\|_{\mathfrak{M},\delta}$ is a monotonically decreasing function of δ.*

5. *$\left\|\mathcal{O}\right\|_{\mathfrak{M},\delta}$ is a monotonically decreasing function of L (number of POVMs) and $N_1,\ldots,N_L$ (number of repetitions of each POVM).*

6. $\|\mathcal{O}\|_{\mathfrak{M},\delta}$ is a continuous and convex function of $\mathcal{O}$.

Prop. 7.5.1 says that the minimax norm is invariant under translations of the observable by a constant matrix. From the point of view of estimation, this means that the estimation error only depends on the *spread* of the eigenvalues of the observable, and not the actual numerical values. An analogy of this property can be drawn with the freedom in shifting the reference energy of a Hamiltonian, which is commonplace in the analysis of physical systems.

Proof. 2. Dropping the constraint in Eq. (7.1), we obtain

$$\|\mathcal{O}\|_{\mathfrak{M},\delta} \leq \frac{1}{2} \max_{\chi_1,\chi_2 \in \mathscr{X}} (\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)) \leq \frac{1}{2}(\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O})). \tag{7.8}$$

3. From Prop. 7.4.1, we know that the constraint set $\mathscr{C}(\mathfrak{M},\delta)$ is compact and convex. Since the objective function $\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2)$ is affine in (χ_1,χ_2), the optimization in Eq. (7.1) is convex. Since the maximum of an affine function on a compact and convex set is attained at an extreme point of the set, there is an extreme point $(\chi_1^*,\chi_2^*) \in \mathscr{C}(\mathfrak{M},\delta)$ that attains the maximum in Eq. (7.1).

4. We have $\mathscr{C}(\mathfrak{M},\delta) \subseteq \mathscr{C}(\mathfrak{M},\delta')$ for $\delta \geq \delta'$. Therefore, the set over which $(\mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2))/2$ is maximized in Eq. (7.1) shrinks as δ increases, and subsequently, the value of the minimax norm decreases, proving the claim.

5. For a fixed L, as the value of $N_1,\ldots,N_L$ increases, the value of $(\mathrm{BC}_{\mathfrak{M}}(\chi_1,\chi_2))^N = \prod_{i=1}^L [\mathrm{BC}(p_{\chi_1}^{(i)}, p_{\chi_2}^{(i)})]^{N_i}$ decreases, as the Bhattacharyya coefficient is bounded between 0 and 1. Similarly, if we add additional POVMs (i.e., increase L), then the value of $(\mathrm{BC}_{\mathfrak{M}}(\chi_1,\chi_2))^N = \prod_{i=1}^L [\mathrm{BC}(p_{\chi_1}^{(i)}, p_{\chi_2}^{(i)})]^{N_i}$ decreases. As a result, the constraint set $\mathscr{C}(\mathfrak{M},\delta) = \{(\chi_1,\chi_2) \in \mathscr{X}\times\mathscr{X} \mid \prod_{i=1}^L [\mathrm{BC}(p_{\chi_1}^{(i)}, p_{\chi_2}^{(i)})]^{N_i} \geq \delta/2\}$ shrinks in size, from which the claim follows.

6. Since $\|\cdot\|_{\mathfrak{M},\delta}$ is a seminorm, it is convex. Since any real-valued convex function on a finite-dimensional vector space is continuous [8, Cor. 8.40], $\|\cdot\|_{\mathfrak{M},\delta}$ is continuous. □

We now show that there is a duality between the characteristic function of the constraint-difference set $\Delta\mathscr{C}(\mathfrak{M},\delta)$ and the minimax norm. Since there is a bijective correspondence between

a set and its characteristic function, this can be thought of as a duality between $\Delta\mathscr{C}(\mathfrak{M},\delta)$ and $\|\cdot\|_{\mathfrak{M},\delta}$.

Proposition 7.6. *Fix the measurement protocol $\mathfrak{M}$ and the confidence level $1-\delta \in (0,1)$. Then the following statements hold.*

1. *With $S_{\Delta\mathscr{C}(\mathfrak{M},\delta)}$ defined to be the support function of $\Delta\mathscr{C}(\mathfrak{M},\delta)$ according to Eq. (2.7), we have*

$$\|\cdot\|_{\mathfrak{M},\delta} = S_{\Delta\mathscr{C}(\mathfrak{M},\delta)}(\cdot). \tag{7.9}$$

2. *With $\chi_{\Delta\mathscr{C}(\mathfrak{M},\delta)}$ defined to be the characteristic function of $\Delta\mathscr{C}(\mathfrak{M},\delta)$ according to Eq. (2.6), the convex conjugate of $\|\cdot\|_{\mathfrak{M},\delta}$ is given as*

$$\|\cdot\|^*_{\mathfrak{M},\delta} = \chi_{\Delta\mathscr{C}(\mathfrak{M},\delta)}(\cdot). \tag{7.10}$$

3. *The minimax norm satisfies $\|\cdot\|^{**}_{\mathfrak{M},\delta} = \|\cdot\|_{\mathfrak{M},\delta}$.*

4. *The constraint-difference set can be expressed as*

$$\Delta\mathscr{C}(\mathfrak{M},\delta) = \{\mathcal{O}' \in \mathbb{S}_d \mid (\forall \mathcal{O} \in \mathbb{S}_d)\ \mathrm{Tr}(\mathcal{O}\mathcal{O}') \leq \|\mathcal{O}\|_{\mathfrak{M},\delta}\}. \tag{7.11}$$

Proof. 1. Follows from the definitions.

2. We always have $\chi^*_{\Delta\mathscr{C}(\mathfrak{M},\delta)} = S_{\Delta\mathscr{C}(\mathfrak{M},\delta)}$. Since $\Delta\mathscr{C}(\mathfrak{M},\delta)$ is closed and convex (Prop. 7.4), $\chi_{\Delta\mathscr{C}(\mathfrak{M},\delta)}$ is a proper lsc convex function, and thus self-dual [8, Thm. 13.37]. It follows that $(S_{\Delta\mathscr{C}(\mathfrak{M},\delta)})^* = (\chi^*_{\Delta\mathscr{C}(\mathfrak{M},\delta)})^* = \chi_{\Delta\mathscr{C}(\mathfrak{M},\delta)}$.

3. Follows from Eq. (7.9) and Eq. (7.10), and the fact that $\chi^*_{\Delta\mathscr{C}(\mathfrak{M},\delta)} = S_{\Delta\mathscr{C}(\mathfrak{M},\delta)}$.

4. We have

$$\begin{aligned}\{\mathscr{O}' \in \mathbb{S}_d \mid (\forall \mathscr{O} \in \mathbb{S}_d)\ \mathrm{Tr}(\mathscr{O}\mathscr{O}') \leq \|\mathscr{O}\|_{\mathfrak{M},\delta}\} &= \{\mathscr{O}' \in \mathbb{S}_d \mid \sup_{\mathscr{O}\in\mathbb{S}_d} (\mathrm{Tr}(\mathscr{O}\mathscr{O}') - \|\mathscr{O}\|_{\mathfrak{M},\delta}) \leq 0\} \\ &= \{\mathscr{O}' \in \mathbb{S}_d \mid \|\mathscr{O}'\|^*_{\mathfrak{M},\delta} \leq 0\} \\ &= \Delta\mathscr{C}(\mathfrak{M},\delta),\end{aligned} \tag{7.12}$$

where the second equality follows from the definition of convex conjugate, and the last equality follows from Eq. (7.10). □

The fact that $\|\cdot\|_{\mathfrak{M},\delta}$ is the support function of $\Delta\mathscr{C}(\mathfrak{M},\delta)$ gives us a geometric interpretation of the minimax norm. This is shown as a schematic in Fig. 2, which we adapt from [8, Fig. (7.1)].

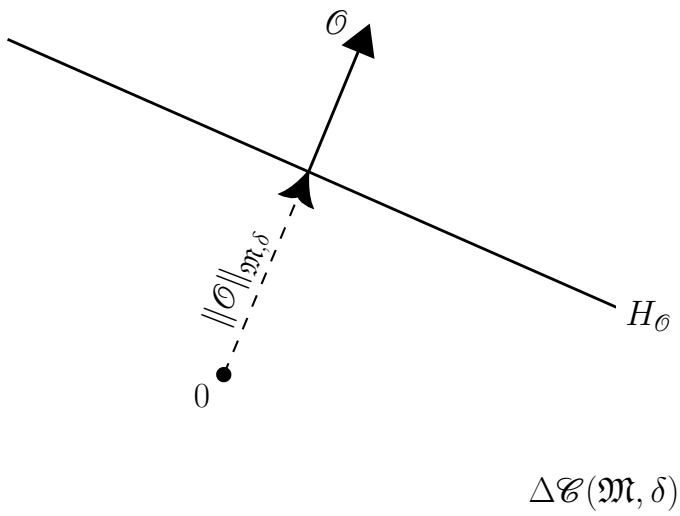

Figure 2: Interpretation of the minimax norm as the support function of the constraint-difference set $\Delta\mathscr{C}(\mathfrak{M},\delta)$. The observable $\mathscr{O}$ is normalized such that $\|\mathscr{O}\|_{\mathrm{HS}} = 1$. The minimax norm measures the distance of the supporting hyperplane $H_{\mathscr{O}} = \{\mathscr{O}' \in \mathbb{S}_d \mid \mathrm{Tr}(\mathscr{O}'\mathscr{O}) = \|\mathscr{O}\|_{\mathfrak{M},\delta}\}$ of $\Delta\mathscr{C}(\mathfrak{M},\delta)$ from the origin.

Below, we describe another useful symmetry property of the minimax norm. By a symmetry,

we mean a unitary that permutes the measurement settings in the measurement protocol. In particular, this includes unitaries that leave the measurement settings invariant. We make this notion precise below. In Def. 7.7, a permutation of $[M]$ is a bijective function $\sigma\colon [M] \to [M]$.

Definition 7.7 (Measurement symmetry). A unitary U is said to be a symmetry of the measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ if for every $i \in [L]$, there is some $j \in [L]$ and a permutation σ_i of $[M_i]$ such that

$$\{UE^{(i)}_{\sigma_i(k)}U^\dagger\}_{k=1}^{M_i} = \boldsymbol{E}^{(j)}, \tag{7.13}$$

and $N_i = N_j$. The set of symmetries of $\mathfrak{M}$ is denoted by $\mathscr{U}_{\mathfrak{M}}$. □

Observe that if i and j are related by a unitary as in Eq. (7.13), we must have $M_i = M_j$. It can be verified that $\mathscr{U}_{\mathfrak{M}}$ forms a group under multiplication. As an example, consider the measurement protocol where (the eigenvalue of) every Pauli operator is measured the same number of times. Since any unitary that takes Pauli operators to Pauli operators under conjugation is Clifford by definition, the set of symmetries of this measurement protocol is just the Clifford group. Another relevant example is a measurement protocol that corresponds to measuring a single POVM many times (this encompasses randomized measurements, for example). The measurement symmetries for such a protocol are those unitaries that leave the POVM invariant. Below, we show that the minimax norm is invariant under the action of a measurement symmetry.

Proposition 7.8. *Let $\mathcal{O}$ be any observable, $\mathfrak{M}$ be a measurement protocol, and $1-\delta$ be the confidence level. Then, for all $U \in \mathscr{U}_{\mathfrak{M}}$, the following statements hold.*

1. $\mathrm{BD}_{\mathfrak{M}}(U\chi_1U^\dagger, U\chi_2U^\dagger) = \mathrm{BD}_{\mathfrak{M}}(\chi_1, \chi_2)$ *for all* $\chi_1, \chi_2 \in \mathscr{X}$.

2. $U\mathscr{C}(\mathfrak{M}, \delta)U^\dagger \equiv \{(U\chi_1U^\dagger, U\chi_2U^\dagger) \mid (\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)\} = \mathscr{C}(\mathfrak{M}, \delta)$.

3. $U\Delta\mathscr{C}(\mathfrak{M}, \delta)U^\dagger \equiv \{U\mathcal{O}'U^\dagger \mid \mathcal{O}' \in \Delta\mathscr{C}(\mathfrak{M}, \delta)\} = \Delta\mathscr{C}(\mathfrak{M}, \delta)$.

4. $\left\|U\mathcal{O}U^\dagger\right\|_{\mathfrak{M},\delta} = \|\mathcal{O}\|_{\mathfrak{M},\delta}$.

Proof. 1. If $U \in \mathscr{U}_{\mathfrak{M}}$, then $U^\dagger \in \mathscr{U}_{\mathfrak{M}}$. By definition of measurement symmetry, and noting that POVMs are taken to be distinct in the definition of a measurement protocol, for all $i \in [L]$, there is exactly one $j \in [L]$ and a permutation σ_i of $[M_i]$, such that $N_i = N_j$ and $\mathrm{Tr}(E^{(i)}_{\sigma_i(k)} U\chi U^\dagger) = \mathrm{Tr}(U^\dagger E^{(i)}_{\sigma_i(k)} U\chi) = \mathrm{Tr}(E^{(j)}_k \chi)$ for all $k \in [M_i]$ and all $\chi \in \mathscr{X}$. Then, $\mathrm{BD}_{\mathfrak{M}}(U\chi_1 U^\dagger, U\chi_2 U^\dagger) = \mathrm{BD}_{\mathfrak{M}}(\chi_1, \chi_2)$ for all $\chi_1, \chi_2 \in \mathscr{X}$ follows from the definition of $\mathrm{BD}_{\mathfrak{M}}$.

2. If $(\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)$, then $(U\chi_1 U^\dagger, U\chi_2 U^\dagger) \in \mathscr{C}(\mathfrak{M}, \delta)$ by part 1. and the definition of $\mathscr{C}(\mathfrak{M}, \delta)$. Therefore, $U\mathscr{C}(\mathfrak{M}, \delta)U^\dagger \subseteq \mathscr{C}(\mathfrak{M}, \delta)$. Since we also have $U^\dagger \in \mathscr{U}_{\mathfrak{M}}$, for any $(\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)$, we have $(U^\dagger \chi_1 U, U^\dagger \chi_2 U) \in \mathscr{C}(\mathfrak{M}, \delta)$. Consequently, $(U(U^\dagger \chi_1 U)U^\dagger, U(U^\dagger \chi_1 U)U^\dagger) = (\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)$. Thus, we have $\mathscr{C}(\mathfrak{M}, \delta) \subseteq U\mathscr{C}(\mathfrak{M}, \delta)U^\dagger$.

3. By part 2., we have $(\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta)$ iff $(U\chi_1 U^\dagger, U\chi_2 U^\dagger) \in \mathscr{C}(\mathfrak{M}, \delta)$. It follows that $U\Delta\mathscr{C}(\mathfrak{M}, \delta)U^\dagger = \Delta\mathscr{C}(\mathfrak{M}, \delta)$ by definition of $\Delta\mathscr{C}(\mathfrak{M}, \delta)$.

4. By the definition of minimax norm, we have

$$\begin{aligned}
\left\| U\mathcal{O}U^\dagger \right\|_{\mathfrak{M},\delta} &= \frac{1}{2} \max \left\{ \mathrm{Tr}(U\mathcal{O}U^\dagger \chi_1) - \mathrm{Tr}(U\mathcal{O}U^\dagger \chi_2) \,\middle|\, (\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta) \right\} \\
&= \frac{1}{2} \max \left\{ \mathrm{Tr}(\mathcal{O}U^\dagger \chi_1 U) - \mathrm{Tr}(\mathcal{O}U^\dagger \chi_2 U) \,\middle|\, (\chi_1, \chi_2) \in \mathscr{C}(\mathfrak{M}, \delta) \right\} \\
&= \frac{1}{2} \max \left\{ \mathrm{Tr}(\mathcal{O}\chi_1') - \mathrm{Tr}(\mathcal{O}\chi_2') \,\middle|\, (\chi_1', \chi_2') \in U^\dagger \mathscr{C}(\mathfrak{M}, \delta) U \right\} \\
&= \frac{1}{2} \max \left\{ \mathrm{Tr}(\mathcal{O}\chi_1') - \mathrm{Tr}(\mathcal{O}\chi_2') \,\middle|\, (\chi_1', \chi_2') \in \mathscr{C}(\mathfrak{M}, \delta) \right\} \\
&= \|\mathcal{O}\|_{\mathfrak{M},\delta} \,,
\end{aligned} \tag{7.14}$$

where the third equality follows from the definition of $U^\dagger \mathscr{C}(\mathfrak{M}, \delta)U$ and the fourth equality follows from part 2. □

Finally, we prove a data-processing inequality for the minimax norm. Suppose that an experimentalist implements the POVM $\boldsymbol{E}$, and then classically processes the outcome observed after the measurement. This classical processing, which can be deterministic or random, is described mathematically as a classical channel. A classical channel describes a process that takes an input probability distribution on $[M]$ to an output probability distribution on $[M']$. We can think of a

classical channel in terms of its transition matrix $\mathcal{N}$, which is an $M' \times M$ matrix with non-negative entries, where the columns sum to 1. If $p_{\boldsymbol{E},\chi}$ is the distribution after measuring $\boldsymbol{E}$ in the state χ, then the post-processing step is described by a classical channel $\mathcal{N}$ that acts on $p_{\boldsymbol{E},\chi}$. This gives rise to a new POVM $\boldsymbol{F}$ with M' elements, where $F_j = \sum_{k=1}^{M} \mathcal{N}_{jk} E_k$ for $j \in [M']$. It can be verified that $p_{\boldsymbol{F},\chi} = \mathcal{N} p_{\boldsymbol{E},\chi}$ for all $\chi \in \mathscr{X}$. Thus, the POVM $\boldsymbol{F}$ is the effective POVM that describes the measurement $\boldsymbol{E}$ and the classical post-processing. We show that performing a classical data processing on the measurement outcomes does not help with estimation.

Proposition 7.9 (Data-processing inequality for minimax norm). *Let $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}$ be a measurement protocol, and let $\mathfrak{M}^{\mathrm{pp}} = \{(\boldsymbol{F}^{(i)}, N_i)\}$, where for all $i \in [L]$, the POVM $\boldsymbol{F}^{(i)}$ is the effective POVM describing the measurement $\boldsymbol{E}^{(i)}$ followed by a classical post-processing. Then, for all observables $\mathcal{O}$ and all $1 - \delta \in (0, 1)$, we have*

$$\|\mathcal{O}\|_{\mathfrak{M}^{\mathrm{pp}},\delta} \geq \|\mathcal{O}\|_{\mathfrak{M},\delta} \,. \tag{7.15}$$

Proof. Since for all $i \in [L]$, $\boldsymbol{F}^{(i)}$ is obtained by classically post-processing $\boldsymbol{E}^{(i)}$, there is some channel $\mathcal{N}^{(i)}$ such that $p_{\boldsymbol{F}^{(i)},\chi} = \mathcal{N}^{(i)} p_{\boldsymbol{E}^{(i)},\chi}$ for all $\chi \in \mathscr{X}$. From [107, Ex. (9.2.8)], we have that for all quantum channels $\mathcal{Q}$ and all states χ_1, χ_2, we have $F(\mathcal{Q}(\chi_1), \mathcal{Q}(\chi_2)) \geq F(\chi_1, \chi_2)$. Since a classical channel acting on discrete probability distributions is a special case of a quantum channel acting on quantum states (see [107, Sec. (4.6.4)]), we have $\mathrm{FC}(p_{\boldsymbol{F}^{(i)},\chi_1}, p_{\boldsymbol{F}^{(i)},\chi_2}) \geq \mathrm{FC}(p_{\boldsymbol{E}^{(i)},\chi_1}, p_{\boldsymbol{E}^{(i)},\chi_2})$ for all $i \in [L]$ and all χ_1, χ_2. It follows that $\mathrm{FC}_{\mathfrak{M}^{\mathrm{pp}}}(\chi_1, \chi_2) \geq \mathrm{FC}_{\mathfrak{M}}(\chi_1, \chi_2)$ for all χ_1, χ_2. Then, Eq. (7.15) follows from the expression for minimax norm given in Eq. (7.2). □

As a consequence of Prop. 7.9, we can conclude that forgetting the sampled POVM in a randomized measurement protocol does not help with estimation.

7.2 Lower bound on the error for a given measurement protocol

In this section, we derive a lower bound on the error of learning the expectation value of an observable for a fixed measurement protocol using the minimax norm. We also show that our lower

bounds can be achieved by $\mathbb{TOOL}$ to within a small factor, thus proving that our bounds are tight, and that we have a constructive estimation procedure for achieving the lower bound to within a constant factor.

We first reinterpret the estimation error of $\mathbb{TOOL}$ given in Box 3 as the minimax norm of the observable whose expectation value we wish to learn with respect to the perturbed measurement protocol defined in Def. 5.4.

Proposition 7.10. *1. The estimation error ε_* of the estimator constructed by $\mathbb{TOOL}$ with parameter $\epsilon_\circ > 0$ for learning the expectation value of $\mathcal{O}$ using outcomes of $\mathfrak{M}$ to a confidence level of $1-\delta \in (0,1)$ can be expressed as*

$$\varepsilon_* = \|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta}\,, \tag{7.16}$$

where $\mathfrak{M}(\epsilon_\circ)$ is the perturbed measurement protocol defined in Def. 5.4.

2. If $\widehat{\mathcal{O}}_*$ is the estimator constructed by $\mathbb{TOOL}$, then we have the guarantee

$$\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}\left(|\widehat{\mathcal{O}}_* - \mathrm{Tr}(\mathcal{O}\rho)| \leq \|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta}\right) \geq 1-\delta \tag{7.17}$$

for all $\rho \in \mathscr{X}$.

3. For $1-\delta \in (0.75, 1)$, we have

$$\mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta) \leq \|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta} \leq \frac{2\log(2/\delta)}{\log(1/(4\delta))}\mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta), \tag{7.18}$$

where $\mathscr{R}_*(\mathcal{O}, \mathfrak{M}, \delta)$ is the minimax optimal risk for learning the expectation value of $\mathcal{O}$ using $\mathfrak{M}$, defined in Eq. (5.2).

Proof. Eq. (7.16) follows from Eq. (5.18) and the definition of minimax norm in Eq. (7.1). Eq. (7.17) follows from Eq. (5.13) and Eq. (7.16). Eq. (7.18) follows from Eq. (7.16) and Eq. (5.16). □

We begin by showing that $\|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta}$ converges to $\|\mathcal{O}\|_{\mathfrak{M},\delta}$ from above, as $\epsilon_\circ \to 0$. For simplifying notation, in Lem. 7.11, we take $\lambda = \epsilon_\circ/(1+\epsilon_\circ)$ in the definition of $\mathfrak{M}(\epsilon_\circ)$ in Def. 5.4.

Lemma 7.11. *Let $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$ be a measurement protocol. For $\lambda \in [0,1]$, define the measurement protocol $\mathfrak{M}(\lambda)$ to consist of the POVMs*

$$\left\{(1-\lambda)E_1^{(i)} + \lambda\frac{\mathbb{I}}{M_i}, \ldots, (1-\lambda)E_{M_i}^{(i)} + \lambda\frac{\mathbb{I}}{M_i}\right\} \tag{7.19}$$

for $i \in [L]$, where ith POVM is measured N_i times. Then, for any observable $\mathcal{O}$ and confidence level $1-\delta \in (0,1)$, the following statements hold.

1. *For all $\lambda \in [0,1]$, we have $\|\mathcal{O}\|_{\mathfrak{M}(\lambda),\delta} \geq \|\mathcal{O}\|_{\mathfrak{M},\delta}$.*

2. *For all $C \geq 0$, we have $\lim_{\lambda\to 0}\|\mathcal{O}\|_{\mathfrak{M}(\lambda),\delta-C\lambda} = \|\mathcal{O}\|_{\mathfrak{M},\delta}$.*

Proof. 1. Let p, q be discrete probability distributions over M symbols. Let $e = (1/M, \ldots, 1/M)$ denote the uniform probability distribution over M symbols. Then, by joint concavity of the Bhattacharyya coefficient [106, Corollary 3.26], we have

$$\mathrm{BC}((1-\lambda)p + \lambda e, (1-\lambda)q + \lambda e) \geq (1-\lambda)\mathrm{BC}(p,q) + \lambda\mathrm{BC}(e,e) \geq \mathrm{BC}(p,q), \tag{7.20}$$

where in the last step, we used $1 = \mathrm{BC}(e,e) \geq \mathrm{BC}(p,q)$.

Now, given state χ, denote $p_\chi^{(i)} = (\mathrm{Tr}(E_1^{(i)}\chi), \ldots, \mathrm{Tr}(E_{M_i}^{(i)}\chi))$ obtained from the POVM $\{E_k^{(i)}\}_{k=1}^{M_i}$. Then, if $p_\chi^{(i)}(\lambda)$ is the distribution obtained from the perturbed POVM in Eq. (7.19), we can write $p_\chi^{(i)}(\lambda) = (1-\lambda)p_\chi^{(i)} + \lambda e^{(i)}$, where $e^{(i)} = (1/M_i, \ldots, 1/M_i)$. Then, we have $B(p_{\chi_1}^{(i)}(\lambda), p_{\chi_2}^{(i)}(\lambda)) \geq B(p_{\chi_1}^{(i)}, p_{\chi_2}^{(i)})$ for all $i \in [L]$, and consequently,

$$\mathrm{BC}_{\mathfrak{M}(\lambda)}(\chi_1,\chi_2) = \prod_{i=1}^{L}[B(p_{\chi_1}^{(i)}(\lambda), p_{\chi_2}^{(i)}(\lambda))]^{N_i/N} \geq \prod_{i=1}^{L}[B(p_{\chi_1}^{(i)}, p_{\chi_2}^{(i)})]^{N_i/N} = \mathrm{BC}_{\mathfrak{M}}(\chi_1,\chi_2). \tag{7.21}$$

It follows from Eq. (7.2) that $\|\mathcal{O}\|_{\mathfrak{M}(\lambda),\delta} \geq \|\mathcal{O}\|_{\mathfrak{M},\delta}$.

2. For $C \geq 0$, define $\lambda_\circ = \min\{\delta/2C, 1\}$ and $\Lambda = [0, \lambda_\circ]$. Observe that $\delta - \lambda C \in (0,1)$ for all

$\lambda \in \Lambda$. Consider the set-valued function $\mathscr{C}\colon \Lambda \to 2^{\mathscr{X}\times\mathscr{X}}$ defined as

$$\mathscr{C}(\lambda) = \left\{(\chi_1,\chi_2) \in \mathscr{X}\times\mathscr{X} \mid (\mathrm{BC}_{\mathfrak{M}(\lambda)}(\chi_1,\chi_2))^N \geq \frac{\delta - C\lambda}{2}\right\}, \tag{7.22}$$

where $2^{\mathscr{X}\times\mathscr{X}}$ is the power set of $\mathscr{X}\times\mathscr{X}$ and $N = \sum_{i=1}^{L} N_i$ is the total number of samples. For each $\lambda \in \Lambda$, $\mathscr{C}(\lambda) = \mathscr{C}(\mathfrak{M}(\lambda), \delta - C\lambda)$ is the constraint set (Eq. (7.4)) over which the optimization defining the minimax norm $\|\cdot\|_{\mathfrak{M}(\lambda),\delta-C\lambda}$ is performed. Observe that $\mathscr{C}(\lambda)$ is non-empty for each $\lambda \in \Lambda$ because $\mathrm{BC}_{\mathfrak{M}(\lambda)}(\chi,\chi) = 1$ for any density matrix χ.

The graph of the set-valued function $\mathscr{C}$ is given as [3, Def. (17.9)]

$$\begin{aligned}\operatorname{gr}\mathscr{C} &= \{(\lambda,(\chi_1,\chi_2)) \in \Lambda\times(\mathscr{X}\times\mathscr{X}) \mid (\chi_1,\chi_2)\in\mathscr{C}(\lambda)\}\\ &= \left\{(\lambda,(\chi_1,\chi_2)) \in \Lambda\times(\mathscr{X}\times\mathscr{X}) \mid (\mathrm{BC}_{\mathfrak{M}(\lambda)}(\chi_1,\chi_2))^N + C\lambda/2 \geq \delta/2\right\}.\end{aligned} \tag{7.23}$$

Since $(\mathrm{BC}_{\mathfrak{M}(\lambda)}(\chi_1,\chi_2))^N + C\lambda/2$ is a continuous function of $\lambda\in\Lambda$ and $(\chi_1,\chi_2)\in\mathscr{X}\times\mathscr{X}$, $\operatorname{gr}\mathscr{C}$ is a closed subset of $\Lambda\times(\mathscr{X}\times\mathscr{X})$. Now, the set-valued function $\mathscr{C}$ is said to be upper hemicontinuous if for every closed subset F of $\mathscr{X}\times\mathscr{X}$, the set $\{\lambda\in\Lambda \mid \mathscr{C}(\lambda)\cap F \neq \varnothing\}$ is a closed subset of Λ [3, Lem. (17.4)]. Since $\mathscr{X}\times\mathscr{X}$ is a compact subset of a Hilbert space, we have from [3, Thm. (17.11)] that $\mathscr{C}$ is an upper hemicontinuous set-valued function. Moreover, for each $\lambda\in\Lambda$, $\mathscr{C}(\lambda)$ is a non-empty compact set (see Prop. 7.4). Since $(\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2))/2$ is a continuous function of $(\lambda,(\chi_1,\chi_2)) \in \Lambda\times(\mathscr{X}\times\mathscr{X})$, the minimax norm $\|\mathscr{O}\|_{\mathfrak{M}(\lambda),\delta-C\lambda} = \max_{(\chi_1,\chi_2)\in\mathscr{C}(\lambda)}(\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2))/2$ is an upper semicontinuous function of $\lambda\in\Lambda$ [3, Lem. (17.30)]. By definition of upper semicontinuity [3, Lem. (2.42)], we have $\limsup_{\lambda\to 0}\|\mathscr{O}\|_{\mathfrak{M}(\lambda),\delta-C\lambda} \leq \|\mathscr{O}\|_{\mathfrak{M},\delta}$, where we used the fact that $\mathfrak{M}(\lambda=0) = \mathfrak{M}$. Then, since $\|\mathscr{O}\|_{\mathfrak{M},\delta} \leq \|\mathscr{O}\|_{\mathfrak{M},\delta-C\lambda} \leq \|\mathscr{O}\|_{\mathfrak{M}(\lambda),\delta-C\lambda}$, from Prop. 7.5.4 and Lem. 7.11.1, we have the chain of inequalities

$$\|\mathscr{O}\|_{\mathfrak{M},\delta} \leq \liminf_{\lambda\to 0}\|\mathscr{O}\|_{\mathfrak{M}(\lambda),\delta-C\lambda} \leq \limsup_{\lambda\to 0}\|\mathscr{O}\|_{\mathfrak{M}(\lambda),\delta-C\lambda} \leq \|\mathscr{O}\|_{\mathfrak{M},\delta}. \tag{7.24}$$

This implies $\lim_{\lambda\to 0}\|\mathscr{O}\|_{\mathfrak{M}(\lambda),\delta-C\lambda} = \|\mathscr{O}\|_{\mathfrak{M},\delta}$. □

Next, we show that we can lower bound the minimax optimal risk defined in Eq. (5.2) for the

measurement protocol $\mathfrak{M}$ by minimizing the minimax optimal risk for $\mathfrak{M}(\epsilon_\circ)$ over all $\epsilon_\circ > 0$.

Lemma 7.12. *Let $\mathfrak{M}$ be a fixed measurement protocol, and for $\epsilon_\circ > 0$, let $\mathfrak{M}(\epsilon_\circ)$ be the perturbed measurement protocol defined in Def. 5.4. Then, for learning the expectation value of the observable $\mathcal{O}$ with confidence $1-\delta \in (0,1)$, we have*

$$\inf_{\epsilon_\circ > 0} \mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta) \leq \mathscr{R}_*(\mathcal{O}, \mathfrak{M}, \delta). \tag{7.25}$$

Proof. We begin by noting that for any given family of non-empty sets $\{F_i\}_{i\in\mathscr{I}}$ indexed by some set $\mathscr{I}$ and any function $f\colon \cup_{i\in\mathscr{I}} F_i \to \mathbb{R}$, we have

$$\inf_{r\in\cup_{i\in\mathscr{I}} F_i} f(r) = \inf_{i\in\mathscr{I}} \inf_{r\in F_i} f(r). \tag{7.26}$$

To see this, observe that for all $i \in \mathscr{I}$, we have $\inf_{r\in\cup_{i\in\mathscr{I}}F_i} f(r) \leq \inf_{r\in F_i} f(r)$, so that $\inf_{r\in\cup_{i\in\mathscr{I}}F_i} f(r) \leq \inf_{i\in\mathscr{I}} \inf_{r\in F_i} f(r)$. On the other hand, for each $r \in \cup_{i\in\mathscr{I}} F_i$, there is some $i \in \mathscr{I}$ such that $r \in F_i$, so that $\inf_{i\in\mathscr{I}} \inf_{r\in F_i} f(r) \leq \inf_{r\in F_i} f(r) \leq f(r)$, and therefore, $\inf_{i\in\mathscr{I}} \inf_{r\in F_i} f(r) \leq \inf_{r\in\cup_{i\in\mathscr{I}}F_i} f(r)$. In particular, when $F_i \subseteq \mathbb{R}$ for all $i \in \mathscr{I}$, we have $\inf \cup_{i\in\mathscr{I}} F_i = \inf_{i\in\mathscr{I}} \inf F_i$.

To proceed, recall that $\mathbb{P}_{\mathfrak{M},\sigma}$ denotes the probability over outcomes defined by the measurement protocol $\mathfrak{M}$ and the state σ. Given an estimator $\widehat{\mathcal{O}}$ and error $\varepsilon > 0$, define the set $A_{\widehat{\mathcal{O}}}(\varepsilon) = \{|\widehat{\mathcal{O}} - \mathrm{Tr}(\mathcal{O}\sigma)| \leq \varepsilon\}$. Then, using Eq. (7.26) along with Eq. (5.1) and Eq. (5.2), we can write

$$\mathscr{R}_*(\mathcal{O}, \mathfrak{M}, \delta) = \inf \bigcup_{\widehat{\mathcal{O}}} \left\{ \varepsilon > 0 \mid \inf_\sigma \mathbb{P}_{\mathfrak{M},\sigma}\left(A_{\widehat{\mathcal{O}}}(\varepsilon)\right) > 1-\delta \right\}, \tag{7.27}$$

and

$$\inf_{\epsilon_\circ>0} \mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta) = \inf \bigcup_{(\epsilon_\circ, \widehat{\mathcal{O}})} \left\{ \varepsilon > 0 \mid \inf_\sigma \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}\left(A_{\widehat{\mathcal{O}}}(\varepsilon)\right) > 1-\delta \right\}, \tag{7.28}$$

where the union in the second equation is over all estimators $\widehat{\mathcal{O}}$ and all positive numbers $\epsilon_\circ > 0$.

Since $\mathbb{P}_{\mathfrak{M},\sigma}$ and $\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}$ are product distributions, by subadditivity of total variation distance

for product distributions, for all states σ and all $\epsilon_\circ > 0$, we have

$$\begin{aligned}
\left\|\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma} - \mathbb{P}_{\mathfrak{M},\sigma}\right\|_{\mathrm{TV}} &\leq \frac{1}{2}\sum_{i=1}^{L} N_i \sum_{j=1}^{M_i} \left| \frac{\mathrm{Tr}(E_j^{(i)}\sigma) + \frac{\epsilon_\circ}{M_i}}{1+\epsilon_\circ} - \mathrm{Tr}(E_j^{(i)}\sigma) \right| \\
&\leq \frac{\epsilon_\circ}{1+\epsilon_\circ}\frac{1}{2}\sum_{i=1}^{L} N_i \sum_{j=1}^{M_i} \left| \mathrm{Tr}(E_j^{(i)}\sigma) - \frac{1}{M_i} \right| \\
&\leq N\frac{\epsilon_\circ}{1+\epsilon_\circ} \\
&< N\epsilon_\circ,
\end{aligned} \tag{7.29}$$

where $N = \sum_{i=1}^{L} N_i$ is the total number of samples, and the last inequality follows from the fact that $\sum_{j=1}^{M_i} |\mathrm{Tr}(E_j^{(i)}\sigma) - 1/M_i| \leq 2$. By the definition of total variation distance that $\left\|\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma} - \mathbb{P}_{\mathfrak{M},\sigma}\right\|_{\mathrm{TV}} = \sup_H |\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}(H) - \mathbb{P}_{\mathfrak{M},\sigma}(H)|$, where the supremum is over all events H, we have $|\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}(A_{\widehat{\mathscr{O}}}(\varepsilon)) - \mathbb{P}_{\mathfrak{M},\sigma}(A_{\widehat{\mathscr{O}}}(\varepsilon))| < N\epsilon_\circ$ for all $\epsilon_\circ > 0$, all $\varepsilon > 0$, all estimators $\widehat{\mathscr{O}}$, and all states σ.

Now, consider an arbitrary element $\varepsilon' \in \bigcup_{\widehat{\mathscr{O}}} \left\{\varepsilon > 0 \mid \inf_\sigma \mathbb{P}_{\mathfrak{M},\sigma}\left(A_{\widehat{\mathscr{O}}}(\varepsilon)\right) > 1-\delta\right\}$. We claim that $\varepsilon' \in \bigcup_{(\epsilon_\circ,\widehat{\mathscr{O}})} \left\{\varepsilon > 0 \mid \inf_\sigma \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}\left(A_{\widehat{\mathscr{O}}}(\varepsilon)\right) > 1-\delta\right\}$. For if this does not hold, then for all $\epsilon_\circ > 0$ and all estimators $\widehat{\mathscr{O}}$, we have $\inf_\sigma \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}(A_{\widehat{\mathscr{O}}}(\varepsilon')) \leq 1-\delta$. Since $\inf_\sigma \mathbb{P}_{\mathfrak{M},\sigma}(A_{\widehat{\mathscr{O}}}(\varepsilon')) \leq \inf_\sigma \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}(A_{\widehat{\mathscr{O}}}(\varepsilon')) + N\epsilon_\circ$, we obtain $\inf_\sigma \mathbb{P}_{\mathfrak{M},\sigma}(A_{\widehat{\mathscr{O}}}(\varepsilon')) \leq 1-\delta+N\epsilon_\circ$. Because this inequality holds for all $\epsilon_\circ > 0$, we obtain $\inf_\sigma \mathbb{P}_{\mathfrak{M},\sigma}(A_{\widehat{\mathscr{O}}}(\varepsilon')) \leq 1-\delta$ for all estimators $\widehat{\mathscr{O}}$, contradicting the assumption that $\varepsilon' \in \bigcup_{\widehat{\mathscr{O}}} \left\{\varepsilon > 0 \mid \inf_\sigma \mathbb{P}_{\mathfrak{M},\sigma}\left(A_{\widehat{\mathscr{O}}}(\varepsilon)\right) > 1-\delta\right\}$. Consequently, we have

$$\bigcup_{\widehat{\mathscr{O}}} \left\{\varepsilon > 0 \mid \inf_\sigma \mathbb{P}_{\mathfrak{M},\sigma}\left(A_{\widehat{\mathscr{O}}}(\varepsilon)\right) > 1-\delta\right\} \subseteq \bigcup_{(\epsilon_\circ,\widehat{\mathscr{O}})} \left\{\varepsilon > 0 \mid \inf_\sigma \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\sigma}\left(A_{\widehat{\mathscr{O}}}(\varepsilon)\right) > 1-\delta\right\}, \tag{7.30}$$

from which the claim follows. □

Using Lem. 7.11 and Lem. 7.12, we show that the minimax norm lower bounds the estimation error up to a small factor. We also show that $\mathbb{TOOL}$ can achieve this lower bound to within a small constant factor.

Theorem 7.13. *Every estimation protocol that learns the expectation value of the observable $\mathcal{O}$ using outcomes of $\mathfrak{M}$ to within an error of $\varepsilon > 0$ with a confidence level of $1-\delta \in (0.75, 1)$ satisfies*

$$\varepsilon \geq \mathscr{c}(\delta) \left\| \mathcal{O} \right\|_{\mathfrak{M},\delta}, \tag{7.31}$$

where

$$\mathscr{c}(\delta) = \frac{\log_2(1/\delta) - 2}{2\log_2(1/\delta) + 2}. \tag{7.32}$$

Moreover, denoting N to be the total number of samples used by $\mathfrak{M}$, for all $\eta > 0$, there is some $0 < \epsilon_\circ \leq \delta/(2N)$, such that the estimator $\widehat{\mathcal{O}}_*$ constructed by $\mathbb{TOOL}$ with parameter $\epsilon_\circ$ for a confidence level of $1-(\delta - N\epsilon_\circ)$ satisfies

$$\mathbb{P}_{\mathfrak{M},\rho}\left(|\widehat{\mathcal{O}}_* - \mathrm{Tr}(\mathcal{O}\rho)| \leq (1+\eta) \left\| \mathcal{O} \right\|_{\mathfrak{M},\delta} \right) > 1 - \delta \tag{7.33}$$

for all states ρ.

Proof. For all $\epsilon_\circ > 0$ and $\delta \in (0, 0.25)$, we know from Eq. (7.18) that

$$\left\| \mathcal{O} \right\|_{\mathfrak{M}(\epsilon_\circ),\delta} \leq \frac{1}{\mathscr{c}(\delta)} \mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta), \tag{7.34}$$

where we used the fact that

$$\frac{2\log(2/\delta)}{\log(1/(4\delta))} = \frac{2(1+\log_2(1/\delta))}{\log_2(1/\delta) - 2} = \frac{1}{\mathscr{c}(\delta)}. \tag{7.35}$$

Minimizing this inequality over all $\epsilon_\circ > 0$, we obtain

$$\inf_{\epsilon_\circ > 0} \left\| \mathcal{O} \right\|_{\mathfrak{M}(\epsilon_\circ),\delta} \leq \frac{1}{\mathscr{c}(\delta)} \inf_{\epsilon_\circ > 0} \mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta). \tag{7.36}$$

Now, let $\lambda = \epsilon_\circ/(1+\epsilon_\circ)$, so that $\mathfrak{M}(\lambda)$ defined in Lem. 7.11 coincides with $\mathfrak{M}(\epsilon_\circ)$ defined in

Def. 5.4. Then, from Lem. 7.11.1, we have

$$\|\mathcal{O}\|_{\mathfrak{M},\delta} \leq \inf_{\epsilon_\circ > 0} \|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta}\,. \tag{7.37}$$

Therefore, we obtain

$$\|\mathcal{O}\|_{\mathfrak{M},\delta} \leq \inf_{\epsilon_\circ > 0} \|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta} \leq \frac{1}{\mathscr{c}(\delta)} \inf_{\epsilon_\circ > 0} \mathscr{R}_*(\mathcal{O}, \mathfrak{M}(\epsilon_\circ), \delta) \leq \frac{1}{\mathscr{c}(\delta)} \mathscr{R}_*(\mathcal{O}, \mathfrak{M}, \delta) \leq \frac{1}{\mathscr{c}(\delta)} \varepsilon, \tag{7.38}$$

where the second inequality follows from Eq. (7.36), the third inequality follows from Lem. 7.12, and the last inequality follows from Def. 5.2.

Now, we show that $\mathbb{TOOL}$ achieves this lower bound to a small constant factor. Fix $\eta > 0$. Since $\lim_{\epsilon_\circ \to 0} \|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta - N\epsilon_\circ} = \|\mathcal{O}\|_{\mathfrak{M},\delta}$ (Lem. 7.11.2), we can find a small enough $0 < \epsilon_\circ \leq \delta/(2N)$ such that $\|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta - N\epsilon_\circ} \leq (1+\eta)\,\|\mathcal{O}\|_{\mathfrak{M},\delta}$. Note that this bound holds even when $\|\mathcal{O}\|_{\mathfrak{M},\delta} = 0$, since by Prop. 7.2, we have $\|\mathcal{O}\|_{\mathfrak{M},\delta} = 0$ iff $\mathcal{O} = c\mathbb{1}$ iff $\|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta - N\epsilon_\circ} = 0$. Let $\widehat{\mathcal{O}}_*$ be the estimator constructed by $\mathbb{TOOL}$ with parameter $\epsilon_\circ$, for a confidence level of $1 - (\delta - N\epsilon_\circ)$. Then, from Eq. (7.17), we have

$$\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}\left(|\widehat{\mathcal{O}}_* - \mathrm{Tr}(\mathcal{O}\rho)| \leq (1+\eta)\,\|\mathcal{O}\|_{\mathfrak{M},\delta}\right) \geq \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}\left(|\widehat{\mathcal{O}}_* - \mathrm{Tr}(\mathcal{O}\rho)| \leq \|\mathcal{O}\|_{\mathfrak{M}(\epsilon_\circ),\delta - N\epsilon_\circ}\right) \geq 1 - (\delta - N\epsilon_\circ) \tag{7.39}$$

for all states ρ. From Eq. (7.29), we know that $\left\|\mathscr{P}_{\mathfrak{M}(\epsilon_\circ),\rho} - \mathbb{P}_{\mathfrak{M},\rho}\right\|_{\mathrm{TV}} \leq N\epsilon_\circ/(1+\epsilon_\circ)$ for all ρ. It follows from the definition of total variation distance that

$$\begin{aligned}
\mathbb{P}_{\mathfrak{M},\rho}(\{|\widehat{\mathcal{O}}_* - \mathrm{Tr}(\mathcal{O}\rho)| \leq (1+\eta)\,\|\mathcal{O}\|_{\mathfrak{M},\delta}\}) &\geq \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}(\{|\widehat{\mathcal{O}}_* - \mathrm{Tr}(\mathcal{O}\rho)| \leq (1+\eta)\,\|\mathcal{O}\|_{\mathfrak{M},\delta}\}) - N\frac{\epsilon_\circ}{1+\epsilon_\circ} \\
&\geq 1 - \delta + \frac{N\epsilon_\circ^2}{1+\epsilon_\circ} \\
&> 1 - \delta
\end{aligned} \tag{7.40}$$

for all ρ. □

Since $\eta > 0$ can be made arbitrarily small in Eq. (7.33), the lower bound in Eq. (7.31) is tight

to within a factor of $1/\mathscr{c}(\delta)$. For confidence levels 90% or more, we have $\mathscr{c}(\delta) \in (0.15, 0.5)$, and therefore, the lower bound in Eq. (7.31) is fairly tight.

7.3 General lower bound on the estimation error

Our goal in this section is to derive a lower bound on the error of estimating the expectation value of an observable that does not depend on the measurement protocol that is implemented. We also show that our lower bound is tight, in the sense that there is some measurement protocol (and estimation procedure) for which we can achieve the lower bound on the estimation error to within a constant factor. We find that measuring in the eigenbasis of the observable gives optimal performance for learning the expectation value of that observable. This result should be intuitive and familiar to many readers, and we give a formal proof using the results derived in the previous section.

Since the minimax norm gives a lower bound on the estimation error for any given measurement protocol, it suffices to obtain a lower bound on the minimax norm that holds for all measurement protocols using a fixed number of samples.

Lemma 7.14. *Fix the observable $\mathcal{O}$ and confidence level $1-\delta \in (0,1)$. Then, for all measurement protocols $\mathfrak{M}$ using N samples, we have*

$$\|\mathcal{O}\|_{\mathfrak{M},\delta} \geq \frac{(\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O}))}{2} \sqrt{1 - \left(\frac{\delta}{2}\right)^{2/N}}. \tag{7.41}$$

Proof. We prove this statement by generalizing the strategy in Ref. [92, Thm. II.1]. In the proof below, we denote $\lambda_{\max}(\mathcal{O})$ as $\lambda_{\max}$ and $\lambda_{\min}(\mathcal{O})$ as $\lambda_{\min}$ for simplicity. $\mathcal{O}$ is a multiple of identity iff $\lambda_{\max} = \lambda_{\min}$, in which case Eq. (7.41) holds trivially. Thus, we assume for the rest of the proof that $\mathcal{O}$ is not a multiple of identity.

Since by Prop. 3.7, $\mathrm{FC}_{\mathfrak{M}}(\chi_1, \chi_2) \geq F(\chi_1, \chi_2)$ for all $\chi_1, \chi_2 \in \mathscr{X}$, we have from Eq. (7.2) the lower bound

$$\|\mathcal{O}\|_{\mathfrak{M},\delta} \geq \frac{1}{2} \max_{\chi_1,\chi_2 \in \mathscr{X}} \left\{ \mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2) \,\middle|\, F(\chi_1,\chi_2) \geq \left(\frac{\delta}{2}\right)^{\frac{2}{N}} \right\}, \tag{7.42}$$

where $N = \sum_{i=1}^{L} N_i$ is the total number of samples. We proceed to evaluating this lower bound. Denote $\gamma = (\delta/2)^{2/N}$, so that the constraint in the above equation becomes $F(\chi_1, \chi_2) \geq \gamma$. Recall that the trace distance between two states χ_1, χ_2 can be expressed as [107, Lem. (9.1.1)]

$$\|\chi_1 - \chi_2\|_{\mathrm{tr}} = \max_{0 \leq \Lambda \leq \mathbb{I}} \mathrm{Tr}(\Lambda(\chi_1 - \chi_2)). \tag{7.43}$$

Therefore, we have

$$\begin{aligned} \mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2) &\leq (\lambda_{\max} - \lambda_{\min}) \|\chi_1 - \chi_2\|_{\mathrm{tr}} && (7.44) \\ &\leq (\lambda_{\max} - \lambda_{\min}) \sqrt{1 - F(\chi_1, \chi_2)} \\ &\leq (\lambda_{\max} - \lambda_{\min}) \sqrt{1 - \gamma}, && (7.45) \end{aligned}$$

where the second inequality follows from the Fuchs-van de Graaf inequality (Eq. (3.35)), and the last inequality holds when $F(\chi_1, \chi_2) \geq \gamma$.

We show that the upper bound in Eq. (7.45) can be achieved by explicitly constructing the density matrices χ_1^* and χ_2^* achieving this bound and satisfying $F(\chi_1^*, \chi_2^*) \geq \gamma$. For this purpose, let $|\lambda_{\min}\rangle$ and $|\lambda_{\max}\rangle$ denote orthonormal eigenvectors corresponding to the eigenvalues $\lambda_{\min}$ and $\lambda_{\max}$ respectively. Define

$$\begin{aligned} \chi_1^* &= \frac{1 + \sqrt{1-\gamma}}{2} |\lambda_{\max}\rangle \langle\lambda_{\max}| + \frac{1 - \sqrt{1-\gamma}}{2} |\lambda_{\min}\rangle \langle\lambda_{\min}|, \\ \chi_2^* &= \frac{1 - \sqrt{1-\gamma}}{2} |\lambda_{\max}\rangle \langle\lambda_{\max}| + \frac{1 + \sqrt{1-\gamma}}{2} |\lambda_{\min}\rangle \langle\lambda_{\min}|. \end{aligned} \tag{7.46}$$

Observe that χ_1^* and χ_2^* are diagonal in the eigenbasis of $\mathcal{O}$. Since $0 < \gamma < 1$, the diagonal entries of these matrices are non-negative and they sum to 1, so that χ_1^* and χ_2^* are density matrices. Since they are diagonal in the same basis, by Prop. 3.6.6, the fidelity between them is given by

$$F(\chi_1^*, \chi_2^*) = \left(2 \sqrt{\left(\frac{1 + \sqrt{1-\gamma}}{2} \right) \left(\frac{1 - \sqrt{1-\gamma}}{2} \right)} \right)^2 = \gamma. \tag{7.47}$$

Furthermore, we have

$$\mathrm{Tr}(\mathcal{O}\chi_1^*) - \mathrm{Tr}(\mathcal{O}\chi_2^*) = (\lambda_{\max} - \lambda_{\min})\sqrt{1-\gamma}. \tag{7.48}$$

As a result, we obtain

$$\max_{\chi_1,\chi_2\in\mathscr{X}} \left\{ \mathrm{Tr}(\mathcal{O}\chi_1) - \mathrm{Tr}(\mathcal{O}\chi_2) \,\middle|\, F(\chi_1,\chi_2) \geq \gamma \right\} = (\lambda_{\max} - \lambda_{\min})\sqrt{1-\gamma}. \tag{7.49}$$

Combining this with Eq. (7.42) gives Eq. (7.41). □

Combining the lower bound on the minimax norm in Eq. (7.41) with the lower bound on estimation error derived in Thm. 7.13 gives the following result.

Theorem 7.15. *Every estimation procedure that learns the expectation value of the observable $\mathcal{O}$ using N samples of a non-adaptive measurement to an error of ε and a confidence level $1-\delta \in (0.75, 1)$ satisfies*

$$\varepsilon \geq \mathscr{c}(\delta)\frac{(\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O}))}{2}\sqrt{1-\left(\frac{\delta}{2}\right)^{2/N}}, \tag{7.50}$$

where $\mathscr{c}(\delta)$ is defined in Eq. (7.32). Equivalently, every estimation procedure needs at least

$$N \geq \frac{2\log(2/\delta)}{\left|\log\left(1 - \frac{4\varepsilon^2}{\mathscr{c}(\delta)^2(\lambda_{\max}(\mathcal{O})-\lambda_{\min}(\mathcal{O}))^2}\right)\right|} \tag{7.51}$$

samples to learn the expectation value of $\mathcal{O}$ to within an error of $\varepsilon \in (0, 0.5)$ and confidence level of $1-\delta \in (0.75, 1)$.

Note that since $\log(1+x) \leq x$ for $x > -1$ and $\log(1+x) \geq 2x$ for $x \in [-1/2, 0]$, we have

$$\frac{\mathscr{c}(\delta)^2}{4}\frac{(\lambda_{\max}-\lambda_{\min})^2}{\varepsilon^2}\log\left(\frac{2}{\delta}\right) \leq \frac{2\log(2/\delta)}{\left|\log\left(1 - \frac{4\varepsilon^2}{\mathscr{c}(\delta)^2(\lambda_{\max}-\lambda_{\min})^2}\right)\right|} \leq \frac{\mathscr{c}(\delta)^2}{2}\frac{(\lambda_{\max}-\lambda_{\min})^2}{\varepsilon^2}\log\left(\frac{2}{\delta}\right) \tag{7.52}$$

for $0 < \varepsilon \leq \mathscr{c}(\delta)(\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O}))/4$.

Proof. From Thm. 7.13, we know that

$$\mathscr{R}_*(\mathcal{O}, \mathfrak{M}, \delta) \geq \mathscr{c}(\delta) \left\|\mathcal{O}\right\|_{\mathfrak{M},\delta} \tag{7.53}$$

for any given measurement protocol $\mathfrak{M}$. Minimizing over all measurement protocols using N samples, we obtain

$$\mathscr{R}_*(\mathcal{O}, N, \delta) \geq \mathscr{c}(\delta) \inf_{\mathfrak{M}} \left\|\mathcal{O}\right\|_{\mathfrak{M},\delta}, \tag{7.54}$$

where $\mathscr{R}_*(\mathcal{O}, N, \delta)$ is the minimax optimal risk defined in Eq. (5.3). Then, from Lem. 7.14, we obtain

$$\varepsilon \geq \mathscr{R}_*(\mathcal{O}, N, \delta) \geq \mathscr{c}(\delta) \frac{(\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O}))}{2} \sqrt{1 - \left(\frac{\delta}{2}\right)^{2/N}}, \tag{7.55}$$

where the first inequality follows from the definition of $\mathscr{R}_*(\mathcal{O}, N, \delta)$. Rearranging Eq. (7.50) gives Eq. (7.51). □

Aaronson [1] has proven a lower bound on the sample complexity of learning the expectation value of an observable similar to Eq. (7.51) that holds in the worst-case over all observables with bounded operator norm. The worst-case lower bound can be too large for a given observable, when the operator norm of the observable is large but the difference between the maximum and minimum eigenvalues is small.

It remains to prove that the lower bound derived in Eq. (7.50) is tight. As one would intuitively expect, this can be achieved by measuring in the eigenbasis of $\mathcal{O}$, as we show below.

Proposition 7.16. $\mathbb{TOOL}$ *can learn the expectation value of an observable* $\mathcal{O}$ *to an error of* $\varepsilon > 0$ *and a confidence level of* $1 - \delta \in (0.75, 1)$ *using at most*

$$\begin{aligned} &\frac{2\log(2/\delta)}{\left|\log\left(1 - \frac{4\varepsilon^2}{(\lambda_{\max} - \lambda_{\min})^2}\right)\right|} \\ &\approx \frac{\mathscr{c}(\delta)^2}{2} \frac{(\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O}))^2}{\varepsilon^2} \log\left(\frac{2}{\delta}\right) \quad \text{for } \varepsilon \ll \mathscr{c}(\delta)(\lambda_{\max}(\mathcal{O}) - \lambda_{\min}(\mathcal{O})) \end{aligned} \tag{7.56}$$

outcomes obtained by measuring in the eigenbasis of $\mathcal{O}$.

Proof. Let $\{|\lambda_1\rangle, \ldots, |\lambda_d\rangle\}$ denote an orthonormal eigenbasis of $\mathcal{O}$. Let the measurement protocol $\mathfrak{M}$ consist of measuring the POVM $\{|\lambda_1\rangle\langle\lambda_1|, \ldots, |\lambda_d\rangle\langle\lambda_d|\}$ N times. For all $\eta > 0$, we know from Thm. 7.13 that $\mathbb{TOOL}$ can learn $\langle\mathcal{O}\rangle$ to an error of $(1+\eta)\|\mathcal{O}\|_{\mathfrak{M},\delta}$ and a confidence level of $1-\delta \in (0.75, 1)$. Thus, it suffices to compute $\|\mathcal{O}\|_{\mathfrak{M},\delta}$. We denote $\lambda_{\max}(\mathcal{O}) = \lambda_{\max}$ and $\lambda_{\min}(\mathcal{O}) = \lambda_{\min}$ in the proof.

From Prop. 7.5.1, we have $\|\mathcal{O}\|_{\mathfrak{M},\delta} = \|\mathcal{O}'\|_{\mathfrak{M},\delta}$ for $\mathcal{O}' = \mathcal{O} - \lambda_{\min}\mathbb{I}$. $\mathcal{O}'$ has eigenvalues $\lambda'_k = \lambda_k - \lambda_{\min}$ for $k \in [d]$ and the same eigenvectors as $\mathcal{O}$. Next, note that $\mathrm{Tr}(\mathcal{O}'\chi) = \sum_{k=1}^d \lambda'_k p_\chi(k)$ for all states χ, where $p_\chi(k) = \langle\lambda_k|\chi|\lambda_k\rangle$. Moreover, we have $\mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2) = \sum_{k=1}^d \sqrt{p_{\chi_1}(k)p_{\chi_2}(k)} = \mathrm{BC}(p_{\chi_1}, p_{\chi_2})$ for all states χ_1, χ_2. Thus, the optimization defining the minimax norm in Eq. (7.2) becomes

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} = \frac{1}{2} \max_{p,q\in\Delta_d} \quad & \sum_{k=1}^d \lambda'_k (p_k - q_k) \\ \text{s.t.} \quad & \sum_{k=1}^d \sqrt{p_k q_k} \geq \left(\frac{\delta}{2}\right)^{1/N}. \end{aligned} \tag{7.57}$$

We wish to derive an upper bound on $\|\mathcal{O}\|_{\mathfrak{M},\delta}$. For this purpose, note that if $p = q$, then $\sum_{k=1}^d \lambda'_k(p_k - q_k) = 0$, so that the minimax norm is zero. Thus, we focus on distributions $p \neq q$. In this case, the sets $I_+ = \{k \in [d] \mid p_k - q_k \geq 0\}$ and $I_- = \{k \in [d] \mid p_k - q_k < 0\}$ are non-empty. Observe that $\sum_{k=1}^d \lambda'_k(p_k - q_k) = \sum_{k\in I_+} \lambda'_k(p_k - q_k) + \sum_{k\in I_-} \lambda'_k(p_k - q_k) \leq \lambda'_{\max} \sum_{k\in I_+}(p_k - q_k)$, where the inequality follows from the fact that $\lambda'_k \geq 0$ for all k. Since $\sum_{k=1}^d (p_k - q_k) = 0$, we have $\sum_{k\in I_-}(p_k - q_k) = -\sum_{k\in I_+}(p_k - q_k)$. Noting that $\|p - q\|_1 = \sum_{k\in I_+}(p_k - q_k) - \sum_{k\in I_-}(p_k - q_k)$, we obtain $\sum_{k\in I_+}(p_k - q_k) = \|p - q\|_1 / 2 = \|p - q\|_{\mathrm{TV}}$. Thus, we have the upper bound $\sum_{k=1}^d \lambda'_k(p_k - q_k) \leq \lambda'_{\max}\|p - q\|_{\mathrm{TV}} = (\lambda_{\max} - \lambda_{\min})\|p - q\|_{\mathrm{TV}}$. From Fuchs-van de Graaf inequality (Eq. (3.36)) and the constraint $\mathrm{BC}(p, q) \geq (\delta/2)^{1/N}$, we obtain $\|p - q\|_{\mathrm{TV}} \leq \sqrt{1 - (\delta/2)^{2/N}}$. It follows that $\|\mathcal{O}\|_{\mathfrak{M},\delta} \leq ((\lambda_{\max} - \lambda_{\min})/2)\sqrt{1 - (\delta/2)^{2/N}}$.

Thus, to learn $\langle\mathcal{O}\rangle$ to an error of ε, we set $(1+\eta)\|\mathcal{O}\|_{\mathfrak{M},\delta} = \varepsilon$, so that $\varepsilon \leq (1+\eta)((\lambda_{\max} - \lambda_{\min})/2)\sqrt{1 - (\delta/2)^{2/N}}$. Solving this for N and noting this holds for arbitrarily small $\eta > 0$ gives the desired result. □

Prop. 7.16 shows that the lower bound on sample complexity derived in Thm. 7.15 is tight to within a factor of $1/c(\delta)^2$. While we used $\mathbb{TOOL}$ to obtain an upper bound in Prop. 7.16, one can also use Hoeffding's inequality [52] to get obtain an upper bound on the sample complexity.

Prop. 7.16 says that the sample complexity scales only with the difference in maximum and minimum eigenvalues of the observable, and not the dimension of the system. Thus, for many observables of interest, it is, in principle, possible to efficiently estimate their expectation values. The problem, however, is that this requires measuring in the eigenbasis of $\mathcal{O}$, which can be very challenging in practice depending on $\mathcal{O}$. Usually, one works with a class of measurement protocols that are relatively easy to implement in the underlying quantum computing architecture. This motivates the importance of Thm. 7.13, which gives tight bounds on learning the expectation value of an observable for a measurement protocol chosen according to experimental constraints.

7.4 Lower and upper bounds on the error for shadow tomography

In this section, we obtain lower bounds on the error of simultaneously estimating the expectation value of many observables. This problem is called shadow tomography, and was first studied by Aaronson [1]. In addition to providing lower bounds for shadow tomography, we give corresponding upper bounds on the error achieved by $\mathbb{TOOL}$.

To learn the expectation values of many observables simultaneously using $\mathbb{TOOL}$, we just learn these expectation values separately and use the union bound to combine the estimates. We describe the general procedure below, and give theoretical guarantees in Prop. 7.17.

Box 7: Shadow tomography with $\mathbb{TOOL}$

Input: Observables $\mathcal{O}_1, \ldots, \mathcal{O}_R$, measurement protocol $\mathfrak{M}$,

confidence level $1-\delta \in (0,1)$, parameter $0 < \epsilon_\circ \ll 1$

Estimator construction:

For $i \in [R]$, compute the estimator $\widehat{\mathcal{O}}_*^{(i)}$ and the error $\varepsilon_*^{(i)}$ for learning $\langle \mathcal{O}_i \rangle$ using Box 3, with measurement protocol $\mathfrak{M}$, confidence level $1-\delta/R$, and parameter $\epsilon_\circ$.

Output: estimators $\widehat{\mathcal{O}}_*^{(1)}, \ldots, \widehat{\mathcal{O}}_*^{(R)}$, estimation error $\varepsilon_* = \max_{i\in[R]} \varepsilon_*^{(i)}$.

The main thing to note in the above procedure is that we need to implement $\mathbb{TOOL}$ for each observable for a confidence level of $1-\delta/R$ instead of $1-\delta$. We now prove a lower bound on learning the expectation values of many observables simultaneously, and study the performance of $\mathbb{TOOL}$.

Proposition 7.17. *The error ε of every estimation procedure that can learn the expectation values of the observables $\mathcal{O}_1, \ldots, \mathcal{O}_R$ simultaneously using outcomes of the measurement protocol $\mathfrak{M}$ with confidence level $1-\delta \in (0.75, 1)$is bounded below as*

$$\varepsilon \geq \mathscr{c}(\delta) \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta}\,. \tag{7.58}$$

On the other hand, for all $\eta > 0$, there is some $\epsilon'_\circ \in (0, \delta/(2NR)]$, such that for all $\epsilon_\circ \in (0, \epsilon'_\circ]$, using $\mathbb{TOOL}$ with parameter $\epsilon_\circ$ and confidence level $1-(\delta - N\epsilon_\circ)$ according to Box 7 can simultaneously learn the expectation values of $\mathcal{O}_1, \ldots, \mathcal{O}_R$ with error

$$(1+\eta) \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\,, \tag{7.59}$$

to a confidence level of $1-\delta$.

Proof. If for $i \in [L]$, $\widehat{\mathcal{O}}_i$ denotes the estimator used by this procedure for learning $\langle \mathcal{O}_i \rangle$, then

$$\mathbb{P}_{\mathfrak{M},\rho}\left(|\widehat{\mathcal{O}}_i - \langle \mathcal{O}_i\rangle\,| > \varepsilon\right) \leq \mathbb{P}_{\mathfrak{M},\rho}\left(\max_{j\in[R]} |\widehat{\mathcal{O}}_j - \langle \mathcal{O}_j\rangle\,| > \varepsilon\right) < \delta, \tag{7.60}$$

for all states ρ. Then, by Thm. 7.13, we must have $\varepsilon \geq \mathscr{c}(\delta)\,\|\mathcal{O}_i\|_{\mathfrak{M},\delta}$ for all $i \in [L]$, giving Eq. (7.58).

Next, we obtain an upper bound achieved by $\mathbb{TOOL}$. Fix $\eta > 0$. For each $i \in [R]$, by Lem. 7.11.2, we have $\lim_{\epsilon_\circ\to 0} \|\mathcal{O}_i\|_{\mathfrak{M}(\epsilon_\circ),\delta/R-N\epsilon_\circ} = \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}$. Thus, given $\eta > 0$, there is some $0 < \epsilon_\circ < \delta/(2NR)$ such that for all $i \in [R]$, we have $\|\mathcal{O}_i\|_{\mathfrak{M}(\epsilon_\circ),\delta/R-N\epsilon_\circ} \leq (1+\eta)\,\|\mathcal{O}_i\|_{\delta/R}$. We then

follow the strategy of Thm. 7.13. First, from Eq. (7.17), for all $i \in [R]$, we have

$$\begin{aligned}\mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}\left(|\widehat{\mathcal{O}}_i^* - \mathrm{Tr}(\mathcal{O}_i\rho)| \leq (1+\eta)\,\|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\right) &\geq \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}\left(|\widehat{\mathcal{O}}_i^* - \mathrm{Tr}(\mathcal{O}_i\rho)| \leq \|\mathcal{O}_i\|_{\mathfrak{M}(\epsilon_\circ),\delta/R-N\epsilon_\circ}\right)\\ &\geq 1 - \left(\frac{\delta}{R} - N\epsilon_\circ\right)\end{aligned} \tag{7.61}$$

for all states ρ. From Eq. (7.29), we know that $\left\|\mathscr{P}_{\mathfrak{M}(\epsilon_\circ),\rho} - \mathbb{P}_{\mathfrak{M},\rho}\right\|_{\mathrm{TV}} \leq N\epsilon_\circ/(1+\epsilon_\circ)$ for all ρ. It follows from the definition of total variation distance that for all ρ and all $i \in [R]$, we have

$$\begin{aligned}\mathbb{P}_{\mathfrak{M},\rho}(\{|\widehat{\mathcal{O}}_i^* - \mathrm{Tr}(\mathcal{O}_i\rho)| \leq (1+\eta)\,\|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\}) &\geq \mathbb{P}_{\mathfrak{M}(\epsilon_\circ),\rho}(\{|\widehat{\mathcal{O}}_i^* - \mathrm{Tr}(\mathcal{O}_i\rho)| \leq (1+\eta)\,\|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\}) - N\frac{\epsilon_\circ}{1+\epsilon_\circ}\\ &\geq 1 - \delta + N\epsilon_\circ - N\frac{\epsilon_\circ}{1+\epsilon_\circ}\\ &\geq 1 - \frac{\delta}{R} + \frac{N\epsilon_\circ^2}{1+\epsilon_\circ}\\ &> 1 - \frac{\delta}{R}\end{aligned} \tag{7.62}$$

Therefore, for all ρ and all $i \in [R]$, we have

$$\mathbb{P}_{\mathfrak{M},\rho}\left(|\widehat{\mathcal{O}}_i^* - \langle\mathcal{O}_i\rangle| > (1+\eta)\max_{j\in[R]}\|\mathcal{O}_j\|_{\mathfrak{M},\delta/R}\right) \leq \mathbb{P}_{\mathfrak{M},\rho}\left(|\widehat{\mathcal{O}}_i^* - \langle\mathcal{O}_i\rangle| > (1+\eta)\,\|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\right) < \frac{\delta}{R}. \tag{7.63}$$

Then, by the union bound, we can infer that

$$\mathbb{P}_{\mathfrak{M},\rho}\left(\max_{j\in[R]}|\widehat{\mathcal{O}}_j^* - \langle\mathcal{O}_j\rangle| > (1+\eta)\max_{j\in[R]}\|\mathcal{O}_j\|_{\mathfrak{M},\delta/R}\right) < \delta, \tag{7.64}$$

giving the desired result. □

The error in Eq. (7.59) obtained by implementing Box 7 does not always match the lower bound in Eq. (7.58) because we are using the union bound to derive Eq. (7.59). In particular, since $\delta/R < \delta$, from Prop. 7.5, we know that $\|\mathcal{O}_i\|_{\mathfrak{M},\delta/R} \geq \|\mathcal{O}_i\|_{\mathfrak{M},\delta}$ for all $i \in [R]$. Therefore, even though $\eta > 0$ can be made arbitrarily small, Eq. (7.59) can be larger than the lower bound in Eq. (7.58). As to how large the error in Eq. (7.59) is compared to Eq. (7.58) will depend on the observables

and the measurement protocol. While the error achieved by Box 7 is not always within a constant factor of the lower bound in Eq. (7.58), we can prove a slightly weaker optimality result for Box 7 that is sufficient in practice.

Proposition 7.18. *For every estimation procedure that learns the expectation values of $\mathcal{O}_1, \ldots, \mathcal{O}_R$ simultaneously using the outcomes of $\mathfrak{M}$ to a confidence level of $1-\delta \in (0.75, 1)$ for all states by learning each expectation value separately to an error $\varepsilon^{(i)}$ and a confidence level of $1-\delta/R$ and using the union bound to obtain an error of $\varepsilon = \max_{i\in[R]} \varepsilon^{(i)}$ must satisfy*

$$\varepsilon \geq \mathscr{c}(\delta) \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\,. \tag{7.65}$$

Proof. Suppose that for each $i \in [R]$, the estimation procedure learns $\langle \mathcal{O}_i \rangle$ to error $\varepsilon^{(i)}$ to a confidence level of $1-\delta/R$. Then, by Thm. 7.13, we must have

$$\varepsilon^{(i)} \geq \mathscr{c}(\delta)\, \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R} \tag{7.66}$$

for all $i \in [R]$. Taking maximum over $i \in [R]$ gives Eq. (7.65). □

The strategy described in Prop. 7.18 is commonly used by many estimation procedures in practice, including classical shadows [54]. Then, the result of Prop. 7.18 and Prop. 7.17 together imply that 𝕋𝕆𝕆𝕃 performs at least as good as such estimation procedures, up to a factor of $1/\mathscr{c}(\delta)$. A detailed comparison of 𝕋𝕆𝕆𝕃 with classical shadows is presented in Ch. 8.

Prop. 7.18 and Prop. 7.17 derived lower bounds on shadow tomography for a fixed measurement protocol. Now, we derive a lower bound, allowing all measurement protocols.

Proposition 7.19. *Let $\mathbb{M}$ denote a set of measurement protocols that use a fixed number of samples. Then, the error ε of every procedure that simultaneously learns the expectation values of $\mathcal{O}_1, \ldots, \mathcal{O}_R$ using a measurement protocol from the set $\mathbb{M}$, with probability greater than $1-\delta \in (0.75, 1)$ for all*

states is bounded below as

$$\varepsilon \geq \mathscr{c}(\delta) \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta}\,. \tag{7.67}$$

If we focus on estimation procedures that learns the expectation values of $\mathcal{O}_1,\ldots,\mathcal{O}_R$ *simultaneously using the outcomes of* $\mathfrak{M}$ *to a confidence level of* $1-\delta\in(0.75,1)$ *for all states by learning each expectation value separately to an error* $\varepsilon^{(i)}$ *and a confidence level of* $1-\delta/R$ *and using the union bound to obtain an error of* $\varepsilon=\max_{i\in[R]}\varepsilon^{(i)}$*, the lower bound can be improved to*

$$\varepsilon \geq \mathscr{c}(\delta) \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\,. \tag{7.68}$$

On the other hand, for all $\eta>0$*, there is a measurement protocol in* $\mathbb{M}$ *such that* $\mathbb{TOOL}$ *can use the outcomes of this protocol to simultaneously learn the expectation values of* $\mathcal{O}_1,\ldots,\mathcal{O}_R$ *to within an error of*

$$(1+\eta)\inf_{\mathfrak{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R} \tag{7.69}$$

with probability greater than $1-\delta\in(0,1)$*.*

Proof. Given any procedure that implements the measurement protocol $\mathfrak{M}'\in\mathbb{M}$, we know from Prop. 7.17 that the error of simultaneously learning $\langle\mathcal{O}_1\rangle,\ldots,\langle\mathcal{O}_R\rangle$ is bounded below as

$$\varepsilon \geq \mathscr{c}(\delta) \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M}',\delta} \geq \mathscr{c}(\delta) \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta}\,, \tag{7.70}$$

which gives Eq. (7.67). If instead we focus on estimation procedures that use the union bound to learn the expectation values of $\mathcal{O}_1,\ldots,\mathcal{O}_R$ simultaneously, then by Prop. 7.18, the error is bounded below as

$$\varepsilon \geq \mathscr{c}(\delta) \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M}',\delta/R} \geq \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\,. \tag{7.71}$$

Now, assume that at least one of the observables $\mathcal{O}_1,\ldots,\mathcal{O}_R$ is not a multiple of identity, for otherwise, the expectation value of each $\mathcal{O}_i$ is zero and there is nothing to learn. Assuming that the

measurement protocols in $\mathbb{M}$ use N samples, by Thm. 7.15, we have

$$\max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R} \geq \frac{c(\delta)}{2} \max_{i\in[R]} (\lambda_{\max}(\mathcal{O}_i) - \lambda_{\min}(\mathcal{O}_i)) \sqrt{1 - \left(\frac{\delta}{2R}\right)^{2/N}} \tag{7.72}$$

for all $\mathfrak{M} \in \mathbb{M}$, $\delta \in (0,1)$, and $R \geq 1$, and consequently,

$$\inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R} \geq \frac{c(\delta)}{2} \max_{i\in[R]} (\lambda_{\max}(\mathcal{O}_i) - \lambda_{\min}(\mathcal{O}_i)) \sqrt{1 - \left(\frac{\delta}{2R}\right)^{2/N}} > 0. \tag{7.73}$$

Fix $\eta > 0$, and define $\eta_0 = \sqrt{1+\eta} - 1$. Since $\eta_0 > 0$, we have

$$(1+\eta_0) \inf_{\mathfrak{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R} > \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\,, \tag{7.74}$$

and therefore, by the definition of infimum, there is some measurement protocol $\mathfrak{M}_* \in \mathbb{M}$ such that

$$\max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M}_*,\delta/R} < (1+\eta_0) \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\,. \tag{7.75}$$

From Prop. 7.17, we know that the using the measurement protocol $\mathfrak{M}_*$, $\mathbb{TOOL}$ can learn $\langle\mathcal{O}_1\rangle, \ldots, \langle\mathcal{O}_R\rangle$ simultaneously to a confidence level of $1-\delta$ with an error of

$$(1+\eta_0) \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M}_*,\delta/R} < (1+\eta_0)^2 \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R} = (1+\eta) \inf_{\mathfrak{M}\in\mathbb{M}} \max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta/R}\,, \tag{7.76}$$

where $(1+\eta_0)^2 = (1+\eta)$ by definition of η_0. □

As in Prop. 7.18, if we only consider estimation procedures in Prop. 7.19 that simultaneously learn the expectation values of $\mathcal{O}_1, \ldots, \mathcal{O}_R$ using the union bound, then $\mathbb{TOOL}$ is minimax optimal up to a small constant factor.

Before ending this section, we give an alternate expression for $\max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta}$. We begin by introducing a seminorm determined by a given list of observables.

Definition 7.20. Given a list of observables $\boldsymbol{\mathcal{O}} = (\mathcal{O}_1, \ldots, \mathcal{O}_R)$, we define the seminorm on $\mathbb{S}_d$

induced by $\boldsymbol{\mathscr{O}}$ as

$$\|Q\|_{\boldsymbol{\mathscr{O}}} = \frac{1}{2} \max_{i\in[R]} |\mathrm{Tr}(\mathscr{O}_i Q)| \tag{7.77}$$

for $Q \in \mathbb{S}_d$. □

It can be verified that $\|\cdot\|_{\boldsymbol{\mathscr{O}}}$ is a seminorm on $\mathbb{S}_d$. Moreover, if $-\mathbb{I} \leq \mathscr{O}_1, \ldots, \mathscr{O}_R \leq \mathbb{I}$, and

$$\mathfrak{M}(\boldsymbol{\mathscr{O}}) = \{(\{(\mathbb{I} + \mathscr{O}_i)/2, (\mathbb{I} - \mathscr{O}_i)/2\}, N_i)\}_{i=1}^R \tag{7.78}$$

denotes the measurement protocol corresponding to measuring the two-outcome POVMs defined by $\mathscr{O}_1, \ldots, \mathscr{O}_R$, then

$$\|\rho - \sigma\|_{\boldsymbol{\mathscr{O}}} = \|\rho - \sigma\|_{\mathfrak{M}(\boldsymbol{\mathscr{O}}),\max}. \tag{7.79}$$

We now give an expression for the minimax norm in terms of $\|\cdot\|_{\boldsymbol{\mathscr{O}}}$.

Proposition 7.21. *1. The minimax norm of an observable $\mathscr{O}$ given the measurement protocol $\mathfrak{M}$ and confidence level $1-\delta$ can be expressed as*

$$\begin{aligned} \|\mathscr{O}\|_{\mathfrak{M},\delta} = \frac{1}{2} \max_{\chi_1,\chi_2\in\mathscr{X}} \quad & |\mathrm{Tr}(\mathscr{O}\chi_1) - \mathrm{Tr}(\mathscr{O}\chi_2)| \\ \text{s.t.} \quad & \mathrm{BD}_{\mathfrak{M}}(\chi_1,\chi_2) \leq \frac{1}{N}\log\left(\frac{2}{\delta}\right). \end{aligned} \tag{7.80}$$

2. Given a list of observables $\boldsymbol{\mathscr{O}} = (\mathscr{O}_1, \ldots, \mathscr{O}_R)$, a measurement protocol $\mathfrak{M}$, and a confidence level $1-\delta$, we have

$$\begin{aligned} \max_{i\in[R]} \|\mathscr{O}_i\|_{\mathfrak{M},\delta} = \max_{\chi_1,\chi_2\in\mathscr{X}} \quad & \|\chi_1 - \chi_2\|_{\boldsymbol{\mathscr{O}}} \\ \text{s.t.} \quad & \mathrm{BD}_{\mathfrak{M}}(\chi_1,\chi_2) \leq \frac{1}{N}\log\left(\frac{2}{\delta}\right). \end{aligned} \tag{7.81}$$

Proof. 1. Since $(\chi_1,\chi_2) \in \mathscr{C}(\mathfrak{M},\delta)$ iff $(\chi_2,\chi_2) \in \mathscr{C}(\mathfrak{M},\delta)$ (Prop. 7.4.1.iii), we have $\mathrm{Tr}(\mathscr{O}(\chi_1-\chi_2)) \leq \|\mathscr{O}\|_{\mathfrak{M},\delta}$ and $-\mathrm{Tr}(\mathscr{O}(\chi_1-\chi_2)) \leq \|\mathscr{O}\|_{\mathfrak{M},\delta}$ for all $(\chi_1,\chi_2) \in \mathscr{C}(\mathfrak{M},\delta)$. This gives $\mathrm{Tr}(\mathscr{O}(\chi_1-\chi_2)) \leq |\mathrm{Tr}(\mathscr{O}(\chi_1-\chi_2)| \leq \|\mathscr{O}\|_{\mathfrak{M},\delta}$ for all $(\chi_1,\chi_2) \in \mathscr{C}(\mathfrak{M},\delta)$. Optimizing over $(\chi_1,\chi_2) \in \mathscr{C}(\mathfrak{M},\delta)$ gives Eq. (7.80).

2. Eq. (7.81) follows from Eq. (7.80), Def. 7.20, and the fact that maximums commute. □

Eq. (7.81) shows that $\max_{i\in[R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta}$ is related to the constant of domination studied in [71], though with respect to different distance measures than those studied in [71].

Chapter 8

Miscellaneous applications of the lower bounds

In Ch. 7, we showed that the minimax norm provides a tight lower bound for learning the expectation values of observables for any given measurement protocol. Therefore, we can use the minimax norm as a figure of merit to study and compare the performance of different measurement protocols and estimation procedures for learning the expectation values of observables. Motivated by this, we study some upper and lower bounds on the minimax norm for different measurement protocols, in order to prove feasibility and infeasibility results.

In Sec. 8.1, we study the relation of the minimax norm to the shadow norm for measurement protocols described by a single POVM. Since randomized measurements can be described by a single effective POVM, our analysis applies to randomized measurements as well. We first show that the minimax norm is always smaller than the shadow norm up to appropriate constant factors, which implies that $\mathbb{TOOL}$ always performs as well as classical shadows. Subsequently, we show that there are many observables for which the minimax norm is exponentially smaller than the shadow norm, implying an exponential advantage of $\mathbb{TOOL}$ over classical shadows. In Sec. 8.2, we present two no-go theorems, one for estimating the fidelity with stabilizer states, and another for learning the expectation values of arbitrary observables.

8.1 Randomized measurements

In this section, we study the performance of learning the expectation values of observables using measurement protocols described by a single POVM. As discussed in the preliminaries (Sec. 2.2),

randomized measurement protocols can be described by an effective POVM, and therefore, our analysis of single POVMs apply to randomized measurement protocols as well.

To begin with, we give a brief description of the classical shadows estimation procedure, as given in [54]. We discuss the procedure for a system of n-qubits (i.e., $d = 2^n$) following [54], but we note that the procedure can be generalized to other scenarios. We fix a set $\mathscr{U}$ of unitary operators, called a unitary ensemble, and randomly sample a unitary operator from $\mathscr{U}$ according to some probability distribution. If $U \in \mathscr{U}$ was sampled, then we rotate the (unknown) state ρ prepared in the experiment by U, i.e., $\rho \mapsto U\rho U^\dagger$. After this, we perform a computational basis measurement. If the outcome $b \in \{0,1\}^n$ was observed, then we store a classical description of the state $U^\dagger |b\rangle \langle b| U$. This measurement procedure is repeated N times, and the resulting classical description, $\{U_i^\dagger |b_i\rangle \langle b_i| U_i\}_{i=1}^N$, is called a classical shadow. Then, given observables $\mathscr{O}_1, \ldots, \mathscr{O}_R$, for each $r \in [R]$, one implements the median-of-means estimation procedure on $\{\mathrm{Tr}(\mathscr{O}_r U_i^\dagger |b_i\rangle \langle b_i| U_i)\}_{i=1}^N$ to learn $\mathrm{Tr}(\mathscr{O}_r \rho)$ to a confidence level of $1 - \delta/R$. Finally, one uses the union bound to estimate $\mathrm{Tr}(\mathscr{O}_1\rho), \ldots, \mathrm{Tr}(\mathscr{O}_R\rho)$ in l_∞-norm to a confidence level of $1-\delta$. Note that it does not matter whether or not the observables $\mathscr{O}_1, \ldots, \mathscr{O}_R$ are known before the measurements are performed.

In this section, we focus on the case where $\mathscr{U}$ is a finite ensemble, since that covers the ensembles studied by [54]. We call the measurement where $U \in \mathscr{U}$ is sampled according to the probability distribution $p_U \in \Delta_{|\mathscr{U}|}$, the state is rotated by U as $\rho \mapsto U\rho U^\dagger$, and a computational basis measurement is performed on the rotated state as **$\mathscr{U}$-random unitary measurement**. The effective POVM describing a $\mathscr{U}$-random unitary measurement is

$$\{p_U U^\dagger |b\rangle \langle b| U\}_{U\in\mathscr{U}, b\in\{0,1\}^n}. \tag{8.1}$$

Following [54], when we talk about a $\mathscr{U}$-random unitary measurement, we will assume that the effective POVM in Eq. (8.1) is informationally complete. In this case, the map

$$\mathscr{E}(\sigma) = \sum_{U\in\mathscr{U}} \sum_{b\in\{0,1\}^n} p_U \left\langle b|U\sigma U^\dagger|b\right\rangle = \mathbb{E}_{U\sim\mathscr{U}} \sum_{b\in\{0,1\}^n} \left\langle b|U\sigma U^\dagger|b\right\rangle \tag{8.2}$$

for $\sigma \in \mathscr{X}$ is invertible [54]. Using this, we can define the shadow norm of an observable as follows.

Definition 8.1 (Shadow norm). Given a unitary ensemble $\mathscr{U}$, the shadow norm of an observable $\mathcal{O}$ is defined as

$$\|\mathcal{O}\|_{\text{shadow}} = \max_{\sigma \in \mathscr{X}} \sqrt{\mathbb{E}_{U \sim \mathscr{U}} \sum_{b \in \{0,1\}^n} \langle b|U\sigma U^\dagger|b\rangle \langle b|U\mathscr{E}^{-1}(\mathcal{O})U^\dagger|b\rangle^2}. \tag{8.3}$$

□

The shadow norm determines the performance of classical shadows. [54, Thm. 1] shows that the sample complexity of classical shadows for learning the expectation values of $\mathcal{O}_1, \ldots, \mathcal{O}_R$ to an error of $\varepsilon > 0$ and a confidence level of $1 - \delta$ is equal to

$$N = O\left(\frac{\log(R)}{\varepsilon^2} \max_{1 \le i \le R} \left\| \mathcal{O}_i - \frac{\text{Tr}(\mathcal{O}_i)}{d} \mathbb{I} \right\|_{\text{shadow}}^2\right). \tag{8.4}$$

[54] prove the following lower bounds to show that the performance of classical shadows is nearly optimal in the worst case over observables.

Theorem 8.2 (Thm. 5, [54], rephrased). *Fix a measurement protocol $\mathfrak{M}$ using N samples, an estimation error ε, and an integer $R \le \exp(d/32)$, where d is the system dimension. If for all $\mathcal{O}_1, \ldots, \mathcal{O}_R$ with $\max_{i \in [R]} \|\mathcal{O}_i\|_{\text{HS}} \le B$, there are estimators $\widehat{\mathcal{O}}_1, \ldots, \widehat{\mathcal{O}}_R$, such that for all ρ, we have $\max_{r \in [R]} |\widehat{\mathcal{O}}_i - \text{Tr}(\mathcal{O}_i \rho)| \ge \varepsilon$ with high probability, then necessarily*

$$N \ge \Omega\left(\frac{B^2 \log(R)}{\varepsilon^2}\right). \tag{8.5}$$

Thm. 8.2 applies to all measurement protocols. [54] also prove a lower bound that applies specifically to local measurements on a system of n-qubits. A local measurement or a local POVM is a POVM $\boldsymbol{L} = \{w_i d\, |v_i\rangle \langle v_i|\}_{i=1}^M$, where $w \in \Delta_M$, $|v_i\rangle = |v_i^{(1)}\rangle \otimes \cdots \otimes |v_i^{(n)}\rangle$ for all $i \in [M]$, and $d = 2^n$ [54]. We define a local measurement protocol as $\mathfrak{M}_{\text{l}} = \{(\boldsymbol{L}^{(i)}, N_i)\}_{i=1}^L$, where $\boldsymbol{L}^{(i)}$ is a local POVM for all $i \in [R]$. On the other hand, a k-local observable acting on a system of n qubits is an

observable that acts trivially on $n-k$ qubits. Then, [54] prove the following result.

Theorem 8.3 (Thm. 6, [54], rephrased). *Fix a local measurement protocol $\mathfrak{M}$ using N samples, an estimation error ε, number of qubits n, locality k, and an integer $R \leq 3^k \binom{n}{k}$. If for all k-local observables $\mathcal{O}_1, \ldots, \mathcal{O}_R$ with $\max_{i\in[R]} \|\mathcal{O}_i\|_\infty \leq 1$, there are estimators $\widehat{\mathcal{O}}_1, \ldots, \widehat{\mathcal{O}}_R$, such that for all ρ, we have $\max_{r\in[R]} |\widehat{\mathcal{O}}_i - \mathrm{Tr}(\mathcal{O}_i\rho)| \geq \varepsilon$ with high probability, then necessarily*

$$N \geq \Omega\left(\frac{3^k \log(R)}{\varepsilon^2}\right). \tag{8.6}$$

[54] prove that the classical shadows estimation procedure achieves the lower bound in Thm. 8.2 by choosing the unitary ensemble to be global Clifford operators, while classical shadows achieves the lower bound in Thm. 8.3 by choosing the unitary ensemble to be local Clifford operators. Note, however, that lower bounds in Thm. 8.2 and Thm. 8.3 are obtained for the worst case over all states *and* all observables with either a fixed bound on the Hilbert-Schmidt norm or a fixed locality. Thus, in principle, it is possible to improve upon the sample complexity bounds given in Thm. 8.2 and Thm. 8.3 for specific choices of observables.

In this section, we focus on proving three results. First, we prove that $\mathbb{TOOL}$ does at least as well as classical shadows for every unitary ensemble, and more generally, for every measurement protocol that can be described by a single POVM. Second, we find observables for which $\mathbb{TOOL}$ can beat both the lower bound of Thm. 8.3 as well as classical shadows. Since the lower bound of Thm. 8.3 is worst case over all observables with a fixed locality, classical shadows itself may perform better than the lower bound for specific observables. It is therefore necessary to show that $\mathbb{TOOL}$ can do better than both the lower bound and classical shadows. Note that we focus on local measurements here because they are more practical to implement with the current experimental capabilities. Finally, we prove two data-processing inequalities for the minimax norm for randomized measurements.

To begin with, we note that for the measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}, N)\}$ that involves implementing the POVM $\boldsymbol{E} = \{E_k\}_{k=1}^M$ N times, we can simplify the expressions for the classical

distance measures defined in Sec. 3.1. In particular, if we denote $p_\chi(k) = \mathrm{Tr}(E_k\chi)$ for $\chi \in \mathscr{X}$, then for all $\chi_1, \chi_2 \in \mathscr{X}$, we have

$$\mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2) = \sum_{k=1}^{M} \sqrt{p_{\chi_1}(k) p_{\chi_2}(k)} = \mathrm{BC}(p_{\chi_1}, p_{\chi_2}), \tag{8.7}$$

$\mathrm{FC}_{\mathfrak{M}}(\chi_1, \chi_2) = (\mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2))^2$, and $\mathrm{BD}_{\mathfrak{M}}(\chi_1, \chi_2) = -\log(\mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2))$.

We can also generalize the definition of the shadow norm for measurement protocols described by a single POVM. Note that generalization of classical shadows to informationally complete POVMs and frames has been studied in the recent past. We briefly discuss how to define the shadow norm to the case of single POVM measurements. First, note that the map $\mathscr{E}$ given in Eq. (8.2) generalizes to

$$\mathscr{E}(\sigma) = \sum_{k=1}^{M} \mathrm{Tr}(E_k\sigma) E_k. \tag{8.8}$$

We can obtain Eq. (8.2) from Eq. (8.8) by taking $\boldsymbol{E}$ to be the effective POVM given in Eq. (8.1). When the POVM $\boldsymbol{E}$ is informationally complete, the map $\mathscr{E}$ has a left inverse $\mathscr{E}^{-1}$ [88]. It can be verified that the converse also holds, i.e., if $\mathscr{E}$ has a left inverse, then $\boldsymbol{E}$ must be informationally complete. Thus, we can generalize shadow norm to an informationally complete POVM as follows.

Definition 8.4 (Shadow norm for an IC-POVM). Given an informationally complete POVM $\boldsymbol{E}$, we define the shadow norm of an observable $\mathscr{O}$ with respect to $\boldsymbol{E}$ as

$$\|\mathscr{O}\|_{\mathrm{shadow}} = \max_{\sigma \in \mathscr{X}} \sqrt{\sum_{k=1}^{M} \left(\mathrm{Tr}\left((\mathscr{E}^{-1})^\dagger(\mathscr{O}) E_k\right)\right)^2 \mathrm{Tr}(E_k\sigma)}. \tag{8.9}$$

□

It can be verified that Eq. (8.9) defines a norm on $\mathbb{S}_d$ (also see [57, Sec. VI]). We now derive an upper bound on the minimax norm in terms of the shadow norm.

Proposition 8.5. *1. Let $\boldsymbol{E}$ be a POVM and $\mathcal{O}$ be an observable contained in the span of $\boldsymbol{E}$, so that*

$$\mathcal{O} = \sum_{k=1}^{M} \alpha_k E_k \tag{8.10}$$

for some (not necessarily unique) numbers $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$. Then, for $\mathfrak{M} = \{(\boldsymbol{E}, N)\}$, we have

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} &\leq \sqrt{2 - 2\left(\frac{\delta}{2}\right)^{1/N}} \inf_{c\in\mathbb{R}} \max_{\sigma\in\mathscr{X}} \sqrt{\sum_{k=1}^{M} (\alpha_k - c)^2 \mathrm{Tr}(E_k\sigma)} \\ &\leq \sqrt{2 - 2\left(\frac{\delta}{2}\right)^{1/N}} \max_{\sigma\in\mathscr{X}} \sqrt{\sum_{k=1}^{M} \alpha_k^2 \mathrm{Tr}(E_k\sigma)}. \end{aligned} \tag{8.11}$$

2. If $\boldsymbol{E}$ is informationally complete, then

$$\mathcal{O} = \sum_{k=1}^{M} \mathrm{Tr}\left((\mathscr{E}^{-1})^\dagger(\mathcal{O})E_k\right) E_k, \tag{8.12}$$

and for $\mathfrak{M} = \{(\boldsymbol{E}, N)\}$, we have

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} &\leq \sqrt{2 - 2\left(\frac{\delta}{2}\right)^{1/N}} \inf_{c\in\mathbb{R}} \|\mathcal{O} - c\mathbb{I}\|_{\text{shadow}} \\ &\leq \sqrt{2 - 2\left(\frac{\delta}{2}\right)^{1/N}} \|\mathcal{O}\|_{\text{shadow}}. \end{aligned} \tag{8.13}$$

Proof. 1. From Eq. (8.10), we can write $\mathrm{Tr}(\mathcal{O}\chi) = \sum_{k=1}^{M} \alpha_k p_\chi(k)$ for any state χ, where $p_\chi(k) = \mathrm{Tr}(E_k\chi)$. Thus, using Eq. (8.7) and the expression for the minimax norm given in Eq. (7.2), we can write

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} = \frac{1}{2} \max_{\chi_1,\chi_2\in\mathscr{X}} \quad & \sum_{k=1} \alpha_k (p_{\chi_1}(k) - p_{\chi_2}(k)) \\ \text{s.t.} \quad & \mathrm{BC}(p_{\chi_1}, p_{\chi_2}) \geq \left(\frac{\delta}{2}\right)^{1/N}. \end{aligned} \tag{8.14}$$

Since $\sum_k \alpha_k (p_{\chi_1}(k) - p_{\chi_2}(k)) = \sum_k \alpha_k (\sqrt{p_{\chi_1}(k)} + \sqrt{p_{\chi_2}(k)})(\sqrt{p_{\chi_1}(k)} - \sqrt{p_{\chi_2}(k)})$, by Cauchy-

Schwarz inequality, we obtain

$$\sum_{k=1}^{M} \alpha_k (p_{\chi_1}(k) - p_{\chi_2}(k)) \leq \sqrt{2} \left(\sqrt{\sum_{k=1}^{M} \alpha_k^2 p_{\chi_1}(k)} + \sqrt{\sum_{k=1}^{M} \alpha_k^2 p_{\chi_2}(k)} \right) \mathrm{HD}(p_{\chi_1}, p_{\chi_2}),$$

where $\mathrm{HD}(p_{\chi_1}, p_{\chi_2})$ is the Hellinger distance between the distributions p_{χ_1} and p_{χ_2} defined in Eq. (3.21). Since $\mathrm{HD}(p_{\chi_1}, p_{\chi_2}) = \sqrt{1 - \mathrm{BC}(p_{\chi_1}, p_{\chi_2})}$, the constraint $\mathrm{BC}(p_{\chi_1}, p_{\chi_2}) \geq (\delta/2)^{1/N}$ implies $H(p_{\chi_1}, p_{\chi_2}) \leq \sqrt{1 - (\delta/2)^{1/N}}$. Therefore, we have

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} &\leq \frac{1}{2} \max_{\chi_1, \chi_2 \in \mathscr{X}} \left(\sqrt{2} \left(\sqrt{\sum_{k=1}^{M} \alpha_k^2 p_{\chi_1}(k)} + \sqrt{\sum_{k=1}^{M} \alpha_k^2 p_{\chi_2}(k)} \right) \sqrt{1 - \left(\frac{\delta}{2}\right)^{1/N}} \right) \\ &\leq \frac{1}{2} \sqrt{2 - 2\left(\frac{\delta}{2}\right)^{1/N}} \left(\max_{\chi_1 \in \mathscr{X}} \sqrt{\sum_{k=1}^{M} \alpha_k^2 p_{\chi_1}(k)} + \max_{\chi_2 \in \mathscr{X}} \sqrt{\sum_{k=1}^{M} \alpha_k^2 p_{\chi_2}(k)} \right) \\ &= \sqrt{2 - 2\left(\frac{\delta}{2}\right)^{1/N}} \max_{\sigma \in \mathscr{X}} \sqrt{\sum_{k=1}^{M} \alpha_k^2 \mathrm{Tr}(E_k \sigma)}. \end{aligned} \tag{8.15}$$

From Prop. 7.5.1, we have $\|\mathcal{O}\|_{\mathfrak{M},\delta} = \|\mathcal{O} - c\mathbb{1}\|_{\mathfrak{M},\delta}$ for all $c \in \mathbb{R}$. Then, since $\mathcal{O} - c\mathbb{1} = \sum_{k=1}^{m} (\alpha_k - c) E_k$,we obtain Eq. (8.11).

2. Since the map $\mathscr{E}$ defined in Eq. (8.8) has a left inverse $\mathscr{E}^{-1}$ when $\boldsymbol{E}$ is informationally complete, we have $\sigma = \sum_{k=1}^{M} \mathrm{Tr}(E_k \sigma) \mathscr{E}^{-1}(E_k)$ for all $\sigma \in \mathscr{X}$. Consequently, for any observable $\mathcal{O}$, we have

$$\mathrm{Tr}(\mathcal{O}\sigma) = \sum_{k=1}^{M} \mathrm{Tr}\left(\mathcal{O}\mathscr{E}^{-1}(E_k)\right) \mathrm{Tr}(E_k \sigma) = \sum_{k=1}^{M} \mathrm{Tr}\left((\mathscr{E}^{-1})^{\dagger}(\mathcal{O}) E_k\right) \mathrm{Tr}(E_k \sigma). \tag{8.16}$$

Since this holds for every state σ, Eq. (8.12) holds.

Since Eq. (8.11) holds for all $\alpha_1, \ldots, \alpha_M \in \mathbb{R}$ such that $\mathcal{O} = \sum_{k=1}^{M} \alpha_k E_k$, we can take $\alpha_k = \mathrm{Tr}((\mathscr{E}^{-1})^{\dagger}(\mathcal{O}) E_k)$ to obtain $\sqrt{2 - 2(\delta/2)^{1/N}} \, \|\mathcal{O}\|_{\mathrm{shadow}}$. Now, from Prop. 7.5.1, we know that $\|\mathcal{O}\|_{\mathfrak{M},\delta} = \|\mathcal{O} - c\mathbb{1}\|_{\mathfrak{M},\delta}$ for all $c \in \mathbb{R}$. Thus, we also have the stronger inequality $\|\mathcal{O}\|_{\mathfrak{M},\delta} \leq \sqrt{2 - 2(\delta/2)^{1/N}} \inf_{c \in \mathbb{R}} \|\mathcal{O} - c\mathbb{1}\|_{\mathrm{shadow}}$. □

Since $\log(1/x) \geq 1 - x$ for $x > 0$, we have

$$\sqrt{1 - \left(\frac{\delta}{2}\right)^{1/N}} \leq \sqrt{\frac{\log(2/\delta)}{N}}. \tag{8.17}$$

Due to the logarithmic dependence of the upper bound in Eq. (8.13) on $1/\delta$, we can simultaneously learn the expectation values of R observables in l_∞-norm by learning them separately to a confidence level of $1 - \delta/R$ and using the union bound. Then, since the minimax norm determines the performance of $\mathbb{TOOL}$, while the shadow norm determines the performance of classical shadows, we can infer from Eq. (8.13) that $\mathbb{TOOL}$ performs at least as well as classical shadows. We formally prove this below.

Corollary 8.6. *For all informationally complete POVMs $\boldsymbol{E}$, $\mathbb{TOOL}$ can simultaneously learn the expectation values of $\mathcal{O}_1, \ldots, \mathcal{O}_R$ in l_∞-norm to an error of ε and confidence level of $1 - \delta \in (0, 1)$ using at most*

$$\frac{2}{\varepsilon^2} \max_{1 \leq i \leq R} \left\| \mathcal{O}_i - \frac{\operatorname{Tr}(\mathcal{O}_i)}{d} \mathbb{I} \right\|_{\text{shadow}}^2 \log\left(\frac{2R}{\delta}\right) \tag{8.18}$$

outcomes of $\boldsymbol{E}$. In particular, $\mathbb{TOOL}$ performs at least as well as classical shadows for all $\mathcal{U}$-random unitary measurements.

Proof. For all $i \in [R]$, we have from Eq. (8.13) and Eq. (8.17) that

$$\|\mathcal{O}_i\|_{\mathfrak{M},\delta} \leq \sqrt{2 - 2\left(\frac{\delta}{2}\right)^{1/N}} \inf_{c \in \mathbb{R}} \|\mathcal{O}_i - c\mathbb{I}\|_{\text{shadow}} \leq \sqrt{\frac{2\log(2/\delta)}{N}} \left\| \mathcal{O}_i - \frac{\operatorname{Tr}(\mathcal{O}_i)}{d} \mathbb{I} \right\|_{\text{shadow}}. \tag{8.19}$$

From Prop. 7.17, we know that for all $\eta > 0$, $\mathbb{TOOL}$ can simultaneously learn the expectation values of $\mathcal{O}_1, \ldots, \mathcal{O}_R$ to within an error of

$$(1 + \eta) \max_{i \in [R]} \|\mathcal{O}_i\|_{\mathfrak{M},\delta} \leq (1 + \eta) C \sqrt{\frac{2\log(2/\delta)}{N}}, \tag{8.20}$$

where $C = \max_{1 \leq i \leq N} \|\mathcal{O}_i - \operatorname{Tr}(\mathcal{O}_i)\mathbb{I}/d\|_{\text{shadow}}$. Setting $(1 + \eta) C \sqrt{2\log(2/\delta)/N} \leq \varepsilon$, solving for N, and using the fact that $\eta > 0$ can be made arbitrarily small gives Eq. (8.18). □

Cor. 8.6 shows that TOOL does at least as well as classical shadows. We now show that there are many observables for which TOOL performs exponentially better than classical shadows.

For this purpose, we focus on uniformly random Pauli measurements on an n-qubit system that achieves the lower bound worst-case lower bound in Thm. 8.3 (see [54]). To describe this measurement, we denote Cl_1 to be the one-qubit Clifford group. Then, by a **uniformly random Pauli measurement (URPM)**, we mean performing a $\mathscr{U}$-random unitary measurement for $\mathscr{U} = \mathrm{Cl}_1^{\otimes n}$ and $p_U = 1/|\mathscr{U}|$ for all $U \in \mathscr{U}$. To describe the effective POVM for URPM, for $k \in \{X, Y, Z\}^n$ and $b \in \{0, 1\}^n$, we denote $|k, b\rangle$ to be the bth eigenvector of Pauli P_k defined by the string k. For example, if $n = 1$, then $|X, 0\rangle = |+\rangle$, if $n = 2$, then $|ZX, 01\rangle = |0-\rangle$, and so on. Then, the effective POVM for URPM is given by the operators

$$E_{k,b} = \frac{1}{3^n} |k, b\rangle \langle k, b| , \tag{8.21}$$

for $k \in \{X, Y, Z\}^n$ and $b \in \{0, 1\}^n$.

We begin by showing that TOOL can perform exponentially better than classical shadows for learning the expectation value of a single observable.

Proposition 8.7. *For learning the expectation value of $\mathcal{O} = |Z^n, 0^n\rangle \langle Z^n, 0^n|$ to an error of $\varepsilon > 0$ and a confidence level of $1 - \delta \in (0, 1)$) using uniformly random Pauli measurements, classical shadows needs at least*

$$\Omega\left(\left(\frac{3}{2}\right)^n \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right) \tag{8.22}$$

samples, whereas TOOL needs at most

$$O\left(\left(\frac{9}{8}\right)^n \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right) \tag{8.23}$$

samples.

Proof. To obtain a lower bound on the sample complexity of classical shadows, it suffices to find a lower bound on the shadow norm. The inverse map $\mathscr{E}^{-1}$ for $\mathscr{E}$ defined in Eq. (8.2) for URPM is

given by $\mathscr{E}^{-1} = \bigotimes_{i=1}^{n} \mathscr{D}_{1/3}^{-1}$, where $\mathscr{D}_{1/3}^{-1}(A) = 3A - \mathrm{Tr}(A)\mathbb{I}$ for any 2×2 Hermitian matrix A [54]. Since $\mathscr{O}$ is a product state, its shadow norm can be expressed as

$$\|\mathscr{O}\|_{\text{shadow}}^2 = \max_{\sigma \in \mathscr{X}} \mathrm{Tr}\left[\left(\sum_{b\in\{0,1\}} \mathbb{E}_{U\in \mathrm{Cl}_1} U^\dagger |Z,b\rangle \langle Z,b| U \left\langle Z,b|U(3|Z,0\rangle\langle Z,0| - \mathbb{I})U^\dagger|Z,b\right\rangle^2\right)^{\otimes n} \sigma\right]. \tag{8.24}$$

Thus, $\|\mathscr{O}\|_{\text{shadow}}^2$ is the maximum eigenvalue of

$$\left(\sum_{b\in\{0,1\}} \mathbb{E}_{U\in \mathrm{Cl}_1} U^\dagger |Z,b\rangle \langle Z,b| U \left\langle Z,b|U(3|Z,0\rangle\langle Z,0| - \mathbb{I})U^\dagger|Z,b\right\rangle^2\right)^{\otimes n}, \tag{8.25}$$

which is equal to the nth power of the maximum eigenvalue of

$$\sum_{b\in\{0,1\}} \mathbb{E}_{U\in \mathrm{Cl}_1} U^\dagger |Z,b\rangle \langle Z,b| U \left\langle Z,b|U(3|Z,0\rangle\langle Z,0| - \mathbb{I})U^\dagger|Z,b\right\rangle^2, \tag{8.26}$$

which we now compute. To proceed, denote $A = 3|Z,0\rangle\langle Z,0| - (3/2)\mathbb{I}$, so that $3|Z,0\rangle\langle Z,0| - \mathbb{I} = A + (1/2)\mathbb{I}$ and $\mathrm{Tr}(A) = 0$. Thus, $\left\langle Z,b|U(3|Z,0\rangle\langle Z,0| - \mathbb{I})U^\dagger|Z,b\right\rangle^2 = \left\langle Z,b|UAU^\dagger|Z,b\right\rangle^2 + \left\langle Z,b|UAU^\dagger|Z,b\right\rangle + 1/4$. From [54, App. 5], we know that

$$\mathbb{E}_{U\in \mathrm{Cl}_1}\left[U^\dagger |Z,b\rangle \langle Z,b| U \left\langle Z,b|UAU^\dagger|Z,b\right\rangle^2\right] = \frac{\mathrm{Tr}(A^2)\mathbb{I} + 2A^2}{24} = \frac{3}{8}\mathbb{I} \tag{8.27}$$

for $b \in \{0,1\}$. Similarly, we have

$$\mathbb{E}_{U\in \mathrm{Cl}_1}\left[U^\dagger |Z,b\rangle \langle Z,b| U \left\langle Z,b|UAU^\dagger|Z,b\right\rangle\right] = \frac{A}{6} = \frac{1}{2}\left(|Z,0\rangle\langle Z,0| - \frac{1}{2}\mathbb{I}\right) \tag{8.28}$$

for $b \in \{0,1\}$ [54, App. 5]. Finally, we have

$$\mathbb{E}_{U\in \mathrm{Cl}_1}\left[U^\dagger |Z,b\rangle \langle Z,b| U \frac{1}{4}\right] = \frac{1}{8}\mathbb{I} \tag{8.29}$$

for $b \in \{0,1\}$. Putting these observations together, we obtain

$$\sum_{b\in\{0,1\}} \mathbb{E}_{U\in \mathrm{Cl}_1} U^\dagger |Z,b\rangle\langle Z,b| U \left(\mathrm{Tr}((3|Z,0\rangle\langle Z,0| - \mathbb{I})U^\dagger |Z,b\rangle\langle Z,b| U)\right)^2 = \frac{3}{2}|Z,0\rangle\langle Z,0| + \frac{1}{2}|Z,1\rangle\langle Z,1|, \tag{8.30}$$

which has a maximum eigenvalue of $3/2$. As a result, we have $\|\mathcal{O}\|^2_{\mathrm{shadow}} = (3/2)^n$, or $\|\mathcal{O}\|_{\mathrm{shadow}} = \sqrt{(3/2)^n}$. We also have $\|\mathbb{I}/2^n\|_{\mathrm{shadow}} = 1/2^n$. Thus, by triangle inequality and reverse triangle inequality, we have

$$\sqrt{\left(\frac{3}{2}\right)^n} - \frac{1}{2^n} \leq \left\|\mathcal{O} - \frac{1}{2^n}\mathbb{I}\right\|_{\mathrm{shadow}} \leq \sqrt{\left(\frac{3}{2}\right)^n} + \frac{1}{2^n}. \tag{8.31}$$

Since classical shadows need $\Omega(\|\mathcal{O} - \mathbb{I}/2^n\|^2_{\mathrm{shadow}} \log(1/\delta)/\varepsilon^2)$ samples to estimate the fidelity to a precision of ε and a confidence level of $1-\delta$ [54], Eq. (8.22) follows.

Next, we obtain an upper bound on the minimax norm. Using Eq. (7.2), Eq. (8.7), and Eq. (8.21), we have

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} = \frac{1}{2} \max_{\chi_1,\chi_2\in\mathcal{X}} \quad & \mathrm{Tr}(\mathcal{O}(\chi_1 - \chi_2)) \\ \text{s.t.} \quad & \sum_{k\in\{X,Y,Z\}^n} \frac{1}{3^n} \sum_{b\in\{0,1\}^n} \sqrt{p_{k,b}(\chi_1) p_{k,b}(\chi_2)} \geq \left(\frac{\delta}{2}\right)^{1/N}, \end{aligned} \tag{8.32}$$

where we denote $p_{k,b}(\chi) = \langle k,b|\chi|k,b\rangle$ for $\chi \in \mathcal{X}$. We also denote $p_k(\chi)$ to be a 2^n-dimensional (probability) vector with entries $p_{k,b}(\chi)$ for $b \in \{0,1\}^n$. Then, using Fuchs-van de Graaf inequality (Eq. (3.36)), we have $\sum_{b\in\{0,1\}^n} \sqrt{p_{k,b}(\chi_1)p_{k,b}(\chi_2)} \leq \sqrt{1 - \|p_k(\chi_1) - p_k(\chi_2)\|^2_{\mathrm{TV}}}$ for each k. By concavity of square-root, we have

$$\sum_{k\in\{X,Y,Z\}^n} (1/3^n) \sum_{b\in\{0,1\}^n} \sqrt{p_{k,b}(\chi_1)p_{k,b}(\chi_2)} \leq \sqrt{1 - \sum_{k\in\{X,Y,Z\}^n} (1/3^n)\|p_k(\chi_1) - p_k(\chi_2)\|^2_{\mathrm{TV}}}. \tag{8.33}$$

Thus, from Eq. (8.32), we have the bound

$$\begin{aligned} \|\mathcal{O}\|_{\mathfrak{M},\delta} \leq \frac{1}{2} \max_{\chi_1,\chi_2 \in \mathcal{X}} \quad & \mathrm{Tr}(\mathcal{O}(\chi_1 - \chi_2)) \\ \text{s.t.} \quad & \sum_{k \in \{X,Y,Z\}^n} \frac{1}{3^n} \|p_k(\chi_1) - p_k(\chi_2)\|_{\mathrm{TV}}^2 \leq 1 - \left(\frac{\delta}{2}\right)^{2/N}. \end{aligned} \tag{8.34}$$

To proceed, we expand $\chi_1 - \chi_2$ in the Pauli basis as

$$\chi_1 - \chi_2 = \sum_{\ell \in \{I,X,Y,Z\}^n} \beta_\ell P_\ell. \tag{8.35}$$

Fix a $k \in \{X,Y,Z\}^n$, and define $\mathcal{I}(k) = \{\ell \in \{I,X,Y,Z\}^n \mid \forall i \in [n], l_i \in \{I,k_i\}\}$, so that $|\mathcal{I}(k)| = 2^n$. Denoting $\mathbb{Z}_2 = \{0,1\}$ to be the 2-element field, for $\ell \in \{I,X,Y,Z\}^n$, define $\pi(\ell) \in \mathbb{Z}_2^n$ with $\pi_i(\ell) = 1$ if $\ell_i \neq I$ and $\pi_i(\ell) = 0$ otherwise. Also define the weight $\mathrm{wt}(\ell, b) = \langle \pi(\ell), b \rangle = \sum_{i=1}^n \pi_i(\ell) b_i \pmod 2 \in \{0,1\}$ for $\ell \in \{I,X,Y,Z\}^n$ and $b \in \{0,1\}^n$, which is an "inner product" between $\pi(\ell), b \in \mathbb{Z}_2^n$. Then, $\Delta p_{k,b} \equiv p_{k,b}(\chi_1) - p_{k,b}(\chi_2)$ can be expressed as

$$\Delta p_{k,b} = \mathrm{Tr}\left(|k,b\rangle\langle k,b| \left(\chi_1 - \chi_2\right)\right) = \sum_{\ell \in \mathcal{I}(k)} (-1)^{\mathrm{wt}(\ell,b)} \beta_\ell. \tag{8.36}$$

Therefore, the equation relating $\Delta p_{k,b}$ for $b \in \{0,1\}^n$ with β_ℓ for $\ell \in \mathcal{I}(k)$ is the Walsh-Hadamard transform. Taking its inverse, we find that

$$\beta_\ell = \frac{1}{2^n} \sum_{b \in \{0,1\}^n} (-1)^{\mathrm{wt}(\ell,b)} \Delta p_{k,b} \tag{8.37}$$

for $\ell \in \mathcal{I}(k)$.

With future calculations in mind, for a given $\ell \in \{I,X,Y,Z\}^n$, we restrict our attention to all the strings $k \in \{X,Y,Z\}^n$ obtained by replacing identity occurring in ℓ with either X or Y. This defines the set $\mathcal{K}(\ell) = \{k \in \{X,Y,Z\}^n \mid \forall i \in [n], k_i = \ell_i \text{ if } \ell_i \neq I, k_i \in \{X,Y\} \text{ if } \ell_i = I\}$ given any $\ell \in \{I,X,Y,Z\}^n$. Let $H(\ell)$ denote the Hamming weight of ℓ, i.e., the number of characters of the

string ℓ not equal to identity. Then, we have $|\mathscr{K}(\ell)| = 2^{n-H(\ell)}$, and we can write

$$\begin{aligned}\beta_\ell &= \frac{1}{|\mathscr{K}(\ell)|}\sum_{k\in\mathscr{K}(\ell)}\beta_\ell \\ &= \frac{2^{H(\ell)}}{2^n}\sum_{k\in\mathscr{K}(\ell)}\frac{1}{2^n}\sum_{b\in\{0,1\}^n}(-1)^{\mathrm{wt}(\ell,b)}\Delta p_{k,b}.\end{aligned} \tag{8.38}$$

To proceed, observe that for $\mathcal{O} = |Z^n, 0^n\rangle\langle Z^n, 0^n|$, we have

$$\mathrm{Tr}(\mathcal{O}(\chi_1 - \chi_2)) = \sum_{\ell\in\mathscr{J}(Z^n)}\beta_\ell, \tag{8.39}$$

because $\mathrm{wt}(\ell, 0^n) = 0$ for all ℓ. Since $\mathrm{Tr}(\chi_1 - \chi_2) = 0$, we must have $\beta_{I^n} = 0$. Thus, for convenience, we denote $\mathscr{J}_0(Z^n) = \mathscr{J}(Z^n)\setminus\{I^n\}$. Then, using Eq. (8.39) and Eq. (8.38), we obtain

$$\begin{aligned}\frac{1}{2}\mathrm{Tr}(\mathcal{O}(\chi_1 - \chi_2)) &= \frac{1}{2}\sum_{\ell\in\mathscr{J}_0(Z^n)}\frac{2^{H(\ell)}}{2^n}\sum_{k\in\mathscr{K}(\ell)}\frac{1}{2^n}\sum_{b\in\{0,1\}^n}(-1)^{\mathrm{wt}(\ell,b)}\Delta p_{k,b} \\ &\le \sum_{\ell\in\mathscr{J}_0(Z^n)}\frac{2^{H(\ell)}}{2^n}\sum_{k\in\mathscr{K}(\ell)}\frac{1}{2^n}\left\|\Delta p_k\right\|_{\mathrm{TV}},\end{aligned} \tag{8.40}$$

where we denote Δp_k to be the 2^n-dimensional vector with components $\Delta p_{k,b}$ for $b \in \{0,1\}^n$, and use the fact that $\sum_{b\in\{0,1\}^n}(-1)^{\mathrm{wt}(\ell,b)}\Delta p_{k,b} \le \|\Delta p_k\|_1 = 2\|\Delta p_k\|_{\mathrm{TV}}$.

Next, we show that $\{\mathscr{K}(\ell) \mid \ell \in \mathscr{J}_0(Z^n)\}$ partitions $\{X,Y,Z\}^n \setminus \{X,Y\}^n$. First, note that for $\ell, \ell' \in \mathscr{J}_0(Z^n)$, if $\ell \neq \ell'$, then there is at least one index $i \in [n]$ where $\ell_i = Z$ but $\ell'_i = I$ or $\ell'_i = Z$ but $\ell_i = I$, and as a result, $\mathscr{K}(\ell)\cap\mathscr{K}(\ell') = \varnothing$. Further, for all $k \in \{X,Y,Z\}^n \setminus \{X,Y\}^n$, we have $k \in \mathscr{K}(\ell)$ for $\ell \in \mathscr{J}_0(Z^n)$ obtained by replacing all X, Y in k with I. Thus, $\{\mathscr{K}(\ell) \mid \ell \in \mathscr{J}_0(Z^n)\}$ partitions $\{X,Y,Z\}^n \setminus \{X,Y\}^n$. Consequently, for every $k \in \{X,Y,Z\}^n \setminus \{X,Y\}^n$ belonging to a particular $\mathscr{K}(\ell)$, we can uniquely assign an $\ell \in \mathscr{J}_0(Z^n)$, which we denote as $\ell(k)$. Thus, we can

write Eq. (8.40) as

$$\begin{aligned}\frac{1}{2}\mathrm{Tr}(\mathcal{O}(\chi_1-\chi_2)) &\leq \sum_{k\in\{X,Y,Z\}^n\setminus\{X,Y\}^n} \frac{2^{H(\ell(k))}}{2^n}\frac{1}{2^n}\left\|\Delta p_k\right\|_{\mathrm{TV}} \\ &\leq \sqrt{\sum_{k\in\{X,Y,Z\}^n\setminus\{X,Y\}^n}\left(\frac{2^{H(\ell(k))}}{2^n}\frac{1}{2^n}\right)^2 3^n}\sqrt{\sum_{k\in\{X,Y,Z\}^n\setminus\{X,Y\}^n}\frac{1}{3^n}\left\|\Delta p_k\right\|_{\mathrm{TV}}^2},\end{aligned} \tag{8.41}$$

where the second inequality follows from the Cauchy-Schwarz inequality. From the above inequality, we can see that only the weight of $\ell(k)$ contributes to the pre-factor. In $\mathscr{I}_0(Z^n)$, there are $\binom{n}{i}$ strings of Hamming weight i for $i\in[n]$. Each string $\ell\in\mathscr{I}_0(Z^n)$ of Hamming weight i comes from $k\in\mathscr{K}(\ell)$, and we have $|\mathscr{K}(\ell)|=2^{n-i}$. Moreover, we know from Eq. (8.34) that we can bound

$$\sum_{k\in\{X,Y,Z\}^n\setminus\{X,Y\}^n}\frac{1}{3^n}\left\|\Delta p_k\right\|_{\mathrm{TV}}^2 \leq \sum_{k\in\{X,Y,Z\}^n}\frac{1}{3^n}\left\|\Delta p_k\right\|_{\mathrm{TV}}^2 \leq 1-\left(\frac{\delta}{2}\right)^{2/N} \tag{8.42}$$

Thus, we obtain

$$\begin{aligned}\frac{1}{2}\mathrm{Tr}(\mathcal{O}(\chi_1-\chi_2)) &\leq \sqrt{\sum_{i=1}^{n}\binom{n}{i}2^{n-i}\left(\frac{2^i}{2^n}\frac{1}{2^n}\right)^2 3^n}\sqrt{1-\left(\frac{\delta}{2}\right)^{2/N}} \\ &= \sqrt{\frac{3^n}{8^n}(3^n-1)}\sqrt{1-\left(\frac{\delta}{2}\right)^{2/N}} \\ &\leq \sqrt{\left(\frac{9}{8}\right)^n}\sqrt{1-\left(\frac{\delta}{2}\right)^{2/N}}.\end{aligned} \tag{8.43}$$

Therefore, we have

$$\|\mathcal{O}\|_{\mathfrak{M},\delta} \leq \sqrt{\left(\frac{9}{8}\right)^n}\sqrt{1-\left(\frac{\delta}{2}\right)^{2/N}}. \tag{8.44}$$

Consequently, from Thm. 7.13, we know that $\mathbb{TOOL}$ can estimate the expectation value of $\mathcal{O}$ with error at most $(1+\eta)\sqrt{(9/8)^n}\sqrt{1-(\delta/2)^{2/N}}$ for a confidence level of $1-\delta\in(0.75,1)$, for any given $\eta>0$. Thus, using Eq. (8.17), we can infer that to reach a precision of ε, we need at most

$$N = O\left(\left(\frac{9}{8}\right)^n\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right) \tag{8.45}$$

samples. □

Since the observable $\mathcal{O} = |Z^n, 0^n\rangle\langle Z^n, 0^n|$ acts non-trivially on all n-qubits, the worst-case lower bound of [54] (Thm. 8.3) is $\Omega(3^n/\varepsilon^2)$ to learn the expectation value of $\mathcal{O}$ to an error of ε. On the other hand, from Eq. (8.31), we can infer that $\Theta((3/2)^n \log(1/\delta)/\varepsilon^2)$ samples are both necessary and sufficient to learn the expectation value of $\mathcal{O}$ using classical shadows. Therefore, classical shadows already beats the lower bound of Thm. 8.3 by an exponential factor (exponential in the number of qubits). Prop. 8.7 shows that $\mathbb{TOOL}$ beats both classical shadows and the lower bound of Thm. 8.3 by an exponential factor. This shows that there are observables for which classical shadows is not optimal, and can be improved by an exponential factor.

To numerically check the bounds on the minimax norm and the shadow norm derived in Prop. 8.7, we define the rescaled minimax norm as

$$\frac{\|\mathcal{O}\|_{\mathfrak{M},\delta}}{\sqrt{1-(\delta/2)^{2/N}}}. \tag{8.46}$$

From Eq. (8.44), we have the upper bound of $\sqrt{(9/8)^n}$ on the rescaled minimax norm of $\mathcal{O}$. Since the minimax norm is invariant under translations of the observable (Prop. 7.5.1), we also have the upper bound of $\sqrt{(9/8)^n}$ on the rescaled minimax norm of $\mathcal{O} - \mathbb{I}/2^n$. On the other hand, from Eq. (8.31), we have the lower bound $\|\mathcal{O} - \mathbb{I}/2^n\|_{\text{shadow}} \geq \sqrt{(3/2)^n} - 1/2^n$. In Fig. 3, we plot the numerically computed values of the rescaled minimax norm and the shadow norm of $\mathcal{O} - \mathbb{I}/2^n$, along with bounds derived above. We can see that the bounds are valid and improve as the system size n increases. In the numerical computation of the minimax norm, we use $N = 5000$, $\delta = 0.05$, and $\epsilon_\circ = 10^{-5}$. Lem. 7.11.1 guarantees that for $\epsilon_\circ > 0$, the computed value is an upper bound on the minimax norm.

We now show that the exponential advantage of $\mathbb{TOOL}$ over classical shadows persists even if we learn the expectation values of many observables simultaneously.

Corollary 8.8. *For simultaneously learning the fidelity with all the* 6^n *(projectors onto the) eigen-*

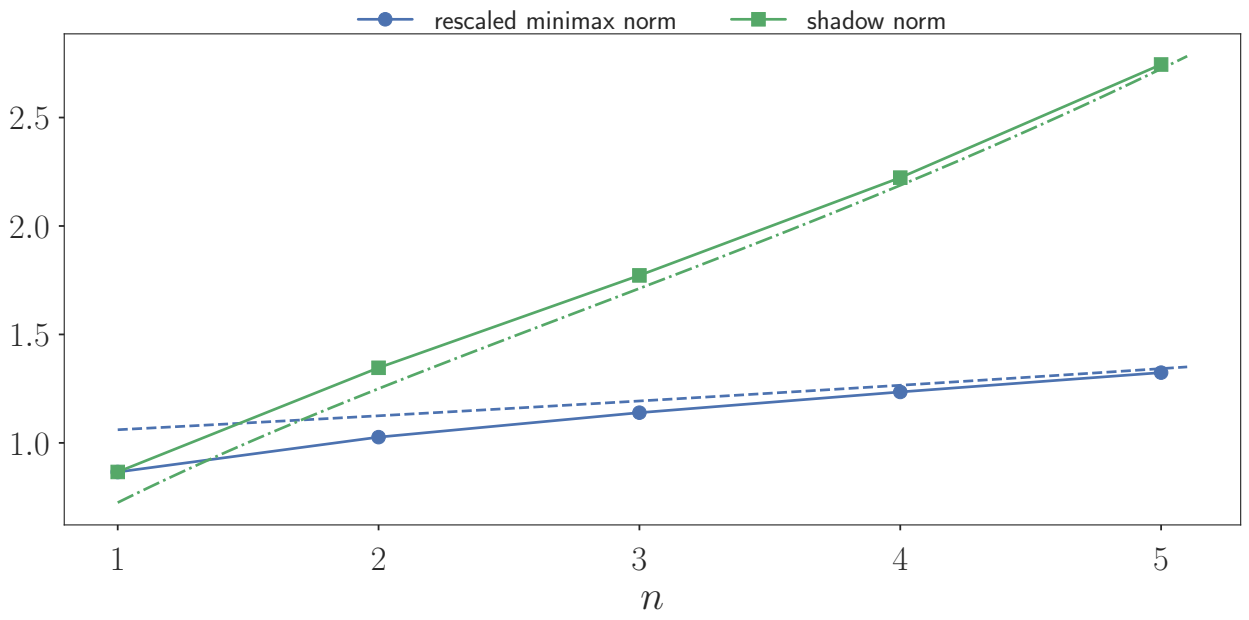

Figure 3: Plot of the rescaled minimax norm (defined in Eq. (8.46)) and the shadow norm of $|Z^n, 0^n\rangle\langle Z^n, 0^n| - \mathbb{I}/2^n$ for uniformly random Pauli measurements as a function of the number of qubits n. The minimax norm is rescaled by a factor of $\sqrt{1-(\delta/2)^{2/N}}$. The analytically computed upper bound of $\sqrt{(9/8)^n}$ on the rescaled minimax norm (dashed line) and lower bound of $\sqrt{(3/2)^n} - 1/2^n$ on the shadow norm (dot-dash line) are plotted for reference.

states $|k,b\rangle\langle k,b|$, $k \in \{X,Y,Z\}^n$ *and* $b \in \{0,1\}^n$, *of weight-n Pauli operators in an n-qubit system to an error of* $\varepsilon > 0$ *and a confidence level of* $1-\delta \in (0,1)$, *classical shadows needs at least*

$$\Omega\left(n\left(\frac{3}{2}\right)^n \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right) \tag{8.47}$$

outcomes of uniformly random Pauli measurements, while $\mathbb{TOOL}$ *needs at most*

$$O\left(n\left(\frac{9}{8}\right)^n \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right) \tag{8.48}$$

outcomes.

Proof. For all $k \in \{X,Y,Z\}^n$ and all $b \in \{0,1\}^n$, there is a unitary $U \in \mathrm{Cl}_1^{\otimes n}$ such that $|k,b\rangle\langle k,b| = U|Z^n,0^n\rangle\langle Z^n,0^n|U^\dagger$. Since $\mathrm{Cl}_1^{\otimes n}$ is a group, it can be verified from the definition of shadow norm

(Def. 8.1) that $\|\|k,b\rangle\langle k,b|\|_{\text{shadow}} = \|\|Z^n,0^n\rangle\langle Z^n,0^n|\|_{\text{shadow}}$.

Similarly, each $U \in \text{Cl}_1^{\otimes n}$ is a measurement symmetry of the uniformly random Pauli measurement (as defined in Def. 7.7), since it only permutes the POVM elements in Eq. (8.21). Thus, from Prop. 7.8, we have $\|\|k,b\rangle\langle k,b|\|_{\mathfrak{M},\delta} = \|\|Z^n,0^n\rangle\langle Z^n,0^n|\|_{\mathfrak{M},\delta}$ for all $k \in \{X,Y,Z\}^n$ and all $b \in \{0,1\}^n$.

It follows that for both classical shadows and $\mathbb{TOOL}$, learning the expectation value of any given $|k,b\rangle\langle k,b|$ for $k \in \{X,Y,Z\}^n$ and $b \in \{0,1\}^n$ using URPM needs exactly as many samples as learning the expectation value of $|Z^n,0^n\rangle\langle Z^n,0^n|$. Since there are a total of $R = 6^n$ observables whose expectation values we want to learn, for classical shadows as well as $\mathbb{TOOL}$, we learn each expectation value to a confidence level of $1 - \delta/R$ and then use the union bound to simultaneously learn them in l_∞-norm. Thus, Eq. (8.47) and Eq. (8.48) follow from Eq. (8.22) and Eq. (8.23), respectively. □

Finally, we prove a data-processing-type inequality for the minimax norm for randomized measurements. First, we show that randomizing a deterministic measurement protocol does not offer any benefits for estimation. We define what we mean by randomization of a measurement protocol below.

Definition 8.9 (Randomization of a measurement protocol). Given a measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}$, its randomization is the measurement protocol $\mathfrak{M}^{\#}$ where the POVM

$$E_{i,k} = \frac{N_i}{N} E_k^{(i)}, k \in [M_i], i \in [L] \tag{8.49}$$

is measured N times, where $N = \sum_{i=1}^{L} N_i$ denotes the total number of samples. □

The measurement protocol $\mathfrak{M}^{\#}$ can be implemented by sampling the ith POVM of $\mathfrak{M}$ with probability $p_i = N_i/N$ and measuring it, and repeating this procedure a total of N times.

Since the minimax norm provides a tight lower bound on the estimation error for any measurement protocol, we can use it to compare the performance of $\mathfrak{M}$ and $\mathfrak{M}^{\#}$ for estimation.

The result below shows that randomizing a measurement protocol according to Def. 8.9 does not offer any benefits for learning the expectation value of an observable.

Proposition 8.10. *Let $\mathfrak{M}$ be a measurement protocol and $\mathfrak{M}^{\#}$ denote its randomization according to Def. 8.9. Then, for all observables $\mathcal{O}$ and confidence levels $1-\delta \in (0,1)$, we have*

$$\|\mathcal{O}\|_{\mathfrak{M}^{\#},\delta} \geq \|\mathcal{O}\|_{\mathfrak{M},\delta}\,. \tag{8.50}$$

Proof. Given any state χ, denote $p_{\chi}^{(i)}(j) = \mathrm{Tr}(E_j^{(i)}\chi)$ to be the probability distribution defined by the ith POVM of the measurement protocol $\mathfrak{M}$. Similarly, given any state χ, denote $q_{\chi}(i,j) = (N_i/N)p_{\chi}^{(i)}(j)$ to be the probability distribution defied by the POVM in Eq. (8.49). Observe that

$$\mathrm{BC}_{\mathfrak{M}^{\#}}(\chi_1,\chi_2) = \mathrm{BC}(q_{\chi_1},q_{\chi_2}) = \sum_{i=1}^{L}\frac{N_i}{N}\mathrm{BC}(p_{\chi_1}^{(i)},p_{\chi_2}^{(i)}) \geq \prod_{i=1}^{L}[\mathrm{BC}(p_{\chi_1}^{(i)},p_{\chi_2}^{(i)})]^{N_i/N} = \mathrm{BC}_{\mathfrak{M}}(\chi_1,\chi_2).$$

Then, Eq. (8.50) follows from the expression for the minimax norm given in Eq. (7.2). □

8.2 Two no-go theorems

In this section, we present two no-go theorems (or propositions), one for fidelity estimation with stabilizer states, and another for learning the expectation values of arbitrary observables.

Our first task is to learn the fidelity with a pure stabilizer state ρ_{target}. We saw in Sec. 6.2, that if we randomly sample elements of the stabilizer group of ρ_{target}, then we can efficiently estimate the fidelity with ρ_{target}. However, we know that all the stabilizer group elements can be obtained from the generators of the stabilizer group $\mathrm{STAB}(\rho_{\mathrm{target}})$ of ρ_{target}. Therefore, one can ask the question if it suffices to measure the stabilizer group elements, and then post-process the measurement outcomes to learn the fidelity with ρ_{target}. Note that the size of the stabilizer group of an n-qubit stabilizer state is 2^n, while there are only n stabilizer generators. Thus, if we randomly sample a fixed number (independent of n) of elements of $\mathrm{STAB}(\rho_{\mathrm{target}})$, we are likely to draw different a element each time we sample. Measuring a different sampled element each time amounts to repeatedly changing the

measurement setting, which is experimentally more time consuming than measuring n stabilizer generators each a fixed number of times. Therefore, whether measuring the stabilizer generators suffices to learn the fidelity with ρ_{target} is also a practically relevant question. First, we consider the case where we measure the stabilizer generators one at a time, that is, we prepare the state ρ, measure a stabilizer generator, record the outcome, and repeat this procedure many times. It turns out that measuring the stabilizer generators in this fashion does not suffice to learn the fidelity with the target stabilizer state to a small enough precision.

Proposition 8.11. *For $n \geq 2$, there is no procedure that can learn the fidelity with an n-qubit pure stabilizer state ρ_{target} to an error $\varepsilon \leq 0.05$ and confidence level $1 - \delta \geq 0.95$ using outcomes from independent projective measurements of the stabilizer generators of ρ_{target}.*

Proof. Let $P_1, \ldots, P_n$ denote stabilizer generators of $\rho_{\text{target}} = |\psi\rangle\langle\psi|$. Note that all the stabilizer generators, and therefore, the stabilizer group elements commute. Since $|\psi\rangle$ is the eigenstate of these generators with eigenvalue $+1$ [66], we can write

$$\rho_{\text{target}} = \prod_{i=1}^{n} \frac{\mathbb{I} + P_i}{2}.$$

$(\mathbb{I} + P_i)/2$ denotes the orthogonal projector onto the $+1$-eigenvalue subspace of P_i for $i \in [n]$, and since these projectors commute, their product (which is ρ_{target}) corresponds to the projector onto the intersection of all the $+1$-eigenvalue subspaces of $P_1, \ldots, P_n$.

Taking inspiration from this observation, we construct states χ_1, χ_2 such that $\text{FC}_{\mathfrak{M}}(\chi_1, \chi_2) = 1$, while $F(\rho_{\text{target}}, \chi_1) - F(\rho_{\text{target}}, \chi_2) = 1 - 1/n$. Here, $\mathfrak{M}$ is the measurement protocol where the POVM $\{(\mathbb{I} + P_i)/2, (\mathbb{I} - P_i)/2\}$ is measured N_i times, for $i \in [n]$, which are eigenvalue/projective measurements of the stabilizer generators. To that end, define

$$\chi_1 = \left(1 - \frac{1}{n}\right) \rho_{\text{target}} + \rho_1^{\perp}$$
$$\chi_2 = \rho_2^{\perp},$$

where

$$\rho_1^{\perp} = \prod_{i=1}^{L} \frac{\mathbb{I} - P_i}{2}$$
$$\rho_2^{\perp} = \sum_{i=1}^{n} \frac{1}{n} \prod_{j=1}^{n} \frac{\mathbb{I} + (-1)^{\delta_{ij}} P_i}{2}.$$

The state $\rho_1^{\perp}$ is the projector onto the simultaneous -1 eigenstate of all $P_1, \ldots, P_n$, whereas the state $\rho_2^{\perp}$ is the uniform mixture over $i = 1, \ldots, n$ of the $+1$-eigenstate of P_j for $j \neq i$ and -1 eigenstate of P_i.

Noting that $((\mathbb{I} \pm P_i)/2)^2 = (\mathbb{I} \pm P_i)/2$ and $((\mathbb{I} + P_i)/2)((\mathbb{I} - P_i)/2) = 0$ for each $i \in [n]$, we obtain

$$\operatorname{Tr}\left[\left(\frac{\mathbb{I} + P_i}{2}\right)\chi_1\right] = \operatorname{Tr}\left[\left(\frac{\mathbb{I} + P_i}{2}\right)\chi_2\right] = 1 - \frac{1}{n}$$
$$\operatorname{Tr}\left[\left(\frac{\mathbb{I} - P_i}{2}\right)\chi_1\right] = \operatorname{Tr}\left[\left(\frac{\mathbb{I} - P_i}{2}\right)\chi_2\right] = \frac{1}{n}.$$

Then, since

$$\mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2) = \prod_{i=1}^{n} \left(\sqrt{\operatorname{Tr}\left[\left(\frac{\mathbb{I} + P_i}{2}\right)\chi_1\right] \operatorname{Tr}\left[\left(\frac{\mathbb{I} + P_i}{2}\right)\chi_2\right]} + \sqrt{\operatorname{Tr}\left[\left(\frac{\mathbb{I} - P_i}{2}\right)\chi_1\right] \operatorname{Tr}\left[\left(\frac{\mathbb{I} - P_i}{2}\right)\chi_2\right]} \right)^{N_i/N},$$

we have $\mathrm{FC}_{\mathfrak{M}}(\chi_1, \chi_2) = (\mathrm{BC}_{\mathfrak{M}}(\chi_1, \chi_2))^2 = 1$. Similarly, since $\rho_1^{\perp}$ and $\rho_2^{\perp}$ are orthogonal to ρ_{target}, we obtain

$$F(\rho_{\mathrm{target}}, \chi_1) - F(\rho_{\mathrm{target}}, \chi_2) = \operatorname{Tr}(\rho_{\mathrm{target}}(\chi_1 - \chi_2)) = 1 - \frac{1}{n}.$$

From Eq. (7.2), we have $\|\rho_{\mathrm{target}}\|_{\mathfrak{M},\delta} = 0.5 - 0.5/n$. Then, from Thm. 7.13, it follows that the error of any procedure is bounded below as $\varepsilon \geq \|\rho_{\mathrm{target}}\|_{\mathfrak{M},\delta} / \vartheta(\delta) = (1 - 1/n)/2\vartheta(\delta)$. For $n \geq 2$ and $\delta \in (0, 0.05)$, we have $1 - 1/n \geq 0.5$ and $\vartheta(\delta) < 4.6$, so that $\varepsilon > 0.05$. □

In Prop. 8.11, we only considered deterministic stabilizer generator measurements. We can use the following general strategy to show that randomized stabilizer generator measurements don't work either.

Suppose that we have shown for an observable $\mathcal{O}$, POVMs $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(L)}$, and a confidence level $1-\delta \in (0.75, 1)$, that $\|\mathcal{O}\|_{\mathfrak{M},\delta} \geq \alpha$ holds for all $N_1, \ldots, N_L$, where $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$. Our goal is to show that for any randomized measurement protocol $\mathfrak{M}'$, obtained by randomly sampling $\boldsymbol{E}^{(i)}$ with probability p_i and measuring it (and repeating this many times), we have $\|\mathcal{O}\|_{\mathfrak{M}',\delta} \geq \alpha$. To see this, let S be a positive number, and define $N_i = \lceil p_i S \rceil$ for $i \in [L]$ and $N = \sum_{i=1}^{L} N_i$. Then, $p_i \approx N_i/N$ for large enough S. Then, from Prop. 8.10, we have $\|\mathcal{O}\|_{\mathfrak{M}^{\#},\delta} \geq \|\mathcal{O}\|_{\mathfrak{M},\delta} \geq \alpha$ for all $S > 0$. Since $\mathrm{BC}(p, q)$ is continuous in p, q, we can infer that for large enough S, we have $\mathrm{BC}_{\mathfrak{M}^{\#}}(\chi_1, \chi_2) \approx \mathrm{BC}_{\mathfrak{M}'}(\chi_1, \chi_2)$ for all χ_1, χ_2. Thus, $\|\mathcal{O}\|_{\mathfrak{M}',\delta} \approx \|\mathcal{O}\|_{\mathfrak{M}^{\#},\delta} \geq \alpha$.

We now look for generalization of the ideas underlying Prop. 8.11. From Prop. 8.11, we see that although the stabilizer group elements can be written as finite products of the stabilizer generators, it is not sufficient to measure the stabilizer generators independently to learn the fidelity with the stabilizer state. This suggests that the algebra structure of $\mathbb{S}_d$ is not relevant to learning the expectation values of observables. We show below that this is indeed the case, as the linear subspace spanned by the measurement protocol determines the observables whose expectation values can be learnt with asymptotically vanishing error.

We describe our result in terms of the projection superoperator, which we define below. Let $\mathcal{U} \subseteq \mathbb{S}_d$ be a subspace. Fix an orthonormal basis $H_1, \ldots, H_{\dim \mathcal{U}}$ of $\mathcal{U}$. Then, the (orthogonal) projection superoperator $\mathscr{P}_{\mathcal{U}} \colon \mathbb{S}_d \to \mathbb{S}_d$ is the linear map

$$\mathscr{P}_{\mathcal{U}}(\mathcal{O}) = \sum_{i=1}^{\dim \mathcal{U}} \mathrm{Tr}(H_i \mathcal{O}) H_i \tag{8.51}$$

for all $\mathcal{O} \in \mathbb{S}_d$. It can be verified that the action of $\mathscr{P}_{\mathcal{U}}$ does not depend on the choice of the orthonormal basis of $\mathcal{U}$. If $\mathcal{U}^{\perp}$ denotes the orthogonal complement of $\mathcal{U}$, then $\mathscr{P}_{\mathcal{U}} + \mathscr{P}_{\mathcal{U}^{\perp}}$ is the identity map on $\mathbb{S}_d$. Thus, every observable $\mathcal{O}$ can be decomposed as $\mathscr{P}_{\mathcal{U}}(\mathcal{O}) + \mathscr{P}_{\mathcal{U}^{\perp}}(\mathcal{O})$. We can

then state our result as follows.

Proposition 8.12. *Fix an observable $\mathcal{O}$ and a measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^{L}$. If $\mathcal{O} \notin \operatorname{span}\mathfrak{M}$, then the error ε of every estimation procedure that learns the expectation value of $\mathcal{O}$ using the outcomes of $\mathfrak{M}$ to a confidence level $1-\delta \in (0.75, 1)$ for all states is bounded below as*

$$\varepsilon \geq \mathscr{c}(\delta) \frac{\left\| \mathscr{P}_{(\operatorname{span}\mathfrak{M})^\perp}(\mathcal{O}) \right\|_{\mathrm{HS}}^2}{\left\| \mathscr{P}_{(\operatorname{span}\mathfrak{M})^\perp}(\mathcal{O}) \right\|_{\infty}}, \tag{8.52}$$

independent of the total number of samples used by $\mathfrak{M}$.

Conversely, if $\mathcal{O} \in \operatorname{span}\mathfrak{M}$, then the expectation value of $\mathcal{O}$ can be learnt to any given error and confidence level using sufficiently many outcomes from measuring $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(L)}$.

Proof. First, assume that $\mathcal{O} \notin \operatorname{span}\mathfrak{M}$. Then, denoting $\mathcal{O}_0 = \mathscr{P}_{\operatorname{span}\mathfrak{M}}(\mathcal{O})$ and $\mathcal{O}_0^\perp = \mathscr{P}_{(\operatorname{span}\mathfrak{M})^\perp}(\mathcal{O})$, we have $\mathcal{O} = \mathcal{O}_0 + \mathcal{O}_0^\perp$. We have $\mathcal{O}_0^\perp \neq 0$, $\mathcal{O}_0^\perp \in (\operatorname{span}\mathfrak{M})^\perp$, and also $\operatorname{Tr}(\mathcal{O}_0^\perp) = 0$ because $\mathbb{I} \in \operatorname{span}\mathfrak{M}$. Then,

$$\chi_\pm = \frac{1}{d}\left(\mathbb{I} \pm \frac{\mathcal{O}_0^\perp}{\left\| \mathcal{O}_0^\perp \right\|_\infty} \right) \tag{8.53}$$

are quantum states, and we have $\operatorname{Tr}(\mathcal{O}(\chi_+ - \chi_-))/2 = \left\| \mathcal{O}_0^\perp \right\|_{\mathrm{HS}}^2 / \left\| \mathcal{O}_0^\perp \right\|_\infty > 0$. Furthermore, if $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}$, then we have $\operatorname{Tr}(E_k^{(i)}\chi_+) = \operatorname{Tr}(E_k^{(i)}\chi_-)$ for all $k \in [M_i]$ and all $i \in [L]$ because $\mathcal{O}_0^\perp \in (\operatorname{span}\mathfrak{M})^\perp$. As a result, we have $\mathrm{BC}_{\mathfrak{M}}(\chi_+, \chi_-) = 1$. Then, from Eq. (7.2), we obtain

$$\|\mathcal{O}\|_{\mathfrak{M},\delta} \geq \frac{\left\| \mathcal{O}_0^\perp \right\|_{\mathrm{HS}}^2}{\left\| \mathcal{O}_0^\perp \right\|_\infty} \tag{8.54}$$

for all $N_1, \ldots, N_L$. By Thm. 7.13, we have that the error of all estimation procedures using outcomes of $\mathfrak{M}$ is bounded below by $\mathscr{c}(\delta) \left\| \mathcal{O}_0^\perp \right\|_{\mathrm{HS}}^2 / \left\| \mathcal{O}_0^\perp \right\|_\infty$, independent of the number of samples.

Conversely, if $\mathcal{O} \in \operatorname{span}\mathfrak{M}$, then we can write $\mathcal{O} = \sum_{i=1}^{L} \sum_{k=1}^{M_i} \alpha_k^{(i)} E_k^{(i)}$, where each $\alpha_k^{(i)} \in \mathbb{R}$. Therefore, $\widehat{\mathcal{O}} = \sum_{i=1}^{L} \sum_{k=1}^{M_i} \alpha_k^{(i)} f_k^{(i)}$, where $f_k^{(i)}$ is the observed frequency (Eq. (5.35)) of $E_k^{(i)}$, is a bounded unbiased estimator of $\langle \mathcal{O} \rangle$. $\langle \mathcal{O} \rangle$ can therefore be estimated using Hoeffding's inequality [52] to any given error and confidence level using sufficiently many outcomes from measuring $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(L)}$.

□

We can informally summarize Prop. 8.12 by saying that the expectation value of an observable $\mathcal{O}$ can be learnt using outcomes of $\mathfrak{M}$ if and only if $\mathcal{O} \in \operatorname{span} \mathfrak{M}$. This parallels the well-known result that for learning a state in trace distance, we need to perform an informationally complete measurement. Indeed, we show that the latter result follows from Prop. 8.12.

Corollary 8.13. *If the measurement protocol $\mathfrak{M}$ is not informationally complete, then there is no estimation procedure that can learn every quantum state using outcomes of $\mathfrak{M}$ to an arbitrarily small error in trace distance with high probability.*

Proof. Suppose that we have an estimator $\widehat{\rho}$ that learns ρ using outcomes of $\mathfrak{M}$ to an error ε in trace distance with confidence level $1-\delta$. If $\mathfrak{M}$ is not informationally complete, then $\operatorname{span} \mathfrak{M} \subsetneq \mathbb{S}_d$ (see Prop. 2.4). Choose any $\mathcal{O} \in (\operatorname{span} \mathfrak{M})^{\perp}$ with $\|\mathcal{O}\|_{\infty} = 1$. Since $\mathbb{P}_{\mathfrak{M},\rho}(\|\widehat{\rho} - \rho\|_{\mathrm{tr}} \geq \varepsilon) < \delta$ by assumption, and $|\operatorname{Tr}(\mathcal{O}(\widehat{\rho} - \rho))| \leq 2\|\widehat{\rho} - \rho\|_{\mathrm{tr}}$ (see [107, Ex. (9.1.6)]), we have $\mathbb{P}_{\mathfrak{M},\rho}(|\operatorname{Tr}(\mathcal{O}\widehat{\rho}) - \operatorname{Tr}(\mathcal{O}\rho)| \geq 2\varepsilon) < \delta$. Then, by Prop. 8.12, we must have $\varepsilon \geq \mathscr{c}(\delta)\|\mathcal{O}\|_{\mathrm{HS}}^2/2$. □

Owing to Prop. 8.12 (and the terminology used in Cor. 8.13), we can call a measurement protocol $\mathfrak{M}$ informationally complete for an observable $\mathcal{O}$ if $\mathcal{O} \in \operatorname{span} \mathfrak{M}$.

Bibliography

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

[2] Atithi Acharya, Siddhartha Saha, and Anirvan M Sengupta. Shadow tomography based on informationally complete positive operator-valued measure. Physical Review A, 104(5):052418, 2021.

[3] Charalambos D Aliprantis and Kim C Border. Infinite Dimensional Analysis: A Hitchhiker's Guide. Springer, 2006.

[4] Anurag Anshu and Srinivasan Arunachalam. A survey on the complexity of learning quantum states. Nature Reviews Physics, 6(1):59–69, 2024.

[5] Koji Azuma, Sophia E. Economou, David Elkouss, Paul Hilaire, Liang Jiang, Hoi-Kwong Lo, and Ilan Tzitrin. Quantum repeaters: From quantum networks to the quantum internet. Rev. Mod. Phys., 95:045006, Dec 2023.

[6] Costin Bădescu and Ryan O'Donnell. Improved quantum data analysis. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pages 1398–1411, 2021.

[7] Bela Bauer, Sergey Bravyi, Mario Motta, and Garnet Kin-Lic Chan. Quantum algorithms for quantum chemistry and quantum materials science. Chemical Reviews, 120(22):12685–12717, 2020.

[8] Heinz H. Bauschke and Patrick L. Combettes. Correction to: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, pages C1–C4. Springer International Publishing, Cham, 2017.

[9] John S Bell. On the einstein podolsky rosen paradox. Physics Physique Fizika, 1(3):195, 1964.

[10] Charles H Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K Wootters. Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. Physical review letters, 70(13):1895, 1993.

[11] Rajendra Bhatia. Matrix analysis, volume 169. Springer Science & Business Media, 2013.

[12] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. Bulletin of the Calcutta Mathematical Society, 35:99–110, 1943.

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

[14] Kostas Blekos, Dean Brand, Andrea Ceschini, Chiao-Hui Chou, Rui-Hao Li, Komal Pandya, and Alessandro Summer. A review on quantum approximate optimization algorithm and its variants. Physics Reports, 1068:1–66, 2024.

[15] Robin Blume-Kohout. Optimal, reliable estimation of quantum states. New J. Phys., 12(4):043034, 2010.

[16] Xavier Bonet-Monroig, Ryan Babbush, and Thomas E O'Brien. Nearly optimal measurement scheduling for partial tomography of quantum states. Physical Review X, 10(3):031064, 2020.

[17] Stephen Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

[18] Gilles Brassard. Quantum communication complexity. Foundations of Physics, 33:1593–1616, 2003.

[19] Yudong Cao, Jonathan Romero, Jonathan P Olson, Matthias Degroote, Peter D Johnson, Mária Kieferová, Ian D Kivlichan, Tim Menke, Borja Peropadre, Nicolas PD Sawaya, et al. Quantum chemistry in the age of quantum computing. Chemical reviews, 119(19):10856–10915, 2019.

[20] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS), pages 574–585. IEEE, 2022.

[21] Sitan Chen, Weiyuan Gong, and Qi Ye. Optimal tradeoffs for estimating pauli observables. arXiv preprint arXiv:2404.19105, 2024.

[22] Sitan Chen, Weiyuan Gong, and Zhihan Zhang. Adaptivity can help exponentially for shadow tomography. arXiv preprint arXiv:2412.19022, 2024.

[23] Dariusz Chruściński and Gniewomir Sarbicki. Entanglement witnesses: construction, analysis and classification. Journal of Physics A: Mathematical and Theoretical, 47(48):483001, 2014.

[24] John B Conway. A course in functional analysis, volume 96. Springer, 2019.

[25] Jordan Cotler and Frank Wilczek. Quantum overlapping tomography. Physical review letters, 124(10):100401, 2020.

[26] Marcus P. da Silva, Olivier Landon-Cardinal, and David Poulin. Practical Characterization of Quantum Devices without Tomography. Phys. Rev. Lett., 107(21):210404, 2011.

[27] Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. Statistical science, pages 189–212, 1996.

[28] Arkopal Dutt, William Kirby, Rudy Raymond, Charles Hadfield, Sarah Sheldon, Isaac L Chuang, and Antonio Mezzacapo. Practical benchmarking of randomized measurement methods for quantum chemistry hamiltonians. arXiv preprint arXiv:2312.07497, 2023.

[29] Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. Nature Reviews Physics, 2(7):382–390, 2020.

[30] Andreas Elben, Steven T Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. Nature Reviews Physics, 5(1):9–24, 2023.

[31] Tim J Evans, Robin Harper, and Steven T Flammia. Scalable bayesian hamiltonian learning. arXiv preprint arXiv:1912.07636, 2019.

[32] Christopher Ferrie and Robin Blume-Kohout. Maximum likelihood quantum state tomography is inadmissible. Preprint at arXiv: 1808.01072, 2018.

[33] Steven T. Flammia and Yi-Kai Liu. Direct Fidelity Estimation from Few Pauli Measurements. Phys. Rev. Lett., 106(23):230501, 2011.

[34] Christopher A Fuchs. Distinguishability and accessible information in quantum theory. arXiv preprint quant-ph/9601020, 1996.

[35] Christopher A Fuchs and Carlton M Caves. Mathematical techniques for quantum communication theory. Open Systems & Information Dynamics, 3(3):345–356, 1995.

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

[37] Valentin Gebhart, Raffaele Santagati, Antonio Andrea Gentile, Erik M Gauger, David Craig, Natalia Ares, Leonardo Banchi, Florian Marquardt, Luca Pezzè, and Cristian Bonato. Learning quantum systems. Nature Reviews Physics, 5(3):141–156, 2023.

[38] Iulia M Georgescu, Sahel Ashhab, and Franco Nori. Quantum simulation. Reviews of Modern Physics, 86(1):153–185, 2014.

[39] Alexei Gilchrist, Nathan K Langford, and Michael A Nielsen. Distance measures to compare real and ideal quantum processes. Physical Review A—Atomic, Molecular, and Optical Physics, 71(6):062310, 2005.

[40] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Advances in quantum metrology. Nature photonics, 5(4):222–229, 2011.

[41] Nicolas Gisin, Grégoire Ribordy, Wolfgang Tittel, and Hugo Zbinden. Quantum cryptography. Reviews of modern physics, 74(1):145, 2002.

[42] Marissa Giustina, Marijn A. M. Versteegh, Sören Wengerowsky, Johannes Handsteiner, Armin Hochrainer, Kevin Phelan, Fabian Steinlechner, Johannes Kofler, Jan-Åke Larsson, Carlos Abellán, Waldimar Amaya, Valerio Pruneri, Morgan W. Mitchell, Jörn Beyer, Thomas Gerrits, Adriana E. Lita, Lynden K. Shalm, Sae Woo Nam, Thomas Scheidl, Rupert Ursin, Bernhard Wittmann, and Anton Zeilinger. Significant-loophole-free test of bell's theorem with entangled photons. Phys. Rev. Lett., 115:250401, Dec 2015.

[43] A. Goldenshluger, A. Juditsky, and A. Nemirovski. Hypothesis testing by convex optimization. Electron. J. Stat., 9(2):1645–1712, 2015.

[44] Weiyuan Gong and Scott Aaronson. Learning distributions over quantum measurement outcomes. In International Conference on Machine Learning, pages 11598–11613. PMLR, 2023.

[45] Daniel Grier, Hakop Pashayan, and Luke Schaeffer. Sample-optimal classical shadows for pure states. Quantum, 8:1373, 2024.

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

[47] Charles Hadfield, Sergey Bravyi, Rudy Raymond, and Antonio Mezzacapo. Measurements of quantum hamiltonians with locally-biased classical shadows. Communications in Mathematical Physics, 391(3):951–967, 2022.

[48] Paul R Halmos. Measure theory, volume 18. Springer, 2013.

[49] Akel Hashim, Long B Nguyen, Noah Goss, Brian Marinelli, Ravi K Naik, Trevor Chistolini, Jordan Hines, JP Marceaux, Yosep Kim, Pranav Gokhale, et al. A practical introduction to benchmarking and characterization of quantum computers. arXiv preprint arXiv:2408.12064, 2024.

[50] Carl W Helstrom. Quantum detection and estimation theory. Journal of Statistical Physics, 1:231–252, 1969.

[51] Bas Hensen, Hannes Bernien, Anaïs E Dréau, Andreas Reiserer, Norbert Kalb, Machiel S Blok, Just Ruitenberg, Raymond FL Vermeulen, Raymond N Schouten, Carlos Abellán, et al. Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres. Nature, 526(7575):682–686, 2015.

[52] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. The collected works of Wassily Hoeffding, pages 409–426, 1994.

[53] Z. Hradil. Quantum-state estimation. Phys. Rev. A, 55(3):R1561–R1564, 1997.

[54] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. Nature Physics, 16(10):1050–1057, 2020.

[55] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Efficient estimation of pauli observables by derandomization. Physical review letters, 127(3):030503, 2021.

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

[57] Luca Innocenti, Salvatore Lorenzo, Ivan Palmisano, Francesco Albarelli, Alessandro Ferraro, Mauro Paternostro, and G Massimo Palma. Shadow tomography on general measurement frames. PRX Quantum, 4(4):040328, 2023.

[58] Matteo Ippoliti. Classical shadows based on locally-entangled measurements. Quantum, 8:1293, 2024.

[59] Daniel F. V. James, Paul G. Kwiat, William J. Munro, and Andrew G. White. Measurement of qubits. Phys. Rev. A, 64(5):052312, 2001.

[60] Zhang Jiang, Amir Kalev, Wojciech Mruczkiewicz, and Hartmut Neven. Optimal fermion-to-qubit mapping via ternary trees with applications to reduced quantum states learning. Quantum, 4:276, 2020.

[61] A. Juditsky and A. Nemirovski. Near-optimal recovery of linear and N-convex functions on unions of convex sets. IMA Inf. Inference, 9(2):423–453, 2019.

[62] Anatoli B. Juditsky and Arkadi S. Nemirovski. Nonparametric estimation by convex programming. Ann. Stat., 37(5A):2278–2300, 2009.

[63] Thomas Kailath. The divergence and bhattacharyya distance measures in signal selection. IEEE transactions on communication technology, 15(1):52–60, 1967.

[64] Olav Kallenberg. Foundations of modern probability, volume 2. Springer, 1997.

[65] Robbie King, David Gosset, Robin Kothari, and Ryan Babbush. Triply efficient shadow tomography. arXiv preprint arXiv:2404.19211, 2024.

[66] Martin Kliesch. Lecture notes: Characterization, certification, and validation of quantum systems, 2020. Heinrich-Heine-Universität Düsseldorf.

[67] Nathan K. Langford. Errors in quantum tomography: diagnosing systematic versus statistical errors. New J. Phys., 15(3):035003, 2013.

[68] Yeong-Cherng Liang, Yu-Hao Yeh, Paulo EMF Mendonça, Run Yan Teh, Margaret D Reid, and Peter D Drummond. Quantum fidelity measures for mixed states. Reports on Progress in Physics, 82(7):076001, 2019.

[69] Angus Lowe and Ashwin Nayak. Lower bounds for learning quantum states with single-copy measurements. arXiv preprint arXiv:2207.14438, 2022.

[70] Filip B Maciejewski, Zbigniew Puchała, and Michał Oszmaniec. Exploring quantum average-case distances: Proofs, properties, and examples. IEEE Transactions on Information Theory, 2023.

[71] William Matthews, Stephanie Wehner, and Andreas Winter. Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding. Communications in Mathematical Physics, 291:813–843, 2009.

[72] Leon Mirsky. A trace inequality of john von neumann. Monatshefte für mathematik, 79(4):303–306, 1975.

[73] Ashley Montanaro. Quantum algorithms: an overview. npj Quantum Information, 2(1):1–8, 2016.

[74] James R. Munkres. Topology. Prentice Hall, Inc., 2000.

[75] Lawrence Narici and Edward Beckenstein. Topological vector spaces. Chapman and Hall/CRC, 2010.

[76] Yurii Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. Ekonomika i Mateaticheskie Metody, 24(3):509–517, 1988.

[77] Michael A Nielsen and Isaac L Chuang. Quantum computation and quantum information. Cambridge university press, 2010.

[78] Ryan O'Donnell and John Wright. Efficient quantum tomography. In Proceedings of the forty-eighth annual ACM symposium on Theory of Computing, pages 899–912, 2016.

[79] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. Technical University of Denmark, 7(15):510, 2008.

[80] Stefano Pirandola, Jens Eisert, Christian Weedbrook, Akira Furusawa, and Samuel L Braunstein. Advances in quantum teleportation. Nature photonics, 9(10):641–652, 2015.

[81] Stefano Pironio, Antonio Acín, Nicolas Brunner, Nicolas Gisin, Serge Massar, and Valerio Scarani. Device-independent quantum key distribution secure against collective attacks. New Journal of Physics, 11(4):045021, 2009.

[82] Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning. Cambridge university press, 2024.

[83] Joseph M Renes, Robin Blume-Kohout, Andrew J Scott, and Carlton M Caves. Symmetric informationally complete quantum measurements. Journal of Mathematical Physics, 45(6):2171–2180, 2004.

[84] Philipp Schindler, Daniel Nigg, Thomas Monz, Julio T. Barreiro, Esteban Martinez, Shannon X. Wang, Stephan Quint, Matthias F. Brandl, Volckmar Nebendahl, Christian F. Roos, Michael Chwalla, Markus Hennrich, and Rainer Blatt. A quantum information processor with trapped ions. New J. Phys., 15(12):123012, 2013.

[85] Travis L Scholten and Robin Blume-Kohout. Behavior of the maximum likelihood in quantum state tomography. New J. Phys., 20(2):023050, 2018.

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

[87] Christian Schwemmer, Lukas Knips, Daniel Richart, Harald Weinfurter, Tobias Moroder, Matthias Kleinmann, and Otfried Gühne. Systematic Errors in Current Quantum State Tomography Tools. Phys. Rev. Lett., 114(8):080403, 2015.

[88] Andrew J Scott. Tight informationally complete quantum measurements. Journal of Physics A: Mathematical and General, 39(43):13507, 2006.

[89] Akshay Seshadri and Stephen Becker. On the computation of a non-parametric estimator by convex optimization. arXiv preprint arXiv:2112.03390, 2021.

[90] Akshay Seshadri, Martin Ringbauer, Jacob Spainhour, Rainer Blatt, Thomas Monz, and Stephen Becker. github.com/akshayseshadri/minimax-fidelity-estimation, 2024.

[91] Akshay Seshadri, Martin Ringbauer, Jacob Spainhour, Rainer Blatt, Thomas Monz, and Stephen Becker. Versatile fidelity estimation with confidence. Physical Review Letters, 133(2):020402, 2024.

[92] Akshay Seshadri, Martin Ringbauer, Jacob Spainhour, Thomas Monz, and Stephen Becker. Theory of versatile fidelity estimation with confidence. Physical Review A, 110(1):012431, 2024.

[93] Lynden K Shalm, Evan Meyer-Scott, Bradley G Christensen, Peter Bierhorst, Michael A Wayne, Martin J Stevens, Thomas Gerrits, Scott Glancy, Deny R Hamel, Michael S Allman, et al. Strong loophole-free test of local realism. Physical review letters, 115(25):250402, 2015.

[94] Matt Shannon. Properties of f-divergences and f-gan training. arXiv preprint arXiv:2009.00757, 2020.

[95] Ariel Shlosberg, Andrew J Jena, Priyanka Mukhopadhyay, Jan F Haase, Felix Leditzky, and Luca Dellantonio. Adaptive estimation of quantum observables. Quantum, 7:906, 2023.

[96] Pulkit Sinha. Dimension independent and computationally efficient shadow tomography. arXiv preprint arXiv:2411.01420, 2024.

[97] Maurice Sion. On general minimax theorems. Pacific Journal of Mathematics, 8(1):171–176, 1958.

[98] Yuki Takeuchi and Tomoyuki Morimae. Verification of many-qubit states. Physical Review X, 8(2):021060, 2018.

[99] Jules Tilly, Hongxiang Chen, Shuxiang Cao, Dario Picozzi, Kanav Setia, Ying Li, Edward Grant, Leonard Wossnig, Ivan Rungger, George H Booth, et al. The variational quantum eigensolver: a review of methods and best practices. Physics Reports, 986:1–128, 2022.

[100] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. submitted to SIAM Journal on Optimization, 2(3), 2008.

[101] Paul Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. Mathematical Programming, 125(2):263–295, 2010.

[102] Armin Uhlmann. The "transition probability" in the state space of a *-algebra. Reports on Mathematical Physics, 9(2):273–279, 1976.

[103] Kaito Wada, Naoki Yamamoto, and Nobuyuki Yoshioka. Heisenberg-limited adaptive gradient estimation for multiple observables. arXiv preprint arXiv:2406.03306, 2024.

[104] Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. arXiv preprint arXiv:1309.1541, 2013.

[105] Yunfei Wang and Junyu Liu. A comprehensive review of quantum machine learning: from nisq to fault tolerance. Reports on Progress in Physics, 2024.

[106] John Watrous. The theory of quantum information. Cambridge university press, 2018.

[107] Mark M Wilde. From classical to quantum shannon theory. arXiv preprint arXiv:1106.1445, 2011.

[108] John Wright. How to learn a quantum state. PhD thesis, Carnegie Mellon University, 2016.

List of symbols

$\mathbb{N}$	Set of natural numbers (excluding zero)	p.14
$\mathbb{R}$	Set of real numbers	p.14
$\overline{\mathbb{R}}$	Set of extended real numbers, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$	p.14
$\mathbb{C}$	Set of complex numbers	p.14
$[M]$	$[M] = \{1, \ldots, M\}, M \in \mathbb{N}$	p.14
δ_{ij}	$\delta_{ij} = 1$ if $i = j$ and zero otherwise	p.14
$\{O_i\}_{i \in \mathscr{I}}$	Indexed family; elements O_i indexed by elements of the set $\mathscr{I}$	p.14
$\langle \cdot, \cdot \rangle$	Inner product	p.15, 19
span	Linear span	p.15
$\oplus$	Direct sum	p.16, 17
$\otimes$	Tensor product or direct product	p.16
$\mathscr{V}/\mathscr{U}$	Quotient space of the vector space $\mathscr{V}$ by the subspace $\mathscr{U} \subseteq \mathscr{V}$	p.16
$\mathscr{A} + \mathscr{B}$	Minkowski sum of subsets $\mathscr{A}, \mathscr{B}$ of a vector space	p.16
ker	Kernel of a linear map	p.17
range	Range of a function	p.17
$\|\cdot\|$	Seminorm or norm	p.18
$\mathbb{K}^n$	Set of n-dimensional vectors with entries from $\mathbb{K}$; $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$	p.18
$\mathbb{K}^{m \times n}$	Set of $m \times n$ matrices with entries from $\mathbb{K}$; $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$	p.18
$\mathbb{S}_n$	Set of $n \times n$ Hermitian matrices	p.18
$\lambda(A)$	Vector of eigenvalues of a (Hermitian) matrix A	p.18
$\lambda_{\max}(A), \lambda_{\min}(A)$	Maximum, minimum eigenvalue of a (Hermitian) matrix A	p.18
$\sigma(A)$	Vector of singular values of a matrix A	p.19
$\sigma_{\max}(A), \sigma_{\min}(A)$	Maximum, minimum singular value of a matrix A	p.19
$\|\cdot\|_p$	l_p or p-norm of a vector; Schatten p-norm of a matrix; $p \in [1, \infty]$	p.19

$\|A\|_{\mathrm{HS}}$	Hilbert-Schmidt norm or Schatten-2 norm of a matrix $A \in \mathbb{S}_n$	p.19
$\|A\|_1$	Trace norm or Schatten-1 of a matrix $A \in \mathbb{S}_n$	p.19
$\|A\|_\infty$	Operator or spectral norm or Schatten-∞ norm of a matrix $A \in \mathbb{S}_n$	p.20
$\langle \mathcal{O} \rangle$	Expectation value of the observable $\mathcal{O}$ with respect to some underlying state ρ; $\langle \mathcal{O} \rangle = \mathrm{Tr}(\mathcal{O}\rho)$	p.20
POVM $\boldsymbol{E}$	Positive operator-valued measure, $\boldsymbol{E} = \{E_1, \ldots, E_M\}$, $(\forall k) E_k \geq 0$, $\sum_{k=1}^M E_k = \mathbb{I}$. $k \in [M]$ is referred to as the label of E_k.	p.21
$p_{\boldsymbol{E},\rho}$	Probability distribution over the labels upon measuring the POVM $\boldsymbol{E}$ with respect to the state ρ given by Born's rule	p.21
$p_\rho^{(i)}$	$p_{\boldsymbol{E}^{(i)},\rho}$, when the POVM $\boldsymbol{E}^{(i)}$ is understood	p.21
$\mathfrak{M}$	Measurement protocol $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^L$, where for each $i \in [L]$, the POVM $\boldsymbol{E}^{(i)}$ is measured N_i times. The POVMs are assumed to be distict.	p.22
$N(\mathfrak{M})$	Total number of samples used by $\mathfrak{M}$; if $\mathfrak{M} = \{(\boldsymbol{E}^{(i)}, N_i)\}_{i=1}^L$, then $N(\mathfrak{M}) = \sum_{i=1}^L N_i$	p.22
$\mathbb{P}_{\mathfrak{M},\rho}$	Joint probability over the labels determined by the measurement protocol $\mathfrak{M}$, and the state ρ; as a vector, $\mathbb{P}_{\mathfrak{M},\rho} = \otimes_{i=1}^L (p_\rho^{(i)})^{\otimes N_i}$	p.22
$\operatorname{span} \mathfrak{M}$	Span of the POVM elements in $\mathfrak{M}$; $\operatorname{span}\{E_k^{(i)} \mid k \in [M_i], i \in [L]\}$	p.24
$\operatorname{int} A$	Interior of the set A	p.26
$\operatorname{cl} A$	Closure of the set A	p.26
$(\Omega, B(\Omega))$	Borel space, Polish space	p.26, 27
m	measure, σ-finite reference measure	p.27, 28
$m_1 \ll m_2$	m_1 is absolutely continuous with respect to m_2	p.27
dm_1/dm_2	Radon-Nikodym derivative of m_1 with respect to m_2	p.27
$\mathbb{P}$	Probability measure or probability distribution	p.27
$\mathbb{E}[X]$	Expected value of the random variable X	p.28
Δ_d	Standard simplex in $\mathbb{R}^d$; set of discrete probability distributions over d symbols	p.29
$\operatorname{supp} p$	Support of a discrete distribution p	p.29
D_f	f-divergence	p.29
$\mathrm{KL}(\mathbb{P}\|\mathbb{Q})$	KL divergence between $\mathbb{P}$ and $\mathbb{Q}$	p.30
$\operatorname{aff} \mathscr{K}$	Affine hull of the set $\mathscr{K}$	p.33
$\operatorname{conv} \mathscr{K}$	Convex hull of the set $\mathscr{K}$	p.33
$\operatorname{relint} \mathscr{K}$	Relative interior of the set $\mathscr{K}$	p.34

lsc	Lower semicontinuous	p.34
usc	Upper semicontinuous	p.34
$\chi_{\mathscr{K}}$	Characteristic function of the set $\mathscr{K}$; $\chi_{\mathscr{K}}(x) = 0$ if $x \in \mathscr{K}$ and ∞ otherwise	p.35
$S_{\mathscr{K}}$	Support function of the set $\mathscr{K}$	p.35
f^*	Convex conjugate of the function f	p.35
BC	Bhattacharyya coefficient between two probability distributions	p.40
FC	Classical fidelity between two probability distributions; $\mathrm{FC}(p,q) = (\mathrm{BC}(p,q))^2$	p.40
BD	Bhattacharyya distance between two probability distributions; $\mathrm{BD}(p,q) = -\log(\mathrm{BC}(p,q))$	p.40
$\mathrm{BD}_{\mathfrak{M}}$	Average Bhattacharyya distance between two states determined by the measurement protocol $\mathfrak{M}$	p.41
$\mathrm{BC}_{\mathfrak{M}}$	Geometric-average Bhattacharyya coefficient between two states determined by the measurement protocol $\mathfrak{M}$	p.42
$\mathrm{FC}_{\mathfrak{M}}$	Geometric-average classical fidelity between two states determined by the measurement protocol $\mathfrak{M}$; $\mathrm{FC}_{\mathfrak{M}}(\rho,\sigma) = \mathrm{BC}^2_{\mathfrak{M}}(\rho,\sigma)$	p.42
$F(\rho,\sigma)$	Fidelity between the states ρ and σ	p.42
$\|p-q\|_{\mathrm{TV}}$	Total variation distance between the probability distributions p and q	p.49
$\mathrm{HD}(p,q)$	Hellinger distance between the probability distributions p and q	p.49
$\mathrm{SDC}(p,q)$	Classical sine distance between the probability distributions p and q	p.49
$\|\rho-\sigma\|_{\mathfrak{M},\mathrm{avg}}$	Average total variation distance between the states ρ and σ	p.50
$\|\rho-\sigma\|_{\mathfrak{M},\max}$	Maximum total variation distance between the states ρ and σ	p.50
$\mathrm{HD}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)$	Average Hellinger distance between the states ρ and σ	p.50
$\mathrm{HD}_{\mathfrak{M},\max}(\rho,\sigma)$	Maximum Hellinger distance between the states ρ and σ	p.50
$\mathrm{SDC}_{\mathfrak{M},\mathrm{avg}}(\rho,\sigma)$	Average classical sine distance between the states ρ and σ	p.50
$\mathrm{SDC}_{\mathfrak{M},\max}(\rho,\sigma)$	Maximum classical sine distance between the states ρ and σ	p.50
$\|\rho-\sigma\|_{\mathrm{tr}}$	Trace distance between the states ρ and σ	p.52
$\mathrm{D}_{\mathrm{Bur}}(\rho,\sigma)$	Bures distance between the states ρ and σ	p.52
$\mathrm{SD}(\rho,\sigma)$	Sine distance between the states ρ and σ	p.52

Index

ϵ-minimax, 32
δ-risk, 61
σ-algebra, 25

absolutely continuous, 27
affine estimator, 60
affine function, 34
affine set, 33
alphabet, 29
average Bhattacharyya distance, 41

Bhattacharyya coefficient, 40
Bhattacharyya distance, 40
Borel space, 27
Bures distance, 52

characteristic function, 35
classical fidelity, 40
classical sine distance, 49
closure, 26
coercive, 34
compact, 26
complementary slackness, 37
complete, 26
concave function, 34
confidence interval, 32
convex conjugate, Legendre-Fenchel transform, 35
convex function, 34
convex hull, 33
convex set, 33

dense, 26
discrete distribution, 29
discrete topology, 28
dual feasibility, 37
dual function, 37
dual optimal value, 37
dual problem, 37
dual variable, 37

effective POVM, 23
event, 28

fidelity, 42
Fuchs-van de Graaf inequality, 53

geometric-average Bhattacharyya coefficient, 42
geometric-average classical fidelity, 42
good pair, 61

Hellinger distance, 49

indexed family, 14
informationally complete (IC), 24
interior, 26
isometry, 17
isomorphism, 17

Karush-Kuhn-Tucker (KKT) conditions, 37

label (of a POVM), 21
Lagrangian, 36
log-concave function, 34
lower semi-continuous, lsc, 34

measurable function, 27
measurable space, measurable set, 25
measure, 27
measurement protocol, 22
metric space, 26
minimax optimal risk, 62

observable, 20
orthogonal complement, 16

parametric density family, 60

perspective (of a) function, 36
Polish space, 26
positive semidefinite (PSD), 17
primal feasibility, 37
primal optimal value, 36
primal problem, 36
primal variable, 37
probability density function, 28
probability measure, probability distribution, 27
probability space, 27
proper function, 34
pseudometric, 51
pure state, 20

quantum state, 20
quotient space, 16

randomized measurement, 23
rank, 19
reference measure, 28
relative interior, 34
relatively open, 34

sample complexity, 33
seminorm, 18
separable, 26
SIC-POVM, 25
simplex, standard simplex, 29
sine distance, 52
Slater's condition, 38
stationarity, 37
strong duality, 37
subspace, 16
support, 18
support function, 35
support of a distribution, 29
symbol, 29

topology, 26
total variation distance, 49
trace distance, 52

upper semi-continuous, usc, 34

weak duality, 37