

# Universal Sample Complexity Bounds in Quantum Learning Theory via Fisher Information Matrix

Hyukgun Kwon<sup>1,\*</sup>, Seok Hyung Lee<sup>2,†</sup>, and Liang Jiang<sup>3,‡</sup>

<sup>1</sup>*Department of Physics and Astronomy, Sejong University,*

*209 Neungdong-ro Gwangjin-gu, Seoul 05006, Republic of Korea*

<sup>2</sup>*Department of Physics, Ulsan National Institute of Science and Technology (UNIST),  
Ulsan 44919, Republic of Korea*

<sup>3</sup>*Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA*



(Received 9 March 2026; revised 27 May 2026; accepted 10 August 2026; published 22 September 2026)

In this work, we show that the sample complexity required in quantum learning theory within a general parametric framework is fundamentally governed by the inverse Fisher information matrix. More specifically, we derive upper and lower bounds on the number of samples required to estimate the parameters of a quantum system within a prescribed small additive error, with high success probability under maximum-likelihood estimation. Notably, both the upper and lower bounds are determined by the supremum of the maximum diagonal entry of the inverse Fisher information matrix, differing only by a logarithmic factor in the number of parameters to be estimated. We then apply the general bounds to Pauli channel learning and Pauli expectation value learning, which serve as representative tasks in quantum channel and state learning, respectively, in the asymptotic small-error regime. Furthermore, we identify the structural origin of exponential sample complexity in Pauli channel learning without entanglement and in Pauli expectation value learning without quantum memory by comparing the quantum Fisher information matrix and the classical Fisher information matrix. We then extend the analysis to an error criterion based on the Euclidean distance between the true parameter values and their estimators, deriving the corresponding upper and lower bounds on the sample complexity, which are likewise characterized by the inverse Fisher information matrix. As an application, we consider Pauli channel learning with entangled probes. We highlight two fundamental contributions to quantum learning theory. First, we establish a systematic framework that determines the task-independent sample complexity under maximum-likelihood estimation. Second, we show that, in the small-error regime, the learning sample complexity is governed by the inverse Fisher information matrix, which is the central quantity in quantum metrology that determines the ultimate achievable mean squared error.

DOI: [10.1103/8k5v-ddtw](https://doi.org/10.1103/8k5v-ddtw)

## I. INTRODUCTION

Characterizing quantum systems is a central prerequisite for the advancement of quantum science and technology. In particular, in quantum information science, accurate characterization enables hardware benchmarking [1,2], facilitates reliable noise modeling for quantum devices [3,4], and informs the design of quantum error correction and error mitigation protocols [5–10].

In this context, quantum learning theory establishes a systematic framework for estimating the parameters that characterize an unknown quantum state or channel of interest [11–20]. More specifically, quantum learning theory aims to efficiently estimate the parameters within a prescribed *additive error*  $\epsilon$  with *success probability* at least  $1 - \delta$ , a requirement commonly referred to as the  $(\epsilon, \delta)$ -criterion. Within this framework, the central objective is to design an efficient estimation protocol that reduces the number of accesses to the given quantum state or channel required to satisfy the  $(\epsilon, \delta)$ -criterion. For a chosen estimation protocol, this quantity is referred to as the *sample complexity*, i.e., the number of state copies or channel uses. Recent results demonstrate that appropriate quantum resources can significantly reduce the sample complexity. In particular, the simultaneous access to multiple copies of unknown quantum states and collective

\*Contact author: [kwon37hg@sejong.ac.kr](mailto:kwon37hg@sejong.ac.kr)

†Contact author: [seokhyung@unist.ac.kr](mailto:seokhyung@unist.ac.kr)

‡Contact author: [liangjiang@uchicago.edu](mailto:liangjiang@uchicago.edu)

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

measurements on these states—simply referred to as the use of quantum memory—can lead to an exponential reduction in the number of samples required for tasks such as learning expectation values of Pauli observables [12,17,21,22], the characteristic function of a bosonic state [23], and the Pauli transfer matrix of a quantum channel [24], compared to protocols that do not employ quantum memory. Moreover, entanglement provides advantages in learning the Pauli eigenvalues of a Pauli channel [19,25–27] and the probability distribution of a random displacement channel [18,28].

However, despite these advances, a comprehensive and systematic framework for characterizing the sample complexity of designed estimation strategies has yet to be established. Existing studies derive sample complexity bounds in a task-dependent manner, invoking distinct proof techniques and information-theoretic quantities tailored to each specific setting. As a result, a unifying method for systematically characterizing the required number of samples in general quantum learning tasks remains elusive. This observation motivates the following fundamental open question: *Is there a unified framework for characterizing the sample complexity of quantum learning problems in general settings?*

Beyond general sample complexity, an additional conceptual question concerns the relationship between quantum learning theory and quantum metrology. Although both fields address parameter estimation in quantum systems, they are typically formulated in terms of distinct performance measures. Quantum metrology aims to achieve quantum-enhanced precision in parameter estimation by exploiting quantum resources [29–31], with performance quantified through the mean squared error of estimators. The fundamental precision limit is determined by the inverse of the Fisher information matrix via the Cramér-Rao bound, which characterizes the minimum achievable mean squared error. On the other hand, quantum learning theory evaluates performance in terms of the sample complexity required to satisfy the  $(\epsilon, \delta)$ -criterion, rather than the mean squared error. This raises a natural fundamental question: *Can the sample complexity required to ensure the  $(\epsilon, \delta)$ -criterion also be characterized in terms of the inverse Fisher information matrix?*

In this work, we establish that the sample complexity required to achieve the  $(\epsilon, \delta)$ -criterion is fundamentally governed by the inverse Fisher information matrix. We consider  $(\epsilon, \delta)$ -criterion-based learning under maximum-likelihood estimation and derive both upper and lower bounds on the sample complexity required to estimate all parameters of a quantum system within additive error  $\epsilon$  with success probability at least  $1 - \delta$ , formalized as  $\ell_\infty$ -distance-based  $(\epsilon, \delta)$ -criterion. Our analysis demonstrates that, for sufficiently small  $\epsilon$ , both bounds are determined by the supremum over the parameter space of the largest diagonal entry of the inverse Fisher

information matrix. These results highlight that the inverse Fisher information matrix is the key quantity governing the sample complexity required to guarantee the  $\ell_\infty$ -distance-based  $(\epsilon, \delta)$ -criterion.

As applications, we consider the learning of Pauli eigenvalues of a given Pauli channel and the learning of Pauli expectation values of a given quantum state, focusing on the asymptotic regime  $\epsilon \rightarrow 0$ . First, based on the general upper and lower bounds established above, we analyze the task of learning the Pauli eigenvalues of a given Pauli channel. We show that the use of entanglement reduces the sample complexity from exponential to polynomial in the number of qubits, a result originally established in Refs. [19,25] using different proof techniques from our FIM approach. Notably, we identify the origin of the exponential sample complexity in the absence of entanglement. A quantum probe must satisfy the purity constraint, which confines its Bloch vector to lie within the Bloch sphere. Crucially, because of this constraint, along certain parameter directions the allowable Bloch vector components are necessarily exponentially small. Consequently, the corresponding diagonal elements of the inverse Fisher information matrix grow exponentially with the number of qubits, which directly results in exponential sample complexity. The necessity of exponential sample complexity for entanglement-free Pauli eigenvalue learning was originally established in Refs. [19,25,27]; our analysis complements these results by providing an FIM-based explanation of its statistical origin. Second, we analyze the task of learning Pauli expectation values. We show that access to quantum memory reduces the sample complexity from exponential to polynomial in the number of qubits, a result originally established in Ref. [17] using a different proof technique from our FIM approach. We further clarify the origin of the exponential cost in the absence of quantum memory. The optimal measurement for estimating the expectation value of a given Pauli operator is a projective measurement in its eigenbasis. However, distinct Pauli operators generally do not commute, and therefore their respective optimal measurements are mutually incompatible. This measurement incompatibility necessarily induces exponential growth in the relevant diagonal elements of the inverse Fisher information matrix, which in turn implies exponential sample complexity. Lastly, we emphasize that Ref. [22] also related these sample-complexity limitations to the commutation structure of Pauli observables using information-theoretic techniques. Our analysis complements this characterization by showing how the associated measurement incompatibility leads to exponentially large diagonal entries of the inverse Fisher information matrix.

Finally, we extend our analysis to  $\ell_2$ -distance-based  $(\epsilon, \delta)$ -criterion, where the estimation error is measured by the Euclidean norm of the difference between the true parameter and the estimator and is required to be bounded by an additive error  $\epsilon$  with high probability at least  $1 - \delta$ .

For this criterion, we derive upper and lower bounds on the sample complexity, which are also characterized by the inverse Fisher information matrix. We then apply our bounds to the task of learning the Pauli eigenvalues of a given Pauli channel using a maximally entangled state as a quantum probe. We show that, when the probe state and the Pauli channel are each used only once, the required sample complexity grows exponentially, in contrast to the  $\ell_\infty$  case.

We emphasize that our results answer the two fundamental questions posed above. First, we resolve the important open problem of characterizing the task-independent sample complexity of quantum learning tasks under the assumption of maximum-likelihood estimation. This yields a general framework that provides new analytical tools for systematically characterizing the sample complexity of general learning protocols. Second, we establish a quantitative connection between quantum metrology and quantum learning: both metrological precision and learning sample complexity are determined by the inverse Fisher information matrix.

This paper is organized as follows. In Sec. II, we review the necessary background on quantum metrology, with a focus on the Fisher information matrix, and on quantum learning theory, including the definition of sample complexity. In Sec. III, we establish general upper and lower bounds on the sample complexity required to satisfy the  $\ell_\infty$ -distance-based  $(\epsilon, \delta)$ -criterion, and apply these bounds to Pauli channel learning and Pauli expectation value learning. In Sec. IV, we derive corresponding upper and lower bounds for the sample complexity under the  $\ell_2$ -distance-based  $(\epsilon, \delta)$ -criterion, and apply the bounds to Pauli channel learning.

## II. PRELIMINARIES OF MULTI-PARAMETER ESTIMATION

### A. Quantum parameter estimation

Let us consider a parametrized quantum channel  $\mathcal{E}_\theta$  where

$$\theta = (\theta_1, \theta_2, \dots, \theta_d)^T \in \Theta \subset \mathbb{R}^d \quad (1)$$

denotes a set of unknown parameters to be estimated,  $d$  is a finite positive integer, and  $\Theta$  denotes the parameter space. To estimate  $\theta$ , one prepares a quantum probe  $\hat{\rho}_0$  and sends it through the channel, resulting in the output state  $\hat{\rho}_\theta := \mathcal{E}_\theta(\hat{\rho}_0)$  that encodes the parameters  $\theta$ . To extract the information about  $\theta$  from  $\hat{\rho}_\theta$ , a measurement described by a positive-operator-valued-measure (POVM) is performed, where each element of  $\{\hat{\Pi}_x\}_x$  corresponds to a possible measurement outcome  $x$  and the elements satisfy the completeness relation  $\sum_x \hat{\Pi}_x = \hat{I}$ . The measurement is repeated independently  $M$  times, resulting in a sequence of measurement outcomes  $\mathbf{x} := (x_1, x_2, \dots, x_M)^T$ . The joint probability distribution for obtaining the measurement

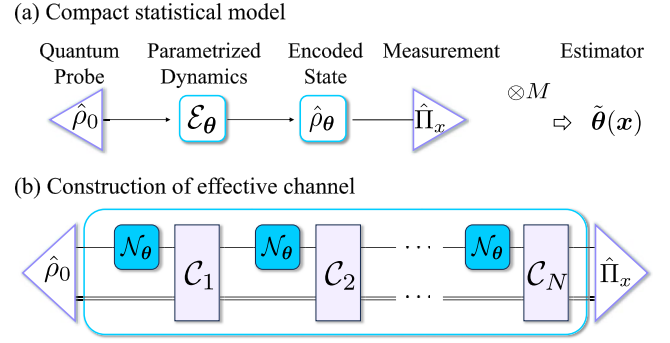


FIG. 1. Schematic of parameter estimation. (a) Compact statistical model for a quantum parameter estimation protocol. A probe state is independently transformed by the same effective parameter-dependent channel and measured over  $M$  repetitions. The resulting  $M$  measurement outcomes are then collectively processed to construct an estimator of the parameter. (b) Construction of the effective channel  $\mathcal{E}_\theta$ . The effective channel is realized by inserting multiple uses of the elementary parameter-dependent channel  $\mathcal{N}_\theta$  into the slots of a  $\theta$ -independent strategy specified by the quantum operations  $\mathcal{C}_i$ . The double dashed line denotes the ancillary system.

outcomes  $\mathbf{x}$  is given by

$$p_\theta(\mathbf{x}) := \prod_{i=1}^M p_\theta(x_i), \quad (2)$$

where  $p_\theta(x) := \text{Tr}[\hat{\Pi}_x \hat{\rho}_\theta]$  is the probability of obtaining outcome  $x$  for a single measurement. Based on the measurement outcomes  $\mathbf{x}$ , we construct estimators for the parameters, denoted by

$$\tilde{\theta}(\mathbf{x}) = (\tilde{\theta}_1(\mathbf{x}), \tilde{\theta}_2(\mathbf{x}), \dots, \tilde{\theta}_d(\mathbf{x}))^T. \quad (3)$$

[See Fig. 1(a) for a schematic illustration of the parameter estimation procedure.]

Notably, the compact statistical model above is not restricted to a fixed physical probe state, a single use of the parameter-dependent channel, or a fixed measurement. Rather, it should be understood as an effective statistical description of one execution of a general finite-resource quantum estimation protocol.

To make the connection with concrete quantum protocols explicit, let  $\mathcal{N}_\theta$  denote the elementary parameter-dependent channel to which the learner has access. To estimate  $\theta$  efficiently, one may construct an effective channel  $\mathcal{E}_\theta$  by invoking  $\mathcal{N}_\theta$  multiple times, possibly together with ancillary systems. These invocations may be arranged in parallel, sequentially, or adaptively, and may be interleaved with arbitrary known  $\theta$ -independent operations, as captured by the standard quantum-comb or quantum-strategy framework [32–37]. Accordingly, one execution of the constructed estimation protocol can be described by an effective channel  $\mathcal{E}_\theta$ , which can be

expressed schematically as

$$\mathcal{E}_\theta = \mathcal{C}_{N_\mathcal{E}} \circ (\mathcal{N}_\theta \otimes \mathcal{I}) \circ \mathcal{C}_{N_\mathcal{E}-1} \circ \cdots \circ \mathcal{C}_1 \circ (\mathcal{N}_\theta \otimes \mathcal{I}). \quad (4)$$

[See Fig. 1(b) for a schematic illustration of the construction of the effective channel  $\mathcal{E}_\theta$ .] Here,  $N_\mathcal{E}$  is the number of uses of the elementary channel  $\mathcal{N}_\theta$  required to implement one execution of the effective channel  $\mathcal{E}_\theta$ . The quantum operations  $\mathcal{C}_i$  are known  $\theta$ -independent CPTP maps. They may include state preparation, ancillary systems, coherent controls, quantum instruments, intermediate measurements, classical memory, and feedforward. The identity channel  $\mathcal{I}$  acts on systems that are not acted upon by  $\mathcal{N}_\theta$ , such as ancillary systems or classical registers. After the effective channel is applied, a final POVM  $\{\hat{\Pi}_x\}_x$  is performed. Hence the induced probability distribution takes the compact form

$$p_\theta(x) = \text{Tr}[\hat{\Pi}_x \mathcal{E}_\theta(\hat{\rho}_0)], \quad (5)$$

where  $\hat{\rho}_0$  is the input state of the effective channel  $\mathcal{E}_\theta$ . We emphasize that different choices of the estimation protocol generally lead to different effective channels  $\mathcal{E}_\theta$ , and therefore to different induced probability distributions  $p_\theta(x)$  and estimation performance. Examples of this dependence are discussed for Pauli-channel estimation and Pauli-expectation-value estimation in Secs. III C and III D, respectively. Lastly, we note that while we assume a fixed number  $N_\mathcal{E}$  of channel uses per protocol execution, the formulation can be extended to more general protocols in which the number of channel uses is determined probabilistically and may therefore vary from one execution to another. In such cases, the channel-use cost may be quantified either by its average value or by the maximum number of channel uses required in the worst case.

## B. Quantum metrology and mean squared error

Under the parameter estimation scheme introduced in Sec. II A, quantum metrology aims to reduce the mean squared error [29–31, 38–40]

$$\delta^2 \theta_i := \sum_x p_\theta(x) (\tilde{\theta}_i(x) - \theta_i)^2, \quad (6)$$

beyond the limits achievable by classical strategies. Such an enhancement is enabled by exploiting quantum resources, including the preparation of an appropriate quantum probe state  $\hat{\rho}_0$ , the construction of the effective channel  $\mathcal{E}_\theta$  from the elementary channel  $\mathcal{N}_\theta$ , and the choice of POVM  $\{\hat{\Pi}_x\}_x$ .

For a fixed estimation protocol, the mean squared error can be quantified by the mean-squared error matrix formalism. Specifically, the mean squared error is characterized by the  $d \times d$  mean-squared error matrix  $\Sigma$ , whose elements are defined as

$$[\Sigma]_{ij} = \sum_x p_\theta(x) (\tilde{\theta}_i(x) - \theta_i)(\tilde{\theta}_j(x) - \theta_j). \quad (7)$$

When all the estimators satisfy the unbiasedness condition

$$\sum_x p_\theta(x) \tilde{\theta}_i(x) = \theta_i, \quad (8)$$

the multi-parameter quantum Cramér-Rao matrix inequality establishes a fundamental lower bound on the mean-squared error matrix [41–44]:

$$\Sigma \geq \frac{1}{M} \mathbf{F}_\theta^{-1}(\{\hat{\Pi}_x\}_x) \geq \frac{1}{M} \mathbf{J}_\theta^{-1}. \quad (9)$$

Here,  $\mathbf{F}_\theta(\{\hat{\Pi}_x\}_x)$  is the Fisher information matrix (FIM) associated with the POVM  $\{\hat{\Pi}_x\}_x$ , while  $\mathbf{J}_\theta$  is the quantum Fisher information matrix (QFIM), which provides an ultimate, measurement-independent bound. Explicitly, the FIM and QFIM are defined as

$$[\mathbf{F}_\theta(\{\hat{\Pi}_x\}_x)]_{ij} = \sum_x \frac{\partial_i p_\theta(x) \partial_j p_\theta(x)}{p_\theta(x)}, \quad (10)$$

$$[\mathbf{J}_\theta]_{ij} := \frac{1}{2} \text{Tr}[\hat{\rho}_\theta \{\hat{L}_i, \hat{L}_j\}], \quad (11)$$

where  $\hat{L}_i$  are the symmetric logarithmic derivative (SLD) operators, satisfying

$$\frac{\partial \hat{\rho}_\theta}{\partial \theta_i} = \frac{1}{2} (\hat{L}_i \hat{\rho}_\theta + \hat{\rho}_\theta \hat{L}_i). \quad (12)$$

The equality in the first matrix inequality  $\Sigma \geq \frac{1}{M} \mathbf{F}_\theta^{-1}$  is attained by a suitable choice of estimator. A paradigmatic example is the maximum-likelihood estimator (MLE), defined as

$$\hat{\theta}^{\text{ML}}(x) := \arg \max_{\theta \in \Theta} \ell_\theta(x), \quad (13)$$

where  $\ell_\theta(x)$  is the log-likelihood function associated with  $M$  independent measurement outcomes, defined as

$$\ell_\theta(x) = \log \prod_{i=1}^M p_\theta(x_i) = \sum_{i=1}^M \log p_\theta(x_i). \quad (14)$$

By construction, the MLE selects the parameter value  $\theta$  that maximizes the log-likelihood of the observed measurement outcomes  $x$ .

In the asymptotic limit  $M \rightarrow \infty$ , the MLE becomes an asymptotically unbiased estimator satisfying Eq. (8) and saturates the first matrix inequality in Eq. (9). As a consequence, the MLE asymptotically achieves

$$\delta^2 \theta_i = \frac{1}{M} [\mathbf{F}_\theta^{-1}(\{\hat{\Pi}_x\}_x)]_{ii}, \quad \forall i \in [d]. \quad (15)$$

In contrast, the equality in the second inequality  $\mathbf{F}_\theta^{-1} \geq \mathbf{J}_\theta^{-1}$  cannot be attained in a general multi-parameter setting [44–47]. The intuitive understanding is that the diagonal elements of the QFIM are obtained by optimizing the measurement independently for each parameter:

$$[\mathbf{J}_\theta]_{ii} = \max_{\{\hat{\Pi}_x\}_x} [\mathbf{F}_\theta(\{\hat{\Pi}_x\}_x)]_{ii}. \quad (16)$$

However, the measurements that are optimal for different parameters are typically incompatible and cannot be implemented simultaneously. As a consequence, there generally does not exist a single POVM that simultaneously achieves

$$\delta^2 \theta_i = \frac{1}{M} [\mathbf{J}_\theta^{-1}]_{ii}, \quad \forall i \in [d]. \quad (17)$$

Lastly, we introduce the QFIM for estimating Pauli expectation values, which is a key ingredient of our applications in Secs. III C and III D. Consider an  $n$ -qubit quantum state expressed in the Pauli basis as

$$\hat{\rho}_\theta := \frac{1}{2^n} \left( \hat{I} + \sum_{i=1}^{4^n-1} \theta_i \hat{P}_i \right), \quad (18)$$

where  $\{\hat{P}_i\}_{i=1}^{4^n-1}$  denotes the set of non-identity  $n$ -qubit Pauli operators and  $\theta_i = \text{Tr}[\hat{\rho}_\theta \hat{P}_i]$  is the corresponding Pauli expectation value. The optimal measurement for minimizing the mean squared error of  $\theta_i$  is the projective measurement in the eigenbasis of  $\hat{P}_i$ . The corresponding diagonal element of the inverse QFIM is given by [48]

$$[\mathbf{J}_\theta^{-1}]_{ii} = 1 - \theta_i^2. \quad (19)$$

### C. Quantum learning theory with $\ell_k$ -distance

We first clarify the notion of *sample complexity*. Let  $N_\mathcal{E}$  denote the number of uses of the elementary channel  $\mathcal{N}_\theta$  required to implement a single instance of the effective parameter-dependent channel  $\mathcal{E}_\theta$ . If the same estimation protocol is repeated independently  $M$  times, the total number of uses of  $\mathcal{N}_\theta$  required to implement the estimation is

$$N_{\text{samp}} := MN_\mathcal{E}. \quad (20)$$

Throughout this paper, we use the following terminology:

$$\begin{aligned} N_{\text{samp}} &: \text{sample complexity} \\ M &: \text{number of repetitions} \\ N_\mathcal{E} &: \text{number of elementary channel uses.} \end{aligned} \quad (21)$$

We note that, once the estimation protocol is designed (i.e.,  $N_\mathcal{E}$  is fixed), determining the sample complexity

$N_{\text{samp}}$  is equivalent to determining the number of repetitions  $M$ .

We introduce quantum learning theory formulated with respect to the  $\ell_k$ -distance. We consider the quantum parameter estimation setting introduced in Sec. II A. Although both quantum metrology and quantum learning theory address parameter estimation, quantum learning theory evaluates performance in terms of the sample complexity required to estimate  $\theta$  within  $\ell_k$ -distance  $\epsilon$  with confidence at least  $1 - \delta$ . To formalize this objective, we adopt the  $(\epsilon, \delta)$ -criterion with respect to the  $\ell_k$ -distance. Within this framework, the performance of an estimation protocol is quantified by the sample complexity required to guarantee that

$$\Pr[\|\tilde{\theta} - \theta\|_k \leq \epsilon] \geq 1 - \delta, \quad (22)$$

uniformly over  $\theta \in \Theta$ . Here,  $\|\mathbf{v}\|_k$  denotes the  $\ell_k$ -norm, defined for a  $d$ -dimensional vector  $\mathbf{v} := (v_1, v_2, \dots, v_d)^T$  as

$$\|\mathbf{v}\|_k := \left( \sum_{i=1}^d |v_i|^k \right)^{1/k}. \quad (23)$$

In the remainder of this paper, we refer to Eq. (22) as the  $(\epsilon, \ell_k, \delta)$ -criterion and focus on  $k = \infty$  and  $k = 2$ . The  $\ell_\infty$ -norm is defined as

$$\|\mathbf{v}\|_\infty := \max_i |v_i|, \quad (24)$$

while the  $\ell_2$ -norm corresponds to the Euclidean norm,

$$\|\mathbf{v}\|_2 := \left( \sum_{i=1}^d |v_i|^2 \right)^{\frac{1}{2}}. \quad (25)$$

Finally, we again emphasize that once the estimation protocol has been specified, the number of elementary channel uses  $N_\mathcal{E}$  is fixed. Consequently, the number of repetitions  $M$  determines the sample complexity  $N_{\text{samp}}$  required to satisfy the  $(\epsilon, \ell_k, \delta)$ -criterion. One may alternatively consider a setting in which  $M$  is fixed where  $N_\mathcal{E}$  determines  $N_{\text{samp}}$ , for example, the extreme one-shot collective-measurement setting where  $M = 1$  and  $N_\mathcal{E} = N_{\text{samp}}$ . In general, however, changing  $N_\mathcal{E}$  amounts to changing the estimation protocol itself. Allowing  $N_\mathcal{E}$  to be chosen separately for each prescribed pair  $(\epsilon, \delta)$  would therefore effectively require designing a different estimation protocol for each target accuracy and confidence level, since different  $(\epsilon, \ell_k, \delta)$ -requirements generally require different sample complexities. To avoid this inefficiency, we adopt a fixed-protocol formulation: once  $\mathcal{E}_\theta$  is chosen,  $N_\mathcal{E}$  is fixed, and the sample-complexity question is to determine the number of independent repetitions  $M$  required to satisfy the desired  $(\epsilon, \ell_k, \delta)$ -criterion.

### III. MAIN RESULT 1: SAMPLE COMPLEXITY OF $\ell_\infty$ -DISTANCE-BASED LEARNING

In this section, we show that the sample complexity required to satisfy the  $(\epsilon, \ell_\infty, \delta)$ -criterion is characterized by the diagonal elements of the inverse FIM. More specifically, we show that both the upper and lower bounds on the number of repetitions of the protocol  $M$  are governed by the supremum over the parameter space  $\Theta$  of the maximum diagonal entry of the inverse FIM. The corresponding bounds on the sample complexity  $N_{\text{samp}}$  are then obtained by simply multiplying these bounds on the number of repetitions by  $N_\mathcal{E}$ , the number of elementary channel uses required to implement one instance of  $\mathcal{E}_\theta$ .

To derive the bounds, we impose the following assumptions on the log-likelihood function  $\ell_\theta(\mathbf{x})$ :

- (A1) Unique maximizer and stationary point:  $\ell_\theta(\mathbf{x})$  has a unique maximizer  $\tilde{\theta}^{\text{ML}}$  in the interior of the parameter domain  $\Theta$ , which is also the unique stationary point.
- (A2) Smoothness:  $\ell_\theta(\mathbf{x})$  is three times continuously differentiable with respect to  $\theta$  on the parameter domain  $\Theta$ .

These assumptions are satisfied by a broad class of regular statistical models under standard parametrizations. Examples include Bernoulli and multinomial models, Poisson models, and Gaussian models with known variance (or covariance), as well as, more generally, exponential family models in their canonical parametrization, as discussed in Appendix H.

For the quantitative bounds derived below, we additionally impose the technical regularity conditions (R1)–(R5), which are stated and discussed in Appendix 3 B 4, to ensure the mathematical rigor of the analysis. Since these assumptions are technical in nature and their detailed presentation would interrupt the flow of the main exposition, their precise statements, together with the corresponding admissible range of finite  $\epsilon$ , are deferred to Appendix 3 B 4. Under these conditions, we now present the main upper and lower bounds.

#### A. Upper bound on $\ell_\infty$ -distance learning

Let us assume that the assumptions (A1) and (A2) and the standard regularity conditions (R1)–(R5) in Appendix 3 B 4 are satisfied. We then obtain the following upper bound on the sample complexity that guarantees the  $(\epsilon, \ell_\infty, \delta)$ -criterion.

*Theorem 1.* (Simplified small-error upper bound). For fixed  $0 < \delta \leq 1$  and  $d < \infty$ , the minimal number of repetitions  $M = M(\epsilon)$  required to guarantee that the MLE satisfies

$$\Pr[\|\tilde{\theta}^{\text{ML}} - \theta\|_\infty \leq \epsilon] \geq 1 - \delta \quad \text{for all } \theta \in \Theta \quad (26)$$

is upper bounded, in the small-error limit  $\epsilon \rightarrow 0$ , as

$$M \lesssim W_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa} \epsilon^{-2}. \quad (27)$$

Equivalently, the minimal sample complexity is upper bounded as

$$N_{\text{samp}} \lesssim N_\mathcal{E} W_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa} \epsilon^{-2}. \quad (28)$$

In Eqs. (27) and (28), the notation  $A_\epsilon \lesssim B_\epsilon$  denotes

$$\lim_{\epsilon \rightarrow 0} \sup \frac{A_\epsilon}{B_\epsilon} \leq 1, \quad (29)$$

i.e.,  $A_\epsilon \leq (1 + o(1))B_\epsilon$  as  $\epsilon \rightarrow 0$  with fixed  $\delta$  and  $d$ .

The proof of Theorem 1 is provided in Appendix C, where we also derive a fully explicit non-asymptotic finite- $\epsilon$  bound.

To clarify the scaling in Eq. (27), we briefly introduce basic properties of the Lambert  $W_0$  function (see Appendix B and Refs. [49,50]). First, it satisfies the elementary bound

$$W_0(x) \leq \log x \quad (30)$$

for  $x \geq e$ . We note that  $W_0(e) = 1$  and  $W_0(x)$  is an increasing function for  $x \geq 0$ . In addition,  $W_0(x)$  admits the asymptotic expansion for large  $x$ ,

$$W_0(x) = \log x - \log \log x + \frac{\log \log x}{\log x} + O\left(\frac{(\log \log x)^2}{(\log x)^2}\right). \quad (31)$$

Applying the bound (30) to our setting, we obtain that, whenever  $8\pi^{-1}\delta^{-2}d^2 \geq e$ ,

$$W_0(8\pi^{-1}\delta^{-2}d^2) \leq \log(8\pi^{-1}\delta^{-2}d^2). \quad (32)$$

Especially when  $8\pi^{-1}\delta^{-2}d^2$  is sufficiently large, the difference between  $W_0(8\pi^{-1}\delta^{-2}d^2)$  and  $\log(8\pi^{-1}\delta^{-2}d^2)$  is approximately  $\log \log(8\pi^{-1}\delta^{-2}d^2)$ .

Finally, we emphasize that the supremum and the maximum are essential in the upper bounds. An upper bound must guarantee the  $(\epsilon, \ell_\infty, \delta)$ -criterion uniformly over the entire parameter space. Since  $\ell_\infty$ -accurate learning requires simultaneous control of every coordinate at every parameter value, as specified in Eq. (24), the number of repetitions  $M$  must be sufficiently large to accommodate the most statistically challenging coordinate at the most unfavorable parameter point. This worst-case uniform requirement is precisely captured by the supremum of the largest diagonal component of the inverse FIM.

### B. Lower bound on $\mathcal{L}_\infty$ -distance learning

We now establish the lower bound under the assumptions (A1) and (A2) and the standard regularity conditions (R1)–(R5) in Appendix 3 B 4 are satisfied.

*Theorem 2.* (Simplified small-error lower bound). For fixed  $0 < \delta < 1/\sqrt{8\pi e}$  and  $d < \infty$ , the minimal number of repetitions  $M = M(\epsilon)$  required to guarantee that the MLE satisfies

$$\Pr[||\tilde{\theta}^{\text{ML}} - \theta||_\infty \leq \epsilon] \geq 1 - \delta \quad \text{for all } \theta \in \Theta \quad (33)$$

is lower bounded, in the small-error limit  $\epsilon \rightarrow 0$ , as

$$M \gtrsim W_0(\delta^{-2}/8\pi) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa} \epsilon^{-2}. \quad (34)$$

Equivalently, the minimal sample complexity is lower bounded as

$$N_{\text{samp}} \gtrsim N_{\mathcal{E}} W_0(\delta^{-2}/8\pi) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa} \epsilon^{-2}. \quad (35)$$

In Eqs. (34) and (35), the notation  $A_\epsilon \gtrsim B_\epsilon$  denotes

$$\liminf_{\epsilon \rightarrow 0} \frac{A_\epsilon}{B_\epsilon} \geq 1, \quad (36)$$

i.e.,  $A_\epsilon \geq (1 + o(1))B_\epsilon$  as  $\epsilon \rightarrow 0$  with fixed  $\delta$  and  $d$ . The proof of Theorem 2 is provided in Appendix D, where we also derive a fully explicit non-asymptotic finite- $\epsilon$  bound.

Combining Theorems 1 and 2, we obtain the following consequence, which shows that the upper and lower bounds nearly match for guarantees that hold uniformly over the entire parameter space.

*Remark 1.* For fixed  $0 < \delta < 1/\sqrt{8\pi e}$  and  $d < \infty$ , the minimal sample complexity  $N_{\text{samp}}$  required to guarantee that the MLE satisfies

$$\Pr[||\tilde{\theta}^{\text{ML}} - \theta||_\infty \leq \epsilon] \geq 1 - \delta \quad \text{for all } \theta \in \Theta \quad (37)$$

is tightly bounded up to  $\log d$ , in the small-error limit  $\epsilon \rightarrow 0$ , as

$$N_{\mathcal{E}} W_0(\delta^{-2}/8\pi) \mathfrak{F} \epsilon^{-2} \lesssim N_{\text{samp}} \lesssim N_{\mathcal{E}} W_0(8\pi^{-1} \delta^{-2} d^2) \mathfrak{F} \epsilon^{-2}, \quad (38)$$

where  $\mathfrak{F} := \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa}$ .

Hence, for a uniform guarantee over the entire parameter space, the upper and lower bounds match in the inverse-FIM-dependent quantity. The only remaining discrepancy between the upper and lower bounds lies in the Lambert  $W_0$  factors. Since

$$8\pi^{-1} \delta^{-2} d^2 = 64d^2(\delta^{-2}/8\pi), \quad (39)$$

the concavity of the Lambert  $W_0$  function implies [49,50],

$$W_0(8\pi^{-1} \delta^{-2} d^2) \leq W_0(\delta^{-2}/8\pi) + 2 \log d + \log 64. \quad (40)$$

Therefore, relative to the lower-bound Lambert factor, the remaining overhead in the upper bound is at most  $\log d$ . Hence, the two bounds match in the inverse FIM dependent quantity and in the leading  $\epsilon^{-2}$  scaling, up to a logarithmic factor in the number of the parameters to be estimated. This demonstrates that the inverse FIM is the fundamental quantity governing the sample-complexity scale of MLE-based  $(\epsilon, \mathcal{L}_\infty, \delta)$ -learning in the asymptotic limit  $\epsilon \rightarrow 0$ .

### C. Application to Pauli channel learning

In this section, we apply Theorems 1 and 2 to the problem of learning the Pauli eigenvalues of an  $n$ -qubit Pauli channel. Our analysis shows that the use of entanglement enables at least an exponential reduction in the sample complexity as a function of the number of qubits for learning the Pauli eigenvalues. This task was originally investigated in Refs. [19,25] using a proof technique different from the FIM-based approach used here.

#### 1. Pauli error rates and eigenvalues

We first introduce notation for  $n$ -qubit Pauli operators. Any  $n$ -qubit Pauli operator can be expressed as

$$\hat{P}_a = \bigotimes_{k=1}^n i^{a_{x,k} a_{z,k}} \hat{X}^{a_{x,k}} \hat{Z}^{a_{z,k}}, \quad (41)$$

where the index  $0 \leq a \leq 4^n - 1$  uniquely labels the Pauli operator, and  $[a]_2 := (a_{x,1}, a_{x,2}, \dots, a_{x,n}, a_{z,1}, a_{z,2}, \dots, a_{z,n})$  denotes the binary representation of  $a$ . Based on this notation, an  $n$ -qubit Pauli channel can be expressed as

$$\Lambda(\hat{\rho}) = \sum_{a=0}^{4^n-1} p_a \hat{P}_a \hat{\rho} \hat{P}_a, \quad (42)$$

where  $\hat{\rho}$  is an arbitrary  $n$ -qubit state, and  $\{p_a\}_{a=0}^{4^n-1}$  are the *Pauli error rates*, satisfying  $\sum_{a=0}^{4^n-1} p_a = 1$ . An equivalent description of a Pauli channel is obtained by considering its action on the Pauli operators. In this representation, the channel is diagonal, satisfying

$$\Lambda(\hat{P}_a) = \lambda_a \hat{P}_a, \quad (43)$$

where  $\{\lambda_a\}_{a=0}^{4^n-1}$  are referred to as the *Pauli eigenvalues*, with  $-1 \leq \lambda_a \leq 1$  for all  $a$  and  $\lambda_0 = 1$ . The Pauli error rates and Pauli eigenvalues are related via the *Walsh-Hadamard transformation*, given by

$$\lambda_b = \sum_{[a]_2 \in \mathbb{Z}^{2n}} (-1)^{\langle [a]_2, [b]_2 \rangle} p_a, \quad (44)$$

where  $\langle [a]_2, [b]_2 \rangle$  denotes the symplectic inner product

$$\langle [a]_2, [b]_2 \rangle := \sum_{k=1}^n a_{x,k} b_{z,k} + a_{z,k} b_{x,k}. \quad (45)$$

In what follows, we consider the task of learning the Pauli eigenvalues  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_{4^n-1})^T$  to additive error  $\epsilon$  in  $\ell_\infty$ -distance. In the Pauli channel learning setting, for simplicity, we omit the dependence on  $\lambda$  in  $\Lambda$  and assume that  $\Lambda$  is the elementary channel  $\mathcal{N}_\lambda$  to which we have access.

## 2. Entanglement-assisted scheme

We first analyze the learning of Pauli eigenvalues assisted by entanglement with a noiseless ancilla mode [see Fig. 2(a) for a schematic illustration]. To estimate the Pauli eigenvalues  $\lambda$ , let us consider the maximally entangled state

$$\hat{\rho}_0 := |\Psi\rangle\langle\Psi| = \frac{1}{4^n} \sum_{a=0}^{4^n-1} \hat{P}_a^S \otimes (\hat{P}_a^A)^T, \quad (46)$$

where  $|\Psi\rangle = 2^{-n/2} \sum_{z \in \{0,1\}^n} |z\rangle_S |z\rangle_A$  denotes the standard maximally entangled state, as a quantum probe. With the Pauli convention in Eq. (41), the transpose of a Pauli operator is given by

$$(\hat{P}_a)^T = \eta_a \hat{P}_a, \quad \eta_a := (-1)^{\sum_{k=1}^n a_{x,k} a_{z,k}},$$

which accounts for the minus sign associated with each  $\hat{Y}$  factor. Here, the superscripts S and A denote the system and the noiseless ancilla, respectively. Next, we inject  $\hat{\rho}_0$  into the  $n$ -qubit Pauli channel acting only on the system.

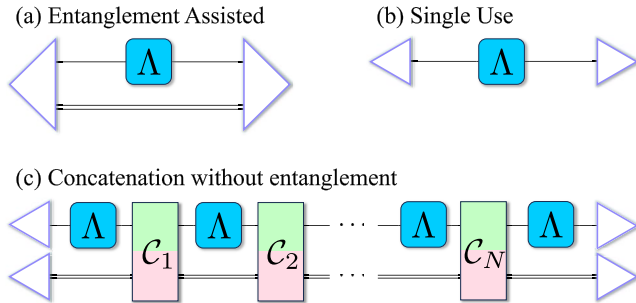


FIG. 2. Schematic of Pauli channel learning. Left and right pointing triangles denote quantum probe and measurement respectively, blue rounded squares represent the Pauli channel  $\Lambda$ , green/pink blocks represent intermediate quantum control operations that do not generate entanglement, and the double dashed line denotes the ancillary system. (a) Entanglement-assisted strategy using a maximally entangled probe state and a Bell measurement on the system and ancilla. (b) Entanglement-free strategy using a single use of the Pauli channel with local state preparation and measurement. (c) Entanglement-free concatenation strategy consisting of consecutive uses of the Pauli channel interleaved with a predetermined sequence of CPTP control maps, with no system-ancilla entanglement generated throughout the protocol.

In this case, the effective channel is defined as  $\mathcal{E}_\lambda := \Lambda^S \otimes \mathcal{I}^A$ , where  $\mathcal{I}^A$  denotes the identity channel. In this estimation protocol,  $N_\mathcal{E} = 1$ ; therefore,  $N_{\text{samp}} = M$ . The encoded state of the Pauli error rates (or Pauli eigenvalues) can be expressed as

$$\hat{\rho}_\lambda = \sum_{a=0}^{4^n-1} p_a (\hat{P}_a^S \otimes \hat{P}_0^A) |\Psi\rangle\langle\Psi| (\hat{P}_a^S \otimes \hat{P}_0^A) \quad (47)$$

$$= \frac{1}{4^n} \sum_{b=0}^{4^n-1} \lambda_b \hat{P}_b^S \otimes (\hat{P}_b^A)^T. \quad (48)$$

Here, we note that  $\hat{\rho}_\lambda$  is a diagonal matrix whose diagonal components are  $\{p_a\}_{a=0}^{4^n-1}$  with respect to the orthonormal basis  $\{(\hat{P}_a^S \otimes \hat{P}_0^A) |\Psi\rangle\}_{a=0}^{4^n-1}$ .

The QFIM with respect to  $\lambda$  is given by

$$[\mathbf{J}_\lambda]_{ij} := \frac{1}{2} \text{Tr}[\hat{\rho}_\lambda \{\hat{L}_i, \hat{L}_j\}]. \quad (49)$$

In particular, the diagonal elements of  $\mathbf{J}_\lambda$  are

$$[\mathbf{J}_\lambda]_{bb} = \text{Tr} \left[ \hat{\rho}_\lambda \frac{\partial \hat{\rho}_\lambda}{\partial \lambda_b} \right]. \quad (50)$$

From Eq. (48), the derivative of the density operator with respect to  $\lambda_b$  is

$$\frac{\partial \hat{\rho}_\lambda}{\partial \lambda_b} = \frac{1}{4^n} \hat{P}_b^S \otimes (\hat{P}_b^A)^T. \quad (51)$$

Importantly, it is straightforward to verify that  $\frac{\partial \hat{\rho}_\lambda}{\partial \lambda_b}$  satisfies

$$\left[ \frac{\partial \hat{\rho}_\lambda}{\partial \lambda_b}, \hat{\rho}_\lambda \right] = 0 \Leftrightarrow \left[ \frac{\partial \hat{\rho}_\lambda}{\partial \lambda_b}, \hat{\rho}_\lambda^{-1} \right] = 0. \quad (52)$$

Since  $\frac{\partial \hat{\rho}_\lambda}{\partial \lambda_b}$  commutes with  $\hat{\rho}_\lambda$ , the SLD admits the simplified form

$$\hat{L}_b = \hat{\rho}_\lambda^{-1} \frac{\partial \hat{\rho}_\lambda}{\partial \lambda_b}. \quad (53)$$

Thus, for all  $b$ , the SLD operators also commute with  $\hat{\rho}_\lambda$  and are simultaneously diagonalizable. Consequently,  $\hat{\rho}_\lambda$  and  $\{\hat{L}_b\}_{b=1}^d$  share the common eigenbasis

$$\{(\hat{P}_i^S \otimes \hat{P}_0^A) |\Psi\rangle\}_{i=0}^{4^n-1}. \quad (54)$$

The Bell measurement is the projective measurement onto the basis

$$|\Psi_x\rangle := (\hat{P}_x^S \otimes \hat{P}_0^A) |\Psi\rangle, \quad \hat{\Pi}_x := |\Psi_x\rangle\langle\Psi_x|. \quad (55)$$

Using Eq. (46), the corresponding projectors can be expressed as

$$\hat{\Pi}_x = \frac{1}{4^n} \sum_{a=0}^{4^n-1} (-1)^{\langle [x]_2, [a]_2 \rangle} \hat{P}_a^S \otimes (\hat{P}_a^A)^T. \quad (56)$$

Hence, a single projective measurement onto this common eigenbasis saturates the quantum Cramér-Rao matrix inequalities in Eq. (9). Operationally, this corresponds to performing a Bell measurement. Therefore, the FIM with the Bell measurement equals the QFIM,

$$\mathbf{F}_\lambda(\{\hat{\Pi}_x\}_x) = \mathbf{J}_\lambda. \quad (57)$$

As a result, from Eq. (19), we have

$$[\mathbf{F}_\lambda^{-1}]_{aa} = [\mathbf{J}_\lambda^{-1}]_{aa} = 1 - \lambda_a^2. \quad (58)$$

It then follows immediately that

$$\sup_{\lambda} \max_{a \in [d]} [\mathbf{F}_\lambda^{-1}]_{aa} = 1. \quad (59)$$

Finally, invoking Theorem 1, we conclude that the sample complexity  $N_{\text{samp}}$  required to satisfy the  $(\epsilon, \ell_\infty, \delta)$ -criterion for Pauli eigenvalue learning using entanglement is upper bounded by

$$N_{\text{samp}} \lesssim W_0(8\pi^{-1}\delta^{-2}4^{2n})e^{-2}. \quad (60)$$

For clarity, we reiterate that

$$W_0(8\pi^{-1}\delta^{-2}4^{2n}) \leq \log \frac{8}{\pi} + 2n \log 4 + 2 \log \delta^{-1}. \quad (61)$$

Hence, in the entangled setting, the sample complexity increases by at most a polynomial factor in the number of qubits  $n$ . This upper bound on the sample complexity is consistent with the upper bound obtained in Ref. [19], which scales as  $O(n \log \delta^{-1} e^{-2})$ .

Lastly, let us inspect the MLE of the Pauli eigenvalues. The probability of obtaining outcome  $x$  from the Bell measurement is given by

$$p_\lambda(x) := \text{Tr}[\hat{\Pi}_x \hat{\rho}_\lambda] = \frac{1}{4^n} \sum_{a=0}^{4^n-1} \lambda_a (-1)^{\langle [x]_2, [a]_2 \rangle}. \quad (62)$$

We perform this measurement independently  $M$  times, obtaining a sequence of outcomes  $\mathbf{x} = (x_1, x_2, \dots, x_M)$ . Let  $n_x$  denote the number of occurrences of outcome  $x$ , with  $\sum_{x=0}^{4^n-1} n_x = M$ . As shown in Appendix H, the log-likelihood associated with the measurement statistics satisfies the standard regularity conditions (A1) and (A2). Under these conditions, the MLE of the Pauli eigenvalues is given by

$$\tilde{\lambda}_a^{\text{ML}} = \sum_{b=0}^{4^n-1} \frac{n_b}{M} (-1)^{\langle [a]_2, [b]_2 \rangle}. \quad (63)$$

### 3. Separable scheme with a single use of Pauli channel without an ancilla

We next investigate learning the Pauli eigenvalues using a single use of the Pauli channel, without entanglement [see Fig. 2(b) for a schematic illustration]. Any  $n$ -qubit quantum state can be expanded in the Pauli basis as

$$\hat{\rho}_0 = \frac{1}{2^n} \sum_{a=0}^{4^n-1} r_a \hat{P}_a, \quad (64)$$

where the coefficients  $r_a := \text{Tr}[\hat{\rho}_0 \hat{P}_a]$  are real-valued and satisfy  $r_0 = 1$  due to the trace condition. We first focus on the simplest setting of a single use of the Pauli channel. In this estimation protocol,  $\Lambda = \mathcal{N}_\lambda = \mathcal{E}_\lambda$  and  $N_\mathcal{E} = 1$ , which implies that  $N_{\text{samp}} = M$ . Let us consider an  $n$ -qubit Pauli channel acting on the quantum probe  $\hat{\rho}_0$ . The resulting state is given by

$$\Lambda(\hat{\rho}_0) = \frac{1}{2^n} \sum_{a=0}^{4^n-1} \lambda_a r_a \hat{P}_a. \quad (65)$$

From Eq. (19) and by applying the chain rule, the diagonal elements of the inverse QFIM are given by

$$[\mathbf{J}_\lambda^{-1}]_{aa} = \frac{1}{r_a^2} (1 - r_a^2 \lambda_a^2). \quad (66)$$

Here,  $[\mathbf{J}_\lambda^{-1}]_{aa}$  satisfies the inequality

$$[\mathbf{J}_\lambda^{-1}]_{aa} = \frac{1}{r_a^2} (1 - r_a^2 \lambda_a^2) \geq \frac{1}{r_a^2} - 1. \quad (67)$$

To proceed, we note that every quantum state must satisfy the purity constraint  $\text{Tr}[\hat{\rho}_0^2] \leq 1$ , which is equivalently expressed as

$$\sum_{a \neq 0} r_a^2 \leq 2^n - 1. \quad (68)$$

Equation (68) directly implies that there always exists an index  $a$  such that  $r_a^2 \leq 2^{-n}$ . As a consequence, there exists at least one  $a$  for which the corresponding diagonal element satisfies

$$[\mathbf{J}_\lambda^{-1}]_{aa} \geq 2^n. \quad (69)$$

According to the quantum Cramér-Rao matrix inequality in Eq. (9), for any POVM  $\{\hat{\Pi}_x\}_x$ , the following inequality holds:

$$\sup_{\lambda} \max_{a \in [d]} [\mathbf{F}_\lambda^{-1}(\{\hat{\Pi}_x\}_x)]_{aa} \geq [\mathbf{F}_\lambda^{-1}(\{\hat{\Pi}_x\}_x)]_{aa} \geq [\mathbf{J}_\lambda^{-1}]_{aa}. \quad (70)$$

As a consequence, Theorem 2 implies that the sample complexity  $N_{\text{samp}}$  required to satisfy the  $(\epsilon, \ell_\infty, \delta)$ -criterion for Pauli eigenvalue learning using a separable scheme, particularly in the single-use setting of the Pauli channel, is lower bounded by

$$N_{\text{samp}} \gtrsim W_0(\delta^{-2}/8\pi)\epsilon^{-2}2^n. \quad (71)$$

For clarity, we reiterate that using Eq. (30), we have

$$1 < W_0(\delta^{-2}/8\pi) \leq 2 \log \delta^{-1} - \log 8\pi, \quad (72)$$

for  $0 < \delta < 1/\sqrt{8\pi e}$ . By comparing Eqs. (60) and (71), we observe that the use of entanglement yields at least an exponential reduction in the required sample complexity with respect to the number of qubits  $n$ . The intuitive origin of the exponential sample complexity is as follows. In the absence of entanglement, the probe state necessarily possesses at least one Pauli-basis component whose magnitude is exponentially small in the number of qubits, as indicated by Eq. (68). This severely limits the amount of information that can be encoded about the corresponding Pauli eigenvalue, resulting in an exponential factor  $2^n$  in the required sample complexity. Lastly, we note that the original analysis of Pauli channel learning in Ref. [19] establishes a lower bound of  $\Omega(n2^n)$  for the fixed choices  $\epsilon = 1/2$  and  $\delta = 1/3$  whereas our lower bound depends explicitly on  $\delta$  and applies in the range  $0 < \delta < 1/\sqrt{8\pi e}$  in the asymptotic small-error regime  $\epsilon \rightarrow 0$ . This leaves a multiplicative gap of order  $n$  compared with our lower bound in Eq. (71). Nevertheless, Eq. (71) is already sufficient to demonstrate that entanglement enables an exponential advantage over protocols restricted to single, non-entangled uses of the Pauli channel.

#### 4. Separable scheme with multiple uses of Pauli channel with an unbounded ancilla

We now show that Pauli-eigenvalue learning without entanglement requires a sample complexity that grows exponentially with the number of qubits  $n$ . We allow an arbitrarily large ancillary system, repeated uses of the Pauli channel, and general parameter-independent processing operations between consecutive channel uses. The only restriction is that the joint system-ancilla state remains separable throughout the protocol; see Fig. 2(c).

Since several parameter-dependent quantum states appear in the following analysis, we make the state dependence of the QFIM explicit. For an arbitrary parameter-dependent state  $\hat{\rho}(\lambda)$ , we denote its QFIM by  $\mathbf{J}(\hat{\rho}(\lambda))$ , whose matrix elements are defined as

$$[\mathbf{J}(\hat{\rho}(\lambda))]_{ab} := \frac{1}{2} \text{Tr}[\hat{\rho}(\lambda)(\hat{L}_a \hat{L}_b + \hat{L}_b \hat{L}_a)], \quad (73)$$

where  $\hat{L}_a$  is the SLD associated with the state appearing in the argument of  $\mathbf{J}$ , defined through

$$\frac{\partial \hat{\rho}(\lambda)}{\partial \lambda_a} = \frac{1}{2}(\hat{L}_a \hat{\rho}(\lambda) + \hat{\rho}(\lambda) \hat{L}_a). \quad (74)$$

In particular, for the final encoded state  $\hat{\rho}_\lambda$ , this notation is consistent with the notation used in the preceding sections:

$$\mathbf{J}_\lambda := \mathbf{J}(\hat{\rho}_\lambda). \quad (75)$$

We now consider a general entanglement-free learning scheme. Since the protocol is entanglement-free, the initial probe state  $\hat{\rho}_0$  is separable across the system and ancillary systems and can therefore be expressed as

$$\hat{\rho}_0 = \sum_{j_0} p_{j_0} \hat{\rho}_{0,j_0}^S \otimes \hat{\rho}_{0,j_0}^A, \quad (76)$$

where  $\{p_{j_0}\}_{j_0}$  is a probability distribution. Between the  $T$ th and  $(T+1)$ th uses of the Pauli channel, with  $1 \leq T \leq N$ , we consider a general  $\lambda$ -independent separable channel

$$\mathcal{C}_T = \sum_{j_T} \mathcal{A}_{T,j_T}^S \otimes \mathcal{B}_{T,j_T}^A, \quad (77)$$

where  $\mathcal{A}_{T,j_T}^S$  and  $\mathcal{B}_{T,j_T}^A$  are completely positive and trace-non-increasing maps acting on the system and ancillary systems, respectively, while  $\mathcal{C}_T$  is trace preserving. Thus, the label  $j_T$  specifies a branch of the separable operation  $\mathcal{C}_T$ . We emphasize that this formulation encompasses general separable schemes, including history-dependent protocols with mid-circuit measurements and adaptive control, provided that no system-ancilla entanglement is generated [25].

For

$$\mathbf{j}_T := (j_0, j_1, \dots, j_T), \quad (78)$$

the multi-index  $\mathbf{j}_T$  records the complete branch history up to the  $T$ th intermediate operation. We set

$$p_{j_0}(\lambda) := p_{j_0}, \quad \hat{\rho}_{j_0}^S(\lambda) := \hat{\rho}_{0,j_0}^S, \quad \hat{\rho}_{j_0}^A := \hat{\rho}_{0,j_0}^A. \quad (79)$$

After  $T$  uses of the Pauli channel and the subsequent operations  $\mathcal{C}_1, \dots, \mathcal{C}_T$ , the joint state is

$$\begin{aligned} \hat{\rho}_T(\lambda) &:= \mathcal{C}_T \circ (\Lambda^S \otimes \mathcal{I}^A) \circ \dots \circ \mathcal{C}_1 \circ (\Lambda^S \otimes \mathcal{I}^A)(\hat{\rho}_0) \\ &= \sum_{j_T} p_{j_T}(\lambda) \hat{\rho}_{j_T}^S(\lambda) \otimes \hat{\rho}_{j_T}^A. \end{aligned} \quad (80)$$

Equivalently, this state obeys the recursion

$$\begin{aligned} \hat{\rho}_T(\lambda) &= \mathcal{C}_T \circ (\Lambda^S \otimes \mathcal{I}^A)(\hat{\rho}_{T-1}(\lambda)) \\ &= \sum_{j_T} \sum_{j_{T-1}} p_{j_{T-1}}(\lambda) \mathcal{A}_{T,j_T}^S(\Lambda^S(\hat{\rho}_{j_{T-1}}^S(\lambda))) \otimes \mathcal{B}_{T,j_T}^A(\hat{\rho}_{j_{T-1}}^A). \end{aligned} \quad (81)$$

For every branch of nonzero probability, the normalized conditional states are therefore

$$\hat{\rho}_{j_T}^S(\lambda) := \frac{\mathcal{A}_{T,j_T}^S(\Lambda^S(\hat{\rho}_{j_{T-1}}^S(\lambda)))}{\text{Tr}[\mathcal{A}_{T,j_T}^S(\Lambda^S(\hat{\rho}_{j_{T-1}}^S(\lambda)))]}, \quad (82)$$

$$\hat{\rho}_{j_T}^A := \frac{\mathcal{B}_{T,j_T}^A(\hat{\rho}_{j_{T-1}}^A)}{\text{Tr}[\mathcal{B}_{T,j_T}^A(\hat{\rho}_{j_{T-1}}^A)]}. \quad (83)$$

Branches with zero probability can be omitted. Since the initial ancillary states and all maps  $\mathcal{B}_{T,j_T}^A$  are independent of  $\lambda$ , the conditional ancillary state  $\hat{\rho}_{j_T}^A$  is also independent of  $\lambda$  for every branch.

The conditional probability of obtaining  $j_T$  given the preceding branch  $j_{T-1}$  is

$$c_{T,j_T|j_{T-1}}(\lambda) := \text{Tr}[\mathcal{A}_{T,j_T}^S(\Lambda^S(\hat{\rho}_{j_{T-1}}^S(\lambda)))] \times \text{Tr}[\mathcal{B}_{T,j_T}^A(\hat{\rho}_{j_{T-1}}^A)]. \quad (84)$$

Accordingly,

$$p_{j_T}(\lambda) = p_{j_{T-1}}(\lambda) c_{T,j_T|j_{T-1}}(\lambda). \quad (85)$$

Because  $\mathcal{C}_T$  is trace preserving, these conditional probabilities satisfy

$$\sum_{j_T} c_{T,j_T|j_{T-1}}(\lambda) = 1. \quad (86)$$

It is useful to express each conditional probability as the outcome probability of a POVM acting only on the system. Define

$$\hat{\Pi}_{T,j_T|j_{T-1}}^S := \text{Tr}[\mathcal{B}_{T,j_T}^A(\hat{\rho}_{j_{T-1}}^A)] \mathcal{A}_{T,j_T}^{S\dagger}(\hat{I}^S). \quad (87)$$

Complete positivity of  $\mathcal{A}_{T,j_T}^S$  implies  $\hat{\Pi}_{T,j_T|j_{T-1}}^S \geq 0$ . Moreover, trace preservation of  $\mathcal{C}_T$  gives

$$\sum_{j_T} \mathcal{A}_{T,j_T}^{S\dagger}(\hat{I}^S) \otimes \mathcal{B}_{T,j_T}^{A\dagger}(\hat{I}^A) = \hat{I}^S \otimes \hat{I}^A. \quad (88)$$

Taking the expectation value of the ancillary part in  $\hat{\rho}_{j_{T-1}}^A$  yields

$$\sum_{j_T} \hat{\Pi}_{T,j_T|j_{T-1}}^S = \hat{I}^S. \quad (89)$$

Hence  $\{\hat{\Pi}_{T,j_T|j_{T-1}}^S\}_{j_T}$  is a valid POVM, and

$$c_{T,j_T|j_{T-1}}(\lambda) = \text{Tr}[\hat{\Pi}_{T,j_T|j_{T-1}}^S \Lambda^S(\hat{\rho}_{j_{T-1}}^S(\lambda))]. \quad (90)$$

Iterating the probability recursion gives

$$p_{j_N}(\lambda) = p_{j_0} \prod_{T=1}^N c_{T,j_T|j_{T-1}}(\lambda). \quad (91)$$

After the final,  $(N+1)$ th, use of the Pauli channel, the encoded state is therefore

$$\hat{\rho}_\lambda = \sum_{j_N} p_{j_N}(\lambda) \Lambda^S(\hat{\rho}_{j_N}^S(\lambda)) \otimes \hat{\rho}_{j_N}^A. \quad (92)$$

We next bound the diagonal elements of the QFIM of  $\hat{\rho}_\lambda$ . By the extended convexity of the QFIM  $[\mathbf{J}_\lambda]_{aa} = [\mathbf{J}(\hat{\rho}_\lambda)]_{aa}$  [43],

$$[\mathbf{J}_\lambda]_{aa} \leq F_{aa}^{\text{cl}}(\lambda) + \sum_{j_N} p_{j_N}(\lambda) [\mathbf{J}(\Lambda^S(\hat{\rho}_{j_N}^S(\lambda)))]_{aa}, \quad (93)$$

where we used the fact that  $\hat{\rho}_{j_N}^A$  is independent of  $\lambda$ . The classical contribution is

$$F_{aa}^{\text{cl}}(\lambda) := \sum_{j_N} \frac{(\partial_{\lambda_a} p_{j_N}(\lambda))^2}{p_{j_N}(\lambda)}. \quad (94)$$

Using the product form of  $p_{j_N}(\lambda)$  in Eq. (91), we have

$$\partial_{\lambda_a} \log p_{j_N}(\lambda) = \sum_{T=1}^N \partial_{\lambda_a} \log c_{T,j_T|j_{T-1}}(\lambda). \quad (95)$$

For every fixed  $j_{T-1}$ ,

$$\begin{aligned} \sum_{j_T} c_{T,j_T|j_{T-1}}(\lambda) \partial_{\lambda_a} \log c_{T,j_T|j_{T-1}}(\lambda) \\ = \partial_{\lambda_a} \sum_{j_T} c_{T,j_T|j_{T-1}}(\lambda) = 0. \end{aligned} \quad (96)$$

Therefore, all cross terms between different values of  $T$  vanish after averaging over the branch distribution, and the classical FIM decomposes exactly as

$$F_{aa}^{\text{cl}}(\lambda) = \sum_{T=1}^N \sum_{j_{T-1}} p_{j_{T-1}}(\lambda) \sum_{j_T} \frac{(\partial_{\lambda_a} c_{T,j_T|j_{T-1}}(\lambda))^2}{c_{T,j_T|j_{T-1}}(\lambda)}. \quad (97)$$

For each fixed  $T$  and  $j_{T-1}$ , the innermost sum is the classical Fisher information obtained by measuring  $\Lambda^S(\hat{\rho}_{j_{T-1}}^S(\lambda))$  with the POVM  $\{\hat{\Pi}_{T,j_T|j_{T-1}}^S\}_{j_T}$ . Hence, by Eq. (16),

$$F_{aa}^{\text{cl}}(\lambda) \leq \sum_{T=1}^N \sum_{j_{T-1}} p_{j_{T-1}}(\lambda) [\mathbf{J}(\Lambda^S(\hat{\rho}_{j_{T-1}}^S(\lambda)))]_{aa}. \quad (98)$$

We now evaluate these bounds at the completely depolarizing Pauli channel,

$$\Lambda^S|_{\lambda=\mathbf{0}} := \Lambda_0^S, \quad (99)$$

where  $\mathbf{0} = (0, 0, \dots, 0)$ . At this point, the output of every use of the Pauli channel is maximally mixed:

$$\Lambda_0^S(\hat{\rho}_{j_T}^S(\mathbf{0})) = \frac{\hat{I}^S}{2^n}, \quad 0 \leq T \leq N. \quad (100)$$

For  $1 \leq a \leq 4^n - 1$ , define

$$r_{j_T,a} := \text{Tr}[\hat{\rho}_{j_T}^S(\mathbf{0}) \hat{P}_a^S]. \quad (101)$$

Notice that  $r_{j_T,a}$  is the Pauli coefficient of the conditional system state entering the next use of the channel. Using the Pauli-channel parametrization and  $\text{Tr}[\partial_{\lambda_a} \hat{\rho}_{j_T}^S(\lambda)] = 0$ , we obtain

$$\partial_{\lambda_a} \Lambda^S(\hat{\rho}_{j_T}^S(\lambda))|_{\lambda=\mathbf{0}} = \frac{1}{2^n} r_{j_T,a} \hat{P}_a^S. \quad (102)$$

The possible  $\lambda$ -dependence of the input state produces no additional contribution at  $\lambda = \mathbf{0}$ , because

$$\Lambda_0^S(\partial_{\lambda_a} \hat{\rho}_{j_T}^S(\lambda)|_{\lambda=\mathbf{0}}) = \frac{\hat{I}^S}{2^n} \text{Tr}[\partial_{\lambda_a} \hat{\rho}_{j_T}^S(\lambda)|_{\lambda=\mathbf{0}}] = 0. \quad (103)$$

For the branch state under consideration, the corresponding SLD operator is therefore

$$\hat{L}_a|_{\lambda=\mathbf{0}} = r_{j_T,a} \hat{P}_a^S, \quad (104)$$

and hence

$$[\mathbf{J}(\Lambda^S(\hat{\rho}_{j_T}^S(\lambda)))]_{aa}|_{\lambda=\mathbf{0}} = (r_{j_T,a})^2. \quad (105)$$

Since  $\hat{\rho}_{j_T}^S(\mathbf{0})$  is a normalized  $n$ -qubit state, its Pauli coefficients satisfy the purity constraint

$$\sum_{a=1}^{4^n-1} (r_{j_T,a})^2 \leq 2^n - 1. \quad (106)$$

It follows that

$$\begin{aligned} \sum_{a=1}^{4^n-1} F_{aa}^{\text{cl}}(\mathbf{0}) &\leq \sum_{T=1}^N \sum_{j_{T-1}} p_{j_{T-1}}(\mathbf{0}) \sum_{a=1}^{4^n-1} (r_{j_{T-1},a})^2 \\ &\leq N(2^n - 1). \end{aligned} \quad (107)$$

Similarly, the QFIM contribution from the final channel use satisfies

$$\begin{aligned} &\sum_{a=1}^{4^n-1} \sum_{j_N} p_{j_N}(\mathbf{0}) [\mathbf{J}(\Lambda^S(\hat{\rho}_{j_N}^S(\lambda)))]_{aa}|_{\lambda=\mathbf{0}} \\ &= \sum_{j_N} p_{j_N}(\mathbf{0}) \sum_{a=1}^{4^n-1} (r_{j_N,a})^2 \leq 2^n - 1. \end{aligned} \quad (108)$$

Combining the preceding bounds gives

$$\sum_{a=1}^{4^n-1} [\mathbf{J}_{\lambda=\mathbf{0}}]_{aa} \leq (N+1)(2^n - 1). \quad (109)$$

Since  $4^n - 1 = (2^n - 1)(2^n + 1)$ , there exists at least one index  $a$  such that

$$[\mathbf{J}_{\lambda=\mathbf{0}}]_{aa} \leq \frac{N+1}{2^n + 1}. \quad (110)$$

This implies that for any final measurement on the encoded state  $\hat{\rho}_{\lambda=\mathbf{0}}$ , the corresponding classical FIM satisfies  $[\mathbf{F}_{\mathbf{0}}]_{aa} \leq [\mathbf{J}(\hat{\rho}_{\mathbf{0}})]_{aa}$ . As a consequence, we have

$$\begin{aligned} \sup_{\lambda} \max_{a \in [4^n-1]} [\mathbf{F}_{\lambda}^{-1}]_{aa} &\geq [\mathbf{F}_{\lambda=\mathbf{0}}^{-1}]_{aa} \geq \frac{1}{[\mathbf{F}_{\lambda=\mathbf{0}}]_{aa}} \\ &\geq \frac{1}{[\mathbf{J}_{\lambda=\mathbf{0}}]_{aa}} \geq \frac{2^n + 1}{N+1}. \end{aligned} \quad (111)$$

One execution of the protocol uses  $N_{\mathcal{E}} = N+1$  elementary Pauli channels. Therefore, Theorem 2 yields

$$\begin{aligned} N_{\text{samp}} &\gtrsim N_{\mathcal{E}} W_0 (\delta^{-2}/8\pi) \epsilon^{-2} \sup_{\lambda} \max_{a \in [4^n-1]} [\mathbf{F}_{\lambda}^{-1}]_{aa} \\ &\geq W_0 (\delta^{-2}/8\pi) (2^n + 1) \epsilon^{-2}. \end{aligned} \quad (112)$$

This reproduces the exponential lower bound established in Ref. [25] by a different method. In contrast to that work, which fixes  $\delta = 1/3$ , the present FIM-based bound retains the explicit dependence on  $\delta$  and applies for  $0 < \delta < 1/\sqrt{8\pi e}$  in the asymptotic small-error regime  $\epsilon \rightarrow 0$ .

#### D. Application to Pauli expectation value estimation

Any  $n$ -qubit quantum state admits an expansion in the Pauli operator basis. Specifically, one can write

$$\hat{\rho}_{\mathbf{c}} := \frac{1}{2^n} \left( \hat{I} + \sum_{a=1}^{4^n-1} c_a \hat{P}_a \right), \quad (113)$$

where the real coefficients  $\mathbf{c} := (c_1, c_2, \dots, c_{4^n-1})^T$  (the Pauli expectation values) uniquely specify  $\hat{\rho}_{\mathbf{c}}$ . We now derive the bounds of the sample complexity required to learn the coefficients  $\mathbf{c}$  to accuracy  $\epsilon$  in the  $\ell_{\infty}$ -distance. The sample complexity of estimating the Pauli expectation values  $\mathbf{c}$ , both with and without quantum memory, was originally analyzed in Ref. [17] using a different proof

technique. Here, quantum memory refers to the capability of simultaneously preparing multiple copies of an unknown quantum state and performing collective measurements across the states.

### 1. Quantum memory and collective measurement

We first consider the setting in which multiple copies of the state  $\hat{\rho}_c$  can be prepared simultaneously and arbitrary collective measurements across these copies are allowed. Realizing this scenario requires quantum memory.

The estimation procedure proposed in Ref. [17] proceeds in two stages. In the first stage, the absolute values  $\{|c_a|\}_a$  of the Pauli expectation values are estimated using a Bell measurement on two copies of  $\hat{\rho}_c$ . In the second stage, the signs of the coefficients are determined by performing measurements on additional copies of the state. Importantly, it is shown that estimating the absolute values  $|c_a|$  dominates the overall sample complexity, whereas determining the signs incurs only a subleading overhead [17]. Motivated by this separation of costs, we restrict our attention to the task of estimating the absolute values of Pauli expectation values.

To estimate the absolute values of the Pauli expectation values, one prepares two identical copies of the state  $\hat{\rho}_c$  and performs a Bell measurement on the joint system, obtaining an outcome  $x$ . In the terminology of Sec. II A, this estimation protocol consumes two copies of the unknown state per run of the protocol, i.e.,  $N_{\mathcal{E}} = 2$ , hence  $N_{\text{samp}} = 2M$ . The probability of observing outcome  $x$  is given by

$$p_c(x) := \text{Tr}[\hat{\Pi}_x(\hat{\rho}_c \otimes \hat{\rho}_c)]. \quad (114)$$

Substituting the Pauli expansion of  $\hat{\rho}_c$  and using the explicit form of the Bell measurement projectors,

$$\hat{\Pi}_x = \frac{1}{4^n} \sum_{k=0}^{4^n-1} (-1)^{\langle [x]_2, [k]_2 \rangle} \hat{P}_k \otimes \hat{P}_k^T,$$

this probability can be expressed as

$$p_c(x) = \frac{1}{16^n} \sum_{a,b=0}^{4^n-1} c_a c_b \times \sum_k \text{Tr}[(\hat{P}_x \otimes \hat{I})(\hat{P}_k \otimes \hat{P}_k^T)(\hat{P}_x \otimes \hat{I})(\hat{P}_a \otimes \hat{P}_b)], \quad (115)$$

with  $c_0 = 1$ . Evaluating the trace using the orthogonality and commutation relations of Pauli operators yields

$$p_c(x) = \frac{1}{4^n} \sum_{a=0}^{4^n-1} \eta_a c_a^2 (-1)^{\langle [x]_2, [a]_2 \rangle}, \quad (116)$$

where  $\eta_a := (-1)^{\sum_{k=1}^n a_{x,k} a_{z,k}}$ . Thus, the Bell-measurement distribution has the same Walsh-Hadamard structure as Eq. (62), with the variables  $C_a := \eta_a c_a^2$ . Since the signs  $\eta_a = \pm 1$  are known *a priori*, estimating  $C_a$  is equivalent to estimating  $c_a^2$ . Consequently, the sample-complexity analysis for Pauli eigenvalue estimation can be directly translated to the estimation of the squared Pauli expectation values. In particular, estimating the known-sign variables  $C_a = \eta_a c_a^2$  is statistically equivalent to estimating Pauli eigenvalues, allowing all corresponding sample-complexity results to be transferred immediately to the problem of learning the magnitudes of Pauli expectation values. Moreover, estimating  $|c_a|^2$  within additive error  $\epsilon^2$  implies estimating  $|c_a|$  within additive error  $\epsilon$ .

Finally, by invoking Theorem 1, we conclude that the sample complexity  $N_{\text{samp}}$  required to satisfy the  $(\epsilon, \ell_{\infty}, \delta)$ -criterion for Pauli expectation value learning with quantum memory is upper bounded by

$$N_{\text{samp}} \lesssim 2W_0(8\pi^{-1}\delta^{-2}4^{2n})\epsilon^{-4}, \quad (117)$$

where  $W_0(8\pi^{-1}\delta^{-2}4^{2n}) \leq \log \frac{8}{\pi} + 2n \log 4 + 2 \log \delta^{-1}$ . Here, we note that the factor of 2 arises from  $N_{\mathcal{E}} = 2$ . This exhibits a polynomial (linear) dependence on the number of qubits  $n$ , consistent with the upper bound established in Ref. [17], which scales as  $O((n + \log \delta^{-1})\epsilon^{-4})$ .

### 2. Single copy of the state

Next, we consider the setting in which only a single copy of the state is available per measurement where  $N_{\mathcal{E}} = 1$ , equivalently,  $N_{\text{samp}} = M$ . Let  $\{\hat{\Pi}_x\}_x$  be a general POVM satisfying  $\sum_x \hat{\Pi}_x = \hat{I}$ . We analyze the estimation problem at the maximally mixed state, i.e.,  $c = \mathbf{0}$ . Since each POVM element is positive semidefinite, it admits a spectral decomposition

$$\hat{\Pi}_x = \sum_j \lambda_{xj} |\phi_{xj}\rangle \langle \phi_{xj}|, \quad (118)$$

where  $\lambda_{xj} \geq 0$ . For the maximally mixed state  $\hat{\rho}_{c=0} = \hat{I}/2^n$ , the probability of obtaining outcome  $x$  is

$$p_0(x) = \text{Tr}[\hat{\Pi}_x \hat{\rho}_{c=0}] = \frac{1}{2^n} \text{Tr}[\hat{\Pi}_x] = \frac{1}{2^n} \sum_j \lambda_{xj}. \quad (119)$$

The derivative of this probability with respect to  $c_a$ , evaluated at  $c = \mathbf{0}$ , is given by

$$\left. \frac{\partial p_c(x)}{\partial c_a} \right|_{c=0} = \frac{1}{2^n} \text{Tr}[\hat{\Pi}_x \hat{P}_a] = \frac{1}{2^n} \sum_j \lambda_{xj} \langle \phi_{xj} | \hat{P}_a | \phi_{xj} \rangle. \quad (120)$$

We now consider the diagonal components of the corresponding FIM evaluated at  $\mathbf{c} = \mathbf{0}$ . The diagonal element corresponding to  $c_a$  takes the form

$$[\mathbf{F}_{\mathbf{c}=\mathbf{0}}]_{aa} = \frac{1}{2^n} \sum_x \frac{\text{Tr}[\hat{\Pi}_x \hat{P}_a]^2}{\text{Tr}[\hat{\Pi}_x]}. \quad (121)$$

Using the Cauchy-Schwarz inequality,

$$\left( \sum_j c_j a_j b_j \right)^2 \leq \left( \sum_j c_j a_j^2 \right) \left( \sum_j c_j b_j^2 \right), \quad (122)$$

we obtain

$$\begin{aligned} (\text{Tr}[\hat{\Pi}_x \hat{P}_a])^2 &= \left( \sum_j \lambda_{xj} \langle \phi_{xj} | \hat{P}_a | \phi_{xj} \rangle \right)^2 \\ &\leq \left( \sum_j \lambda_{xj} \right) \left( \sum_j \lambda_{xj} \langle \phi_{xj} | \hat{P}_a | \phi_{xj} \rangle^2 \right). \end{aligned} \quad (123)$$

Substituting this bound into Eq. (121) yields

$$[\mathbf{F}_{\mathbf{c}=\mathbf{0}}]_{aa} \leq \frac{1}{2^n} \sum_{x,j} \lambda_{xj} \langle \phi_{xj} | \hat{P}_a | \phi_{xj} \rangle^2. \quad (124)$$

Summing over all Pauli indices, we obtain

$$\begin{aligned} \sum_{a=1}^{4^n-1} [\mathbf{F}_{\mathbf{c}=\mathbf{0}}]_{aa} &\leq \frac{1}{2^n} \sum_{x,j} \lambda_{xj} \sum_{a=1}^{4^n-1} \langle \phi_{xj} | \hat{P}_a | \phi_{xj} \rangle^2 \\ &\leq \frac{2^n - 1}{2^n} \sum_{x,j} \lambda_{xj} = \frac{2^n - 1}{2^n} \sum_x \text{Tr}[\hat{\Pi}_x] = 2^n - 1. \end{aligned} \quad (125)$$

Here, we have used the fact that each  $|\phi_{xj}\rangle \langle \phi_{xj}|$  is a pure state and therefore satisfies the purity constraint in Eq. (68)

$$\sum_{a=1}^{4^n-1} \text{Tr}[|\phi_{xj}\rangle \langle \phi_{xj}| \hat{P}_a]^2 \leq 2^n - 1. \quad (126)$$

Therefore, there exists  $a$  such that

$$[\mathbf{F}_{\mathbf{c}=\mathbf{0}}]_{aa} \leq \frac{2^n - 1}{4^n - 1}. \quad (127)$$

Using the basic relation

$$\sup_c \max_a [\mathbf{F}_{\mathbf{c}}^{-1}]_{aa} \geq [\mathbf{F}_{\mathbf{c}=\mathbf{0}}^{-1}]_{aa} \geq 1/[\mathbf{F}_{\mathbf{c}=\mathbf{0}}]_{aa}, \quad (128)$$

together with Theorem 2, we conclude that the sample complexity  $N_{\text{samp}}$  required to satisfy the  $(\epsilon, \mathcal{L}_\infty, \delta)$ -criterion for Pauli expectation value learning without quantum memory is lower bounded by

$$N_{\text{samp}} \gtrsim W_0(\delta^{-2}/8\pi)\epsilon^{-2}2^n, \quad (129)$$

where  $1 < W_0(\delta^{-2}/8\pi) \leq 2 \log \delta^{-1} - \log 8\pi$  for  $0 < \delta < 1/(\sqrt{8\pi e})$ . Here, each measurement round consumes a single copy of the unknown state, so the state-copy count coincides with the sample complexity  $M$ . This matches the exponential lower bound in the number of qubits  $n$ , namely  $\Omega(2^n \epsilon^{-2})$ , a result originally established in Ref. [17] using a different proof technique from our FIM approach. The difference is that Ref. [17] fixes  $\epsilon = 1/2$  and  $\delta = 1/3$ , whereas our lower bound depends explicitly on  $\delta$  and applies in the range  $0 < \delta < 1/\sqrt{8\pi e}$  in the asymptotic small-error regime  $\epsilon \rightarrow 0$ .

Lastly, it is worth investigating the corresponding QFIM. As in Sec. III C 3, one may ask whether an exponentially growing lower bound can already be inferred directly from the QFIM, independently of measurement restrictions. From Eq. (19), the diagonal component of the inverse of the QFIM with respect to  $c_a$  is given by

$$[\mathbf{J}_{\mathbf{c}}^{-1}]_{aa} = 1 - c_a^2. \quad (130)$$

In particular, at the maximally mixed state  $\mathbf{c} = \mathbf{0}$ , this reduces to

$$[\mathbf{J}_{\mathbf{c}=\mathbf{0}}^{-1}]_{aa} = 1, \quad \forall a. \quad (131)$$

Consequently, the QFIM alone does not yield an exponentially large lower bound. This observation highlights an important distinction between the FIM and QFIM analyses in the present setting. While the diagonal components of the QFIM quantify the ultimate sensitivity achievable when estimating a single parameter with an optimal measurement tailored to that parameter, they do not capture the incompatibility between the optimal measurements for different Pauli expectation values. The exponential lower bound derived above therefore does not originate from a lack of quantum sensitivity at the state level, but rather from the fundamental incompatibility of simultaneously estimating many non-commuting parameters using a single measurement strategy.

## IV. MAIN RESULT 2: SAMPLE COMPLEXITY OF $\mathcal{L}_2$ -DISTANCE-BASED LEARNING

In this section, we establish an upper bound on the sample complexity under the  $\mathcal{L}_2$ -criterion. In addition, we show that the lower bound on the sample complexity for learning with respect to the  $\mathcal{L}_2$ -distance can be characterized by the largest eigenvalue of the inverse FIM.

### A. Upper bound on $\mathcal{L}_2$ -distance learning

Let us assume that the assumptions (A1) and (A2) and the standard regularity conditions (R1)–(R5) in Appendix 3 B 4 are satisfied. We establish an upper bound

on the minimal sample complexity required to guarantee the  $(\epsilon, \ell_2, \delta)$ -criterion.

*Theorem 3.* (Simplified small-error upper bound for  $\ell_2$  learning). For fixed  $0 < \delta \leq 1$  and  $d < \infty$ , the minimal number of repetitions  $M = M(\epsilon)$  required to guarantee that the MLE satisfies

$$\Pr [\|\tilde{\theta}^{\text{ML}} - \theta\|_2 \leq \epsilon] \geq 1 - \delta \quad \text{for all } \theta \in \Theta \quad (132)$$

is upper bounded, in the small-error limit  $\epsilon \rightarrow 0$ , as

$$M \lesssim d W_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\theta}^{-1}]_{aa} \epsilon^{-2}. \quad (133)$$

Equivalently, the minimal sample complexity is upper bounded as

$$N_{\text{samp}} \lesssim N_{\mathcal{E}} d W_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\theta}^{-1}]_{aa} \epsilon^{-2}. \quad (134)$$

The proof of Theorem 3 is provided in Appendix E, where we also derive a fully explicit non-asymptotic finite- $\epsilon$  bound.

Theorem 3 shows that, in the small-error regime  $\epsilon \rightarrow 0$ , the sample complexity under the  $\ell_2$ -criterion contains an additional multiplicative factor  $d$  compared with the  $\ell_{\infty}$ -criterion in Theorems 1 and 2. The additional factor  $d$  in Eqs. (133) and (134) originates from the norm conversion used to reduce the  $\ell_2$ -criterion to the  $\ell_{\infty}$ -criterion: it is sufficient to require  $\|\tilde{\theta}^{\text{ML}} - \theta\|_{\infty} \leq \epsilon/\sqrt{d}$ . Since the FIM-based sample complexity scales as the inverse square of the target accuracy, this replacement produces the multiplicative factor  $d$ . This factor is therefore distinct from the logarithmic  $d$ -dependence in the  $\ell_{\infty}$  bounds, which arises from simultaneously controlling the failure probability over all coordinates.

### B. Lower bound on $\ell_2$ -distance learning

We establish a lower bound on the minimal sample complexity required to guarantee the  $(\epsilon, \ell_2, \delta)$ -criterion under the assumptions (A1) and (A2) and the standard regularity conditions (R1)–(R5) in Appendix 3 B 4 are satisfied.

*Theorem 4.* (Simplified small-error lower bound for  $\ell_2$  learning). For fixed  $0 < \delta < 1/\sqrt{8\pi e}$  and  $d < \infty$ , the minimal number of repetitions  $M = M(\epsilon, \delta, d)$  required to guarantee that the MLE satisfies

$$\Pr [\|\tilde{\theta}^{\text{ML}} - \theta\|_2 \leq \epsilon] \geq 1 - \delta \quad \text{for all } \theta \in \Theta \quad (135)$$

is lower bounded, in the small-error limit  $\epsilon \rightarrow 0$ , as

$$M \gtrsim W_0(\delta^{-2}/8\pi) \sup_{\theta \in \Theta} \mu_{\max}(\mathbf{F}_{\theta}^{-1}) \epsilon^{-2}, \quad (136)$$

where  $\mu_{\max}(\mathbf{F}_{\theta}^{-1})$  is the largest eigenvalue of  $\mathbf{F}_{\theta}^{-1}$ . Equivalently, the minimal sample complexity is lower bounded as

$$N_{\text{samp}} \gtrsim N_{\mathcal{E}} W_0(\delta^{-2}/8\pi) \sup_{\theta \in \Theta} \mu_{\max}(\mathbf{F}_{\theta}^{-1}) \epsilon^{-2}. \quad (137)$$

The proof of Theorem 4 is provided in Appendix F, where we also derive a fully explicit non-asymptotic finite- $\epsilon$  bound. The  $\ell_2$  lower bound is determined by the largest eigenvalue of the inverse FIM. This captures the most statistically ill-conditioned direction in the parameter space, which dictates the minimal sample size required to control the overall quadratic error.

## C. Application to Pauli channel learning

### 1. Entanglement-assisted scheme

We consider the same estimation protocol as in Sec. III C 2, for which we have  $N_{\mathcal{E}} = 1$  and

$$\sup_{\lambda} \max_{a \in [d]} [\mathbf{F}_{\lambda}^{-1}]_{aa} = 1. \quad (138)$$

Combining this with Theorem 3, to learn  $\lambda$  under the  $(\epsilon, \ell_2, \delta)$ -criterion, the minimal sample complexity  $N_{\text{samp}}$  is upper bounded as

$$N_{\text{samp}} \lesssim 4^n W_0(8\pi^{-1}\delta^{-2}4^{2n}) \epsilon^{-2}. \quad (139)$$

Next, let us derive the lower bound. As shown in Sec. III C 2, when a maximally entangled state is employed and the Bell measurement is performed, the probability of obtaining outcome  $x$  is given by

$$p_{\lambda}(x) := \frac{1}{4^n} \sum_{a=0}^{4^n-1} \lambda_a (-1)^{\langle [x]_2, [a]_2 \rangle}. \quad (140)$$

Our goal here is to determine the maximum eigenvalue of the inverse FIM with respect to  $\lambda$ .

Although  $\lambda_0$  is fixed by normalization, it is convenient to introduce the following auxiliary full information matrix  $\mathbf{F}_0$ . We first compute the FIM with respect to the full parameter set  $(\lambda_0, \lambda_1, \dots, \lambda_{4^n-1})$ , from which the FIM corresponding to  $\lambda$  can be readily obtained. The FIM associated with this parametrization, which we denote by  $\mathbf{F}_0$ , is defined according to Eq. (10) and admits the decomposition

$$\mathbf{F}_0 = \mathbf{U}^T \mathbf{D} \mathbf{U}. \quad (141)$$

Here,  $\mathbf{U} \in \mathbb{R}^{4^n \times 4^n}$  and  $\mathbf{D} \in \mathbb{R}^{4^n \times 4^n}$  are given element-wise by

$$[\mathbf{U}]_{ab} := \frac{1}{\sqrt{4^n}} (-1)^{\langle [a]_2, [b]_2 \rangle}, \quad [\mathbf{D}]_{ab} := \frac{1}{p_a 4^n} \delta_{ab}. \quad (142)$$

The matrix  $\mathbf{U}$  is orthogonal and corresponds to the normalized Walsh-Hadamard transform, while  $\mathbf{D}$  is diagonal, with entries given by the inverse of the Pauli error rates. As a consequence, the eigenvalues of  $\mathbf{F}_0$  are given by the diagonal elements of  $\mathbf{D}$ .

For notational simplicity, for any positive-definite matrix  $\mathbf{A}$  we denote its eigenvalues in nondecreasing order by  $0 < \mu_1(\mathbf{A}) \leq \mu_2(\mathbf{A}) \leq \dots$ .

The FIM with respect to the reduced parameter vector  $\lambda$  is obtained by restricting to the subspace orthogonal to the all-ones direction (equivalently, by removing the coordinate corresponding to  $\lambda_0$ ). Concretely, the resulting matrix  $\mathbf{F}$  is the principal submatrix of  $\mathbf{F}_0$  obtained by deleting the first row and the first column. Hence, combining Eq. (141) with the eigenvalue interlacing theorem for principal submatrices [51], we have

$$\mu_1(\mathbf{F}_0) \leq \mu_1(\mathbf{F}) \leq \mu_2(\mathbf{F}_0), \quad (143)$$

and therefore

$$\mu_{\max}(\mathbf{F}^{-1}) = \mu_1^{-1}(\mathbf{F}) \geq \mu_2^{-1}(\mathbf{F}_0). \quad (144)$$

Since  $\mathbf{F}_0 = \mathbf{U}^T \mathbf{D} \mathbf{U}$  with  $\mathbf{U}$  orthogonal, the eigenvalues of  $\mathbf{F}_0$  coincide with the diagonal entries of  $\mathbf{D}$ , i.e.,  $\{1/(4^n p_a)\}_a$ . Let  $\{p_{(1)} \geq p_{(2)} \geq \dots\}$  be the Pauli rates sorted in nonincreasing order. Then  $\mu_2(\mathbf{F}_0) = 1/(4^n p_{(2)})$ , and thus

$$\mu_{\max}(\mathbf{F}^{-1}) \geq 4^n p_{(2)}. \quad (145)$$

To obtain an explicit lower bound, consider the valid choice of Pauli rates

$$p_{(1)} = \frac{1}{2}, \quad p_{(2)} = \frac{1}{4}, \quad p_{(a)} = \frac{1}{4(4^n - 2)} \text{ for all } a > 2. \quad (146)$$

For this choice, Eq. (146) yields

$$\mu_{\max}(\mathbf{F}^{-1}) \geq 4^{n-1}. \quad (147)$$

As a consequence, according to Theorem 4, the lower bound is

$$N_{\text{samp}} \gtrsim W_0(\delta^{-2}/8\pi)\epsilon^{-2}4^{n-1}. \quad (148)$$

Therefore, even when entanglement is employed, the estimation protocol still requires an exponentially large sample complexity.

## V. DISCUSSION

In this work, we derive task-independent upper and lower bounds on the sample complexity required to learn the parameters of a given quantum system within asymptotic additive error  $\epsilon \rightarrow 0$  and with success probability at least  $1 - \delta$ , under both the  $\ell_\infty$ - and  $\ell_2$ -criteria. Our bounds apply to general learning protocols and recover tight task-specific bounds, particularly for Pauli eigenvalue learning and Pauli expectation-value learning [17,19,25], in the asymptotic limit  $\epsilon \rightarrow 0$ . However, the finite  $\epsilon$  bounds presented in Theorems 1–4 in the Appendixes are not as tight as the corresponding bounds previously obtained in Refs. [17,19,25]. It would therefore be an interesting direction for future work to investigate whether tighter yet still task-independent bounds can be established without invoking the asymptotic limit  $\epsilon \rightarrow 0$ .

Notably, these bounds are governed by the FIM, a central quantity in quantum metrology. Quantum metrology has a longer history than quantum learning theory, and a wide range of techniques for maximizing Fisher information has been developed. For instance, prior work has demonstrated how symmetries of quantum systems can be exploited to enhance the Fisher information [52]. In addition, the application of quantum error correction to protect Fisher information in noisy settings has been extensively studied [53–62]. It would be worthwhile to investigate whether such metrological tools can systematically inform and strengthen the theoretical foundations of quantum learning.

Our analysis focuses exclusively on sample complexity. Query complexity, namely, the number of times the quantum system must be accessed or interrogated during the learning process, or the total interrogation time, also constitutes a fundamental resource. Hamiltonian learning from dynamical queries provides a representative setting for such an extension. More concretely, one may consider learning the parameters of an unknown Hamiltonian  $\hat{H}_\theta = \sum_a \theta_a \hat{H}_a$  through controlled or uncontrolled queries to the time evolution  $\hat{U}_\theta(t) = e^{-i\hat{H}_\theta t}$  [14,15]. In this context, a central open question is how the query complexity scales with the target precision  $\epsilon$  and failure probability  $\delta$ . Exploring whether query complexity can be characterized through the FIM represents another promising direction for future research.

In establishing our results, we focus primarily on the regime in which the FIM is invertible [63,64]. The singular FIM case is treated separately in Appendix G. This corresponds to situations in which not all parameters are unbiasedly estimable. In such cases, we consider learning restricted to the subspace of parameters that admit unbiased estimation. In this setting, we show that the inverse FIM appearing in the sample complexity bounds can be naturally replaced by its Moore-Penrose pseudoinverse.

Lastly, during the preparation of this manuscript, we became aware of closely related work [65]. Our work and that recent study were conducted independently and without mutual influence. Although both investigate sample complexity in quantum estimation problems, their scope and technical assumptions differ. In particular, Ref. [65] considers shadow tomography without assuming a specific estimator in the lower bound, whereas our analysis addresses general quantum parameter estimation and derives explicit bounds under the assumption of maximum-likelihood estimation. Moreover, the proof techniques employed in the two works are different.

## ACKNOWLEDGMENTS

H. K. is supported by the IITP (Grants No. RS-2025-25464252, No. RS-2025-02219034, and No. RS-2024-00437191) and the NRF (Grants No. RS-2026-25476454, No. RS-2025-25464492, and No. RS-2024-00442710) funded by the Ministry of Science and ICT (MSIT), Korea. S. H. L. is supported by the 2025 Research Fund (Grant No. 1.250007.01) of Ulsan National Institute of Science & Technology (UNIST), Institute of Information & Communications Technology Planning & Evaluation (IITP) (Grants No. RS-2023-00227854 and No. RS-2025-02283189) and National Research Foundation of Korea (Grant No. RS-2025-25464492). L. J. acknowledges support from the ARO (Grant No. W911NF-23-1-0077), ARO MURI (Grant No. W911NF-21-1-0325), AFOSR MURI (Grants No. FA9550-21-1-0209 and No. FA9550-23-1-0338), DARPA (Grants No. HR0011-24-9-0359 and No. HR0011-24-9-0361), NSF (Grants No. ERC-1941583, No. OMA-2137642, No. OSI-2326767, No. CCF-2312755, and No. OSI-2426975), and the Packard Foundation (Grant No. 2020-71479). This work was supported by the faculty research fund of Sejong University in 2025.

## DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

## APPENDIX A: NOTATIONS

In this section, we introduce the notation used throughout the manuscript.

### 1. Single-sample derivatives and sample averages

We begin by introducing the log-likelihood function associated with a single measurement outcome  $x$ . For a given parameter vector  $\boldsymbol{\theta} \in \mathbb{R}^d$ , the log-likelihood is defined as

$$\ell_{\boldsymbol{\theta}}(x) := \log p_{\boldsymbol{\theta}}(x), \quad (\text{A1})$$

where  $p_{\boldsymbol{\theta}}(x)$  denotes the probability of observing the measurement outcome  $x$  conditioned on the parameter vector  $\boldsymbol{\theta}$ . We then denote the first-, second-, and third-order derivatives of the log-likelihood function with respect to the parameter vector  $\boldsymbol{\theta} \in \mathbb{R}^d$  as

$$\ell_{\boldsymbol{\theta}}^{(1)}(x) := \nabla_{\boldsymbol{\theta}} \ell_{\boldsymbol{\theta}}(x) \in \mathbb{R}^d, \quad (\text{A2})$$

$$\ell_{\boldsymbol{\theta}}^{(2)}(x) := \nabla_{\boldsymbol{\theta}}^2 \ell_{\boldsymbol{\theta}}(x) \in \mathbb{R}^{d \times d}, \quad (\text{A3})$$

$$\ell_{\boldsymbol{\theta}}^{(3)}(x) := \nabla_{\boldsymbol{\theta}}^3 \ell_{\boldsymbol{\theta}}(x) \in \mathbb{R}^{d \times d \times d}. \quad (\text{A4})$$

We now extend these definitions to the case of multiple observations. Let  $\mathbf{x} := \{x_i\}_{i=1}^M$  denote an independent and identically distributed (i.i.d.) sample drawn according to  $p_{\boldsymbol{\theta}}(x)$ . The total log-likelihood associated with the sample is then given by

$$\ell_{\boldsymbol{\theta}}(\mathbf{x}) := \sum_{i=1}^M \ell_{\boldsymbol{\theta}}(x_i). \quad (\text{A5})$$

Because the total log-likelihood is additive over samples, its derivatives naturally decompose into sums of the corresponding measurement outcome quantities.

Motivated by this observation, we define the empirical score function, the empirical Hessian, and the empirical third-derivative tensor as

$$\mathbf{S}_{\boldsymbol{\theta}}(\mathbf{x}) := \frac{1}{M} \nabla_{\boldsymbol{\theta}} \ell_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^M \ell_{\boldsymbol{\theta}}^{(1)}(x_i), \quad (\text{A6})$$

$$\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x}) := \frac{1}{M} \nabla_{\boldsymbol{\theta}}^2 \ell_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^M \ell_{\boldsymbol{\theta}}^{(2)}(x_i), \quad (\text{A7})$$

$$\mathbf{R}_{\boldsymbol{\theta}}(\mathbf{x}) := \frac{1}{M} \nabla_{\boldsymbol{\theta}}^3 \ell_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^M \ell_{\boldsymbol{\theta}}^{(3)}(x_i). \quad (\text{A8})$$

Each of these empirical quantities represents an average over the measurement outcomes and converges, under suitable regularity conditions, to the corresponding expectation value in the large-sample limit.

We next summarize the expectation values of the single measurement outcome derivatives, which encode the fundamental statistical structure of the model. Taking the expectation with respect to the probability distribution of the measurement outcomes, we obtain

$$\mathbb{E}[\ell_{\boldsymbol{\theta}}^{(1)}(x)] = \mathbf{0}, \quad (\text{A9})$$

$$\mathbb{E}[\ell_{\boldsymbol{\theta}}^{(1)}(x) \ell_{\boldsymbol{\theta}}^{(1)}(x)^T] = \mathbf{F}_{\boldsymbol{\theta}}(\{\hat{\Pi}_x\}_x), \quad (\text{A10})$$

$$\mathbb{E}[\ell_{\boldsymbol{\theta}}^{(2)}(x)] = -\mathbf{F}_{\boldsymbol{\theta}}(\{\hat{\Pi}_x\}_x). \quad (\text{A11})$$

Here,  $\mathbb{E}$  denotes the expectation with respect to the probability distribution of the measurement outcomes  $x$ .

For notational simplicity, we henceforth suppress the explicit dependence on the measurement outcomes, and denote

$$\mathbf{S}_g(\mathbf{x}) := \mathbf{S}_g, \quad (\text{A12})$$

$$\mathbf{H}_g(\mathbf{x}) := \mathbf{H}_g, \quad (\text{A13})$$

$$\mathbf{R}_g(\mathbf{x}) := \mathbf{R}_g, \quad (\text{A14})$$

$$\mathbf{F}_g(\{\hat{\Pi}_x\}_x) := \mathbf{F}_g. \quad (\text{A15})$$

## 2. Norm conventions

Throughout the manuscript,  $\langle \mathbf{u}, \mathbf{v} \rangle := \mathbf{u}^T \mathbf{v}$  denotes the standard inner product on  $\mathbb{R}^d$ , and  $\|\mathbf{v}\|_2 := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$  denotes the corresponding Euclidean norm.

We first recall the operator norm for linear maps. For a matrix  $\mathbf{B} \in \mathbb{R}^{d \times d}$ , the operator norm induced by  $\|\cdot\|_2$  is defined as

$$\|\mathbf{B}\|_{\text{op}} := \sup_{\mathbf{v} \in \mathbb{R}^d: \|\mathbf{v}\|_2=1} \|\mathbf{B}\mathbf{v}\|_2. \quad (\text{A16})$$

This norm quantifies the maximal amplification of a unit vector under the action of  $\mathbf{B}$ .

We now extend these conventions to third-order tensors. Let  $\mathbf{T} \in \mathbb{R}^{d \times d \times d}$  be a third-order tensor. With respect to the standard basis  $\{\mathbf{e}_i\}_{i=1}^d$  of  $\mathbb{R}^d$ , we define its components by

$$T_{ijk} := \mathbf{T}[\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k], \quad i, j, k \in \{1, \dots, d\}. \quad (\text{A17})$$

Accordingly, for any  $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ , the tensor  $\mathbf{T}$  induces a trilinear form defined by

$$\mathbf{T}[\mathbf{u}, \mathbf{v}, \mathbf{w}] := \sum_{i,j,k=1}^d T_{ijk} u_i v_j w_k, \quad (\text{A18})$$

where the coordinates are given by  $u_i := \langle \mathbf{e}_i, \mathbf{u} \rangle$ ,  $v_j := \langle \mathbf{e}_j, \mathbf{v} \rangle$ , and  $w_k := \langle \mathbf{e}_k, \mathbf{w} \rangle$ .

In direct analogy with the matrix case, we define the operator norm of the tensor  $\mathbf{T}$  as

$$\|\mathbf{T}\|_{\text{op}} := \sup_{\substack{\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^d \\ \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = 1}} |\mathbf{T}[\mathbf{u}, \mathbf{v}, \mathbf{w}]|. \quad (\text{A19})$$

This norm captures the maximal magnitude of the trilinear form evaluated on unit vectors.

Next, we introduce a partially contracted form of the tensor. For fixed vectors  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ , we define the vector-valued contraction  $\mathbf{T}[\mathbf{v}, \mathbf{w}] \in \mathbb{R}^d$  implicitly by the relation

$$\langle \mathbf{u}, \mathbf{T}[\mathbf{v}, \mathbf{w}] \rangle := \mathbf{T}[\mathbf{u}, \mathbf{v}, \mathbf{w}] \quad \forall \mathbf{u} \in \mathbb{R}^d. \quad (\text{A20})$$

By construction, this definition ensures that  $\mathbf{T}[\mathbf{u}, \mathbf{v}, \mathbf{w}]$  depends linearly on each of its arguments.

In coordinates, the components of the contracted vector are given explicitly by

$$\langle \mathbf{e}_i, \mathbf{T}[\mathbf{v}, \mathbf{w}] \rangle = \sum_{j,k=1}^d T_{ijk} v_j w_k, \quad i = 1, \dots, d. \quad (\text{A21})$$

As a direct consequence of the definition of the operator norm, the Euclidean norm of the contraction satisfies

$$\|\mathbf{T}[\mathbf{v}, \mathbf{w}]\|_2 = \sup_{\mathbf{u} \in \mathbb{R}^d: \|\mathbf{u}\|_2=1} |\mathbf{T}[\mathbf{u}, \mathbf{v}, \mathbf{w}]| \leq \|\mathbf{T}\|_{\text{op}} \|\mathbf{v}\|_2 \|\mathbf{w}\|_2. \quad (\text{A22})$$

This inequality will be used repeatedly to control tensor contractions in subsequent proofs.

## APPENDIX B: FREQUENTLY EXPLOITED PROOF TOOLS

### 1. Berry-Esseen theorem

For details, see Refs. [66–68]. The Berry-Esseen theorem is a quantitative refinement of the central limit theorem: it provides an explicit rate at which the distribution of a normalized sum of independent random variables approaches the standard normal distribution. Let  $X := (X_1, X_2, \dots, X_M)$  be i.i.d. random variables with

$$\mathbb{E}[X_1] = 0, \quad \text{Var}(X_1) = \sigma^2 > 0, \quad \rho := \mathbb{E}[|X_1|^3] < \infty. \quad (\text{B1})$$

Next, consider the standardized sum

$$S(X) = \frac{1}{\sigma\sqrt{M}} \sum_{i=1}^M X_i. \quad (\text{B2})$$

The Berry-Esseen theorem states that there exists a universal constant  $C > 0$  such that for all  $M \geq 1$ ,

$$\sup_{x \in \mathbb{R}} |\Pr[S \leq x] - \Phi(x)| \leq C \frac{\rho}{\sigma^3 \sqrt{M}}. \quad (\text{B3})$$

Here,  $\Phi(x)$  denotes the cumulative distribution function of the standard normal distribution  $\mathcal{N}(0, 1)$ ,

$$\Phi(x) = \int_{-\infty}^x \phi(t) dt, \quad \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \quad (\text{B4})$$

### 2. Mills ratio inequality

The Mills ratio inequality provides sharp bounds on the Gaussian tail [69–72]. For all  $x > 0$ , one has

$$\frac{x}{1+x^2} \phi(x) < 1 - \Phi(x) < \frac{\phi(x)}{x}. \quad (\text{B5})$$

### 3. Lambert $W_0$ function

#### a. Definition

See Refs. [49,50] for details. In mathematics, the Lambert  $W$  function is defined as the inverse of the map

$$f(w) = we^w, \quad (\text{B6})$$

where  $w$  is a complex number. The principal branch, denoted by  $W_0(z)$ , is the single-valued branch that is real-valued on its maximal real domain. By definition,  $W_0(z)$  satisfies

$$W_0(z)e^{W_0(z)} = z. \quad (\text{B7})$$

When restricted to real variables, the equation

$$ye^y = x \quad (\text{B8})$$

admits real solutions if and only if

$$x \geq -\frac{1}{e}. \quad (\text{B9})$$

On this domain, the principal branch gives the real solution

$$y = W_0(x), \quad x \geq -\frac{1}{e}, \quad (\text{B10})$$

with

$$W_0\left(-\frac{1}{e}\right) = -1, \quad W_0(0) = 0, \quad W_0(e) = 1. \quad (\text{B11})$$

For  $x \geq e$ , the principal branch satisfies the inequality

$$W_0(x) \leq \log x. \quad (\text{B12})$$

Indeed, if  $y = W_0(x)$ , then  $x = ye^y$ . If  $y = \log x$ , then

$$ye^y = (\log x)x \geq x \quad (x \geq e), \quad (\text{B13})$$

which implies  $y \leq \log x$  by the monotonicity of  $ye^y$  for  $y \geq 1$ .

#### b. Asymptotic regime

For large positive  $z$ , the principal branch grows logarithmically. The Lambert  $W_0$  function has the asymptotic expansion (in the regime  $z \rightarrow \infty$ )

$$W_0(z) = \log z - \log \log z + \frac{\log \log z}{\log z} + O\left(\frac{(\log \log z)^2}{(\log z)^2}\right). \quad (\text{B14})$$

To leading order,

$$W_0(z) \sim \log z - \log \log z. \quad (\text{B15})$$

#### c. Concavity

We first consider the first derivative of  $W_0$ . By differentiating  $We^W = z$  with respect to  $z$ , we obtain

$$\frac{d}{dz}(We^W) = W'e^W + We^WW' = W'e^W(1+W) = 1, \quad (\text{B16})$$

hence

$$W'(z) = \frac{1}{e^{W(z)}(1+W(z))}. \quad (\text{B17})$$

Using  $e^{W(z)} = z/W(z)$  (from  $We^W = z$ ), this becomes

$$W'(z) = \frac{W(z)}{z(1+W(z))}. \quad (\text{B18})$$

We then calculate the second derivative. By differentiating  $W'(z) = \frac{1}{z} \frac{W}{1+W}$  with respect to  $z$ , we obtain

$$\begin{aligned} W''(z) &= \frac{d}{dz} \left( \frac{1}{z} \right) \frac{W}{1+W} + \frac{1}{z} \frac{d}{dz} \left( \frac{W}{1+W} \right) \\ &= -\frac{1}{z^2} \frac{W}{1+W} + \frac{1}{z} \left( \frac{d}{dz} \frac{W}{1+W} \right) W' \\ &= -\frac{1}{z^2} \frac{W}{1+W} + \frac{1}{z} \left( \frac{1}{(1+W)^2} \right) \left( \frac{W}{z(1+W)} \right) \\ &= -\frac{W}{z^2(1+W)} + \frac{W}{z^2(1+W)^3} \\ &= \frac{W}{z^2} \left( -\frac{1}{1+W} + \frac{1}{(1+W)^3} \right) = -\frac{W^2(W+2)}{z^2(1+W)^3}. \end{aligned} \quad (\text{B19})$$

For  $z > 0$  on the principal branch,  $W_0(z) > 0$ , hence  $W_0(z)^2 > 0$  and  $W_0(z) + 2 > 0$ . Moreover,  $z^2(1+W_0(z))^3 > 0$ . Therefore, by (B19),

$$W''_0(z) = -\frac{W_0(z)^2(W_0(z)+2)}{z^2(1+W_0(z))^3} < 0 \quad (z > 0), \quad (\text{B20})$$

so  $W_0$  is strictly concave on  $(0, \infty)$ .

### 4. Standard regularity assumptions for the proof of main theorems

For the proofs of the main theorems, we impose, in addition to assumptions (A1) and (A2) introduced in the main text,

- (A1) Unique maximizer and stationary point:  $\mathcal{L}_\theta(\mathbf{x})$  has a unique maximizer  $\tilde{\theta}^{\text{ML}}$  in the interior of the parameter domain  $\Theta$ , which is also the unique stationary point.

- (A2) Smoothness:  $\ell_\theta(\mathbf{x})$  is three times continuously differentiable with respect to  $\theta$  on the parameter domain  $\Theta$ .

and the following additional standard regularity conditions, which are used in the proofs of the main theorems:

- (R1) Local containment of the Taylor neighborhood in the ambient parameter domain: Let  $\tilde{\Theta}$  be an open ambient parameter domain containing the parameter space  $\Theta$ , i.e.,  $\Theta \subseteq \tilde{\Theta}$ . There exists a constant  $r_0 > 0$  such that

$$\theta + \Delta \in \tilde{\Theta} \quad \text{for all } \theta \in \Theta \quad \text{and} \quad \|\Delta\|_2 \leq r_0. \quad (\text{B21})$$

- (R2) Uniform non-singularity of the Fisher information matrix: The Fisher information matrix  $\mathbf{F}_\theta$  is non-singular for every  $\theta \in \Theta$ , and its inverse is uniformly bounded in operator norm:

$$\sup_{\theta \in \Theta} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} < \infty. \quad (\text{B22})$$

- (R3) Uniform second-moment bound for the centered Hessian: The centered Hessian of the log-likelihood has a uniformly bounded second moment:

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[\text{Tr}\{(\ell_\theta^{(2)}(X) + \mathbf{F}_\theta)^2\}] < \infty. \quad (\text{B23})$$

- (R4) Uniform second-moment bound for the local third-derivative envelope: The local supremum of the third derivative of the log-likelihood has a uniformly bounded second moment:

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[ \left( \sup_{\|\Delta\|_2 \leq r_0} \|\ell_{\theta+\Delta}^{(3)}(X)\|_{\text{op}} \right)^2 \right] < \infty. \quad (\text{B24})$$

- (R5) Uniform bound on the standardized third absolute moment of the projected score: The standardized third absolute moment of the projected

Fisher-preconditioned score is uniformly bounded over all unit directions:

$$\sup_{\theta \in \Theta} \sup_{\|u\|_2=1} \frac{\mathbb{E}_\theta[|u^\top \mathbf{F}_\theta^{-1} \ell_\theta^{(1)}(X)|^3]}{(u^\top \mathbf{F}_\theta^{-1} u)^{3/2}} < \infty. \quad (\text{B25})$$

The regularity conditions above admit the following intuitive interpretations. Condition (R1) guarantees that a uniform neighborhood around each parameter value remains within the ambient parameter domain, thereby allowing local Taylor expansions to be carried out without boundary issues. Condition (R2) ensures that the Fisher information matrix remains uniformly nondegenerate, so that all parameter directions are identifiable and the estimation problem does not become arbitrarily ill-conditioned. Condition (R3) controls the fluctuations of the observed Hessian around its mean, while condition (R4) limits the local variation of the Hessian by imposing a uniform moment bound on the third derivative of the log-likelihood. Finally, condition (R5) prevents the standardized projected score from having excessively heavy tails and provides the moment control required for quantitative Gaussian approximations, such as Berry-Esseen-type bounds.

### APPENDIX C: PROOF OF THE UPPER BOUND FOR $\ell_\infty$

Assume that conditions (A1) and (A2) and the standard regularity conditions (R1)–(R5) stated in Appendix 3 B 4 hold. Then the following theorem applies.

*Theorem 1.* Fix  $d < \infty$ . There exists a constant  $\epsilon_0 > 0$  such that, for every  $0 < \epsilon < \epsilon_0$  and  $0 < \delta \leq 1$ , the following holds. Let  $M_0(\epsilon, \delta, d)$  denote the minimal number of repetitions such that

$$\forall M \geq M_0(\epsilon, \delta, d): \Pr[\|\tilde{\theta}^{\text{ML}} - \theta\|_\infty \leq \epsilon] \geq 1 - \delta \quad (\text{C1})$$

for all  $\theta \in \Theta$ .

$M_0$  is then upper bounded as

$$M_0 \leq \sup_{\theta \in \Theta} \max \left\{ \left( \frac{D}{\tau_{0-}} \right)^2, \left( \frac{2d\eta}{\delta} \right)^2, \left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} + \sqrt{\left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} \right)^2 - \frac{2d\eta}{\delta} \frac{D}{\tau_{0-}}} \right)^2 \right\}, \quad (\text{C2})$$

where

$$W_0 := W_0(8\pi^{-1}\delta^{-2}d^2), \quad (\text{C3})$$

$$\tau_{0-} := \left( 1 - \epsilon \frac{d\mu_R}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \right) \epsilon, \quad (\text{C4})$$

$$D := \left( \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \sqrt{\frac{4V_H}{\delta}} \sqrt{d} + \frac{1}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \sqrt{\frac{4V_R}{\delta}} d \epsilon \right) \epsilon, \quad (\text{C5})$$

$$\eta := \frac{1}{d} \sum_{a=1}^d \frac{2C\rho}{\sigma_{\theta,a}^3}, \quad (\text{C6})$$

$$\rho := \sup_{\theta \in \Theta} \max_{a \in [d]} \mathbb{E}[|\mathbf{e}_a^T \mathbf{F}_\theta^{-1} \mathcal{L}_\theta^{(1)}(x)|^3], \quad (\text{C7})$$

$$\sigma := \sqrt{\sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa}}. \quad (\text{C8})$$

In the small-error limit  $\epsilon \rightarrow 0$ , Eq. (C2) reduces to

$$M_0 \lesssim W_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa} \epsilon^{-2}, \quad (\text{C9})$$

where the notation  $A_\epsilon \lesssim B_\epsilon$  denotes

$$\lim_{\epsilon \rightarrow 0} \sup \frac{A_\epsilon}{B_\epsilon} \leq 1, \quad (\text{C10})$$

i.e.,  $A_\epsilon \leq (1 + o(1))B_\epsilon$  as  $\epsilon \rightarrow 0$  with fixed  $\delta$  and  $d$ .

### 1. Proof sketch

Before beginning the proof, we specify a valid uniform choice of the small-error radius. Define

$$K_F := \sup_{\theta \in \Theta} \|\mathbf{F}_\theta^{-1}\|_{\text{op}}, \quad (\text{C11})$$

and

$$\bar{\mu}_R := \sup_{\theta \in \Theta} \mathbb{E} \left[ \sup_{\|u\|_2 \leq r_0} \|\mathcal{L}_{\theta+u}^{(3)}(X)\|_{\text{op}} \right]. \quad (\text{C12})$$

By condition (R2),  $K_F < \infty$ . Moreover, condition (R4) and the Cauchy-Schwarz inequality imply that

$$\bar{\mu}_R \leq \left\{ \sup_{\theta \in \Theta} \mathbb{E} \left[ \left( \sup_{\|u\|_2 \leq r_0} \|\mathcal{L}_{\theta+u}^{(3)}(X)\|_{\text{op}} \right)^2 \right] \right\}^{1/2} < \infty. \quad (\text{C13})$$

We consider

$$\epsilon_0 := \min \left\{ \frac{r_0}{\sqrt{d}}, \frac{1}{dK_F\bar{\mu}_R} \right\} > 0, \quad (\text{C14})$$

where the second term is understood as  $+\infty$  when  $\bar{\mu}_R = 0$ . This choice depends only on  $d$  and the uniform regularity constants and is independent of  $\theta$ ,  $\delta$ , and  $M$ . Let us define the following accuracy event on the sample space of  $M$  measurement outcomes:

$$A(\epsilon) := \{\mathbf{x}: \|\tilde{\theta}^{\text{ML}}(\mathbf{x}) - \theta\|_\infty \leq \epsilon\}, \quad (\text{C15})$$

where  $\mathbf{x} := \{x_i\}_{i=1}^M$  denotes an i.i.d. sample of measurement outcomes drawn according to the distribution  $p_\theta(\mathbf{x}) = \prod_{i=1}^M p_\theta(x_i)$ . That is,  $A(\epsilon)$  is the event that the MLE constructed from the observed data  $\mathbf{x}$  lies within the  $\ell_\infty$ -ball of radius  $\epsilon$  around the true parameter  $\theta$ . Our goal is to determine the minimal number of repetitions  $M_0(\epsilon, \delta, d)$

such that, for a prescribed accuracy  $\epsilon > 0$  and confidence level  $1 - \delta$ , with  $0 < \delta < 1$ , the MLE  $\tilde{\theta}^{\text{ML}}$  satisfies

$$\forall M \geq M_0(\epsilon, \delta, d): \Pr[A(\epsilon)] \geq 1 - \delta, \quad (\text{C16})$$

or equivalently,

$$\forall M \geq M_0(\epsilon, \delta, d): \Pr[A(\epsilon)^c] \leq \delta. \quad (\text{C17})$$

In this section, we derive an upper bound on  $M_0$  by following the procedure below. First, we identify an event  $U(\epsilon)$  such that  $A(\epsilon)^c \subseteq U(\epsilon)$ , which implies

$$\Pr[A(\epsilon)^c] \leq \Pr[U(\epsilon)]. \quad (\text{C18})$$

We then determine the minimal  $M_U(\epsilon, \delta, d)$  such that

$$\forall M \geq M_U(\epsilon, \delta, d): \Pr[U(\epsilon)] \leq \delta. \quad (\text{C19})$$

Combining (C18) and (C19) yields, for all  $M \geq M_U(\epsilon, \delta, d)$ ,

$$\Pr[A(\epsilon)^c] \leq \Pr[U(\epsilon)] \leq \delta, \quad (\text{C20})$$

which implies that the guarantee in Eq. (C17) holds whenever  $M \geq M_U(\epsilon, \delta, d)$ . Therefore,  $M_U(\epsilon, \delta, d)$  is an upper bound on the minimal number of repetitions  $M_0(\epsilon, \delta, d)$ .

### 2. Taylor expansion of the score function

The proof follows the standard MLE strategy based on a Taylor expansion of the score function around the true parameter. Related finite number of measurement outcomes likelihood results are available in the classical statistics literature [73–75]. We keep the proof self-contained to make explicit the Fisher-information dependence relevant to quantum-learning sample complexity.

We begin by introducing the MLE associated with the observed measurement outcomes  $\mathbf{x}$ . Let

$$\tilde{\theta}^{\text{ML}}(\mathbf{x}) := \arg \max_{\theta} \mathcal{L}_\theta(\mathbf{x}) \quad (\text{C21})$$

denote the MLE, which we assume to be the unique maximizer of the log-likelihood function from the assumption (A1) in the main text. Throughout the proof, we denote the true parameter value by  $\theta$  and define the estimation error vector as

$$\Delta^{\text{ML}} := \tilde{\theta}^{\text{ML}} - \theta. \quad (\text{C22})$$

To this end, we apply Taylor's theorem with integral remainder to the empirical score function around the point  $\theta$ . For an arbitrary estimation error vector  $\Delta \in \mathbb{R}^d$ , the score function admits the expansion

$$S_{\hat{\theta}^{\text{ML}}} = S_{\theta+\Delta} = S_{\theta} + \mathbf{H}_{\theta} \Delta + \mathbf{r}_{\theta}(\Delta), \quad (\text{C23})$$

where the remainder term  $\mathbf{r}_{\theta}(\Delta) \in \mathbb{R}^d$  captures the contribution of third-order derivatives and is given explicitly by

$$\mathbf{r}_{\theta}(\Delta) := \int_0^1 (1-t) \mathbf{R}_{\theta+t\Delta}[\Delta, \Delta] dt. \quad (\text{C24})$$

[See Eqs. (A12)–(A14).] We now specialize this expansion to the MLE. Since  $\hat{\theta}^{\text{ML}}$  is assumed to be the unique stationary point of the log-likelihood from the assumption (A1) in the main text, it satisfies the first-order optimality condition

$$S_{\theta+\Delta^{\text{ML}}} = \mathbf{0}. \quad (\text{C25})$$

Consequently, substituting  $\Delta = \Delta^{\text{ML}}$  into the Taylor expansion of the score function yields

$$\mathbf{0} = S_{\theta} + \mathbf{H}_{\theta} \Delta^{\text{ML}} + \mathbf{r}_{\theta}(\Delta^{\text{ML}}). \quad (\text{C26})$$

We next rearrange (C26) in order to isolate the estimation error vector as

$$\Delta^{\text{ML}} = \mathbf{F}_{\theta}^{-1} S_{\theta} + \mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta^{\text{ML}} + \mathbf{F}_{\theta}^{-1} \mathbf{r}_{\theta}(\Delta^{\text{ML}}). \quad (\text{C27})$$

Motivated by this structure, we define a mapping  $\mathcal{F}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  by

$$\mathcal{F}(\Delta) := \mathbf{F}_{\theta}^{-1} S_{\theta} + \mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta + \mathbf{F}_{\theta}^{-1} \mathbf{r}_{\theta}(\Delta). \quad (\text{C28})$$

By construction, the estimation error  $\Delta^{\text{ML}}$  is a fixed point of the map  $\mathcal{F}$ , a fact that will be exploited in subsequent steps of the proof.

### 3. Application of Chebyshev's inequality to bound the norm

For  $\epsilon > 0$ , let us consider the closed cube

$$\mathcal{C}_{\epsilon}^{(\infty)} := \{\Delta \in \mathbb{R}^d: \|\Delta\|_{\infty} \leq \epsilon\}. \quad (\text{C29})$$

Controlling the estimation error  $\|\Delta^{\text{ML}}\|_{\infty}$  therefore reduces to showing that the fixed-point equation Eq. (C27), derived in Appendix C, admits a solution inside  $\mathcal{C}_{\epsilon}^{(\infty)}$ . Equivalently, it suffices to verify that the maximum-likelihood error vector satisfies  $\Delta^{\text{ML}} \in \mathcal{C}_{\epsilon}^{(\infty)}$ . Using the fixed-point representation (C27), this condition can be rewritten as

$$\|\mathcal{F}(\Delta^{\text{ML}})\|_{\infty} \leq \epsilon. \quad (\text{C30})$$

To establish this inequality, we derive a convenient upper bound on the map  $\mathcal{F}$ . Applying the triangle inequality to the definition of  $\mathcal{F}$ , we obtain

$$\begin{aligned} \|\mathcal{F}(\Delta)\|_{\infty} &\leq \|\mathbf{F}_{\theta}^{-1} S_{\theta}\|_{\infty} + \|\mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta\|_{\infty} + \|\mathbf{F}_{\theta}^{-1} \mathbf{r}_{\theta}(\Delta)\|_{\infty}. \end{aligned} \quad (\text{C31})$$

Inequality (C31) serves as the starting point for the sample complexity analysis. In the remainder of this step, we derive uniform bounds for the second and third terms on the right-hand side over  $\mathcal{C}_{\epsilon}^{(\infty)}$ , while the first term will be handled separately.

(1) Upper bound for  $\|\mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta\|_{\infty}$  on a *good event*.

Invoking the inequality  $\|\cdot\|_{\infty} \leq \|\cdot\|_2$ , together with the submultiplicativity of the operator norm, we obtain

$$\begin{aligned} \|\mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta\|_{\infty} &\leq \|\mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta\|_2 \\ &\leq \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} \|\Delta\|_2. \end{aligned} \quad (\text{C32})$$

For any  $\Delta \in \mathcal{C}_{\epsilon}^{(\infty)}$ , the relation between the  $\ell_2$ - and  $\ell_{\infty}$ -norms implies  $\|\Delta\|_2 \leq \sqrt{d}\epsilon$ . Therefore, the only remaining random quantity in Eq. (C32) is the operator norm  $\|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}}$ , which captures the deviation of the empirical Hessian from its expectation. We note that  $\mathbb{E}[\mathbf{H}_{\theta} + \mathbf{F}_{\theta}] = \mathbf{0}$ . To control this deviation, we introduce the following *good event*:

$$\{\|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} < c_H\}, \quad (\text{C33})$$

where  $c_H > 0$  is a small constant to be specified. On the event  $\{\|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} < c_H\}$ , the bound (C32) implies

$$\sup_{\Delta \in \mathcal{C}_{\epsilon}^{(\infty)}} \|\mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta\|_{\infty} \leq \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} c_H \sqrt{d}\epsilon. \quad (\text{C34})$$

It therefore remains to control the probability of the complement event  $\{\|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} < c_H\}^c$ , which corresponds to atypically large fluctuations of the empirical Hessian around its expectation. To this end, we bound the complement event  $\{\|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} < c_H\}^c$  using Markov's inequality. We define the random matrices

$$\mathbf{B}_{\theta}(x_i) := \ell_{\theta}^{(2)}(x_i) + \mathbf{F}_{\theta}, \quad \mathbb{E}[\mathbf{B}_{\theta}(x_i)] = \mathbf{0}, \quad (\text{C35})$$

so that

$$\mathbf{H}_\theta + \mathbf{F}_\theta = \frac{1}{M} \sum_{i=1}^M \mathbf{B}_\theta(x_i). \quad (\text{C36})$$

We assume that the second moment

$$V_H := \mathbb{E}[\text{Tr}(\mathbf{B}_\theta(x_1)^2)] < \infty \quad (\text{C37})$$

is finite. Under this assumption, the second moment of the empirical average admits a simple expression. Indeed, since the measurement outcomes  $\{x_i\}_{i=1}^M$  are independent and satisfy  $\mathbb{E}[\mathbf{B}_\theta(x_i)] = 0$ , all cross terms vanish when taking expectations. As a result, only the diagonal contributions remain, and we obtain

$$\begin{aligned} \mathbb{E}[\text{Tr}((\mathbf{H}_\theta + \mathbf{F}_\theta)^2)] &= \frac{1}{M^2} \sum_{i=1}^M \mathbb{E}[\text{Tr}(\mathbf{B}_\theta(x_i)^2)] \\ &= \frac{1}{M} \mathbb{E}[\text{Tr}(\mathbf{B}_\theta(x_1)^2)] = \frac{V_H}{M}. \end{aligned} \quad (\text{C38})$$

Since  $\mathbf{H}_\theta + \mathbf{F}_\theta$  is symmetric, the inequality  $\|\mathbf{X}\|_{\text{op}}^2 \leq \text{Tr}(\mathbf{X}^2)$  holds. Applying this inequality together with Markov's inequality, we obtain

$$\begin{aligned} \Pr[\|\mathbf{H}_\theta + \mathbf{F}_\theta\|_{\text{op}} \geq c_H] \\ \leq \frac{\mathbb{E}[\text{Tr}((\mathbf{H}_\theta + \mathbf{F}_\theta)^2)]}{c_H^2} = \frac{V_H}{M c_H^2}. \end{aligned} \quad (\text{C39})$$

Consequently, for any  $\delta_H \in (0, 1)$ , choosing

$$c_H := \sqrt{\frac{V_H}{M \delta_H}} \quad (\text{C40})$$

ensures that

$$\Pr[\|\mathbf{H}_\theta + \mathbf{F}_\theta\|_{\text{op}} \geq c_H] \leq \delta_H. \quad (\text{C41})$$

This establishes a high-probability bound on the deviation of the empirical Hessian from its expectation.

(2) Upper bound for  $\|\mathbf{F}_\theta^{-1} \mathbf{r}_\theta(\Delta)\|_\infty$  on a *good event*.

We next control the nonlinear remainder term arising from the third-order derivatives in the Taylor expansion. We begin by reducing the  $\ell_\infty$ -norm to the  $\ell_2$ -norm. Using the inequality  $\|\cdot\|_\infty \leq \|\cdot\|_2$ , together with the submultiplicativity of the operator norm, we obtain

$$\|\mathbf{F}_\theta^{-1} \mathbf{r}_\theta(\Delta)\|_\infty \leq \|\mathbf{F}_\theta^{-1} \mathbf{r}_\theta(\Delta)\|_2 \leq \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \|\mathbf{r}_\theta(\Delta)\|_2. \quad (\text{C42})$$

We now bound the  $\ell_2$ -norm of the remainder term. Recalling its integral representation and applying Minkowski's inequality, we have

$$\begin{aligned} \|\mathbf{r}_\theta(\Delta)\|_2 &= \left\| \int_0^1 (1-t) \mathbf{R}_{\theta+t\Delta}[\Delta, \Delta] dt \right\|_2 \\ &\leq \int_0^1 (1-t) \|\mathbf{R}_{\theta+t\Delta}[\Delta, \Delta]\|_2 dt \\ &\leq \int_0^1 (1-t) \|\mathbf{R}_{\theta+t\Delta}\|_{\text{op}} \|\Delta\|_2^2 dt. \end{aligned} \quad (\text{C43})$$

For the true parameter  $\theta$ , define the nonnegative envelope

$$r(x) := \sup_{\|u\|_2 \leq r_0} \|\ell_{\theta+u}^{(3)}(x)\|_{\text{op}}, \quad (\text{C44})$$

where the dependence of  $r$  on  $\theta$  is suppressed for notational simplicity. By condition (R1), all points  $\theta + u$  appearing in this supremum belong to the ambient parameter domain, and condition (R4) guarantees that  $r(X)$  has a finite second moment.

Since  $0 < \epsilon < \epsilon_0 \leq r_0/\sqrt{d}$ , we have

$$\sup_{\|\Delta\|_2 \leq \sqrt{d}\epsilon} \|\ell_{\theta+\Delta}^{(3)}(x)\|_{\text{op}} \leq r(x). \quad (\text{C45})$$

In particular, this bound applies to  $t\Delta$  for every  $\Delta \in C_\epsilon^{(\infty)}$  and  $t \in [0, 1]$ . We define

$$\mu_R := \mathbb{E}[r(X)] < \infty, \quad V_R := \mathbb{V}[r(X)] < \infty. \quad (\text{C46})$$

By construction,  $\mu_R \leq \bar{\mu}_R$ , and condition (R4) also implies that  $V_R$  is uniformly bounded over  $\theta \in \Theta$ .

Let us further bound  $\|\mathbf{R}_{\theta+t\Delta}\|_{\text{op}}$  using Chebyshev's inequality. Let  $R(\mathbf{x}) := \frac{1}{M} \sum_{i=1}^M r(x_i)$ . Then  $\mathbb{E}[R(\mathbf{x})] = \mu_R$  and  $\mathbb{V}[R(\mathbf{x})] = V_R/M$ . By Chebyshev's inequality, for any  $\delta_R \in (0, 1)$ ,

$$\Pr \left[ |R(\mathbf{x}) - \mu_R| \geq \sqrt{\frac{V_R}{M \delta_R}} \right] \leq \delta_R. \quad (\text{C47})$$

In particular, since

$$\left\{ R(\mathbf{x}) - \mu_R \geq \sqrt{\frac{V_R}{M \delta_R}} \right\} \subseteq \left\{ |R(\mathbf{x}) - \mu_R| \geq \sqrt{\frac{V_R}{M \delta_R}} \right\}, \quad (\text{C48})$$

we obtain the one-sided bound

$$\Pr[R(\mathbf{x}) \geq c_R] \leq \delta_R, \quad (\text{C49})$$

where

$$c_R := \mu_R + \sqrt{\frac{V_R}{M\delta_R}}. \quad (\text{C50})$$

Indeed, for every  $\Delta \in \mathcal{C}_\epsilon^{(\infty)}$  and  $t \in [0, 1]$ ,

$$\|R_{\theta+t\Delta}\|_{\text{op}} \leq \frac{1}{M} \sum_{i=1}^M \|\ell_{\theta+t\Delta}^{(3)}(x_i)\|_{\text{op}} \leq R(\mathbf{x}). \quad (\text{C51})$$

Consequently, on the event  $\{R(\mathbf{x}) \leq c_R\}$ , inequality (C43) yields

$$\begin{aligned} \|\mathbf{r}_\theta(\Delta)\|_2 &\leq \int_0^1 (1-t) \|\mathbf{R}_{\theta+t\Delta}\|_{\text{op}} \|\Delta\|_2^2 dt \\ &\leq \int_0^1 (1-t) c_R \|\Delta\|_2^2 dt \\ &= c_R \|\Delta\|_2^2 \int_0^1 (1-t) dt \\ &= \frac{1}{2} c_R \|\Delta\|_2^2. \end{aligned} \quad (\text{C52})$$

Therefore, we conclude that, on the *good event*  $\{R(\mathbf{x}) \leq c_R\}$ ,

$$\sup_{\Delta \in \mathcal{C}_\epsilon^{(\infty)}} \|\mathbf{F}_\theta^{-1} \mathbf{r}_\theta(\Delta)\|_\infty \leq \frac{1}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} c_R d \epsilon^2. \quad (\text{C53})$$

#### 4. Brouwer fixed-point theorem

We now combine the bounds obtained in Appendix C to conclude the proof. To this end, we introduce the finite number of measurement outcomes margin

$$\tau_- := \epsilon - \|\mathbf{F}_\theta^{-1}\|_{\text{op}} c_H \sqrt{d} \epsilon - \frac{1}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} c_R d \epsilon^2, \quad (\text{C54})$$

Taking complements and probabilities yields

$$\begin{aligned} \Pr[\|\Delta^{\text{ML}}\|_\infty \geq \epsilon] &\leq \Pr[\mathcal{E}^c] = \Pr[\{\|\mathbf{F}_\theta^{-1} \mathbf{S}_\theta\|_\infty \geq \tau_-\} \cup \{\|\mathbf{H}_\theta + \mathbf{F}_\theta\|_{\text{op}} \geq c_H\} \cup \{R \geq c_R\}] \\ &\leq \Pr[\{\|\mathbf{F}_\theta^{-1} \mathbf{S}_\theta\|_\infty \geq \tau_-\}] + \delta_H + \delta_R. \end{aligned} \quad (\text{C61})$$

Here, by applying the union bound, we further obtain

$$\Pr[\{\|\mathbf{F}_\theta^{-1} \mathbf{S}_\theta\|_\infty \geq \tau_-\}] = \Pr\left[\bigcup_{a=1}^d |e_a^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| \geq \tau_-\right] \leq \sum_{a=1}^d \Pr[|e_a^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| \geq \tau_-]. \quad (\text{C62})$$

which quantifies the residual budget available for the score term after accounting for the linear and nonlinear correction terms. We next define the event

$$\begin{aligned} \mathcal{E} &:= \{\|\mathbf{F}_\theta^{-1} \mathbf{S}_\theta\|_\infty \leq \tau_-\} \cap \{\|\mathbf{H}_\theta + \mathbf{F}_\theta\|_{\text{op}} \leq c_H\} \\ &\quad \cap \{R(\mathbf{x}) \leq c_R\}. \end{aligned} \quad (\text{C55})$$

On this event, the bounds established in Eqs. (C34) and (C53) imply that

$$\sup_{\Delta \in \mathcal{C}_\epsilon^{(\infty)}} \|\mathcal{F}(\Delta)\|_\infty \leq \epsilon. \quad (\text{C56})$$

Equivalently, on the event  $\mathcal{E}$ , the map  $\mathcal{F}$  satisfies

$$\mathcal{F}(\mathcal{C}_\epsilon^{(\infty)}) \subseteq \mathcal{C}_\epsilon^{(\infty)}, \quad (\text{C57})$$

that is,  $\mathcal{F}$  maps  $\mathcal{C}_\epsilon^{(\infty)}$  into itself.

We now verify the conditions required to apply Brouwer's fixed-point theorem. The set  $\mathcal{C}_\epsilon^{(\infty)}$  is nonempty, closed, and bounded in  $\mathbb{R}^d$ ; hence it is compact by the Heine-Borel theorem. Moreover, it is convex, since it is an intersection of  $2d < \infty$  closed half-spaces:

$$\mathcal{C}_\epsilon^{(\infty)} = \bigcap_{a=1}^d \{\Delta: \Delta_a \leq \epsilon\} \cap \{\Delta: -\Delta_a \leq \epsilon\}. \quad (\text{C58})$$

In addition, the map  $\mathcal{F}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  is continuous by construction. Consequently, on the event  $\mathcal{E}$ , Brouwer's fixed-point theorem guarantees the existence of a fixed point

$$\exists \Delta^* \in \mathcal{C}_\epsilon^{(\infty)} \text{ such that } \Delta^* = \mathcal{F}(\Delta^*). \quad (\text{C59})$$

Finally, we identify this fixed point. By assumption, the log-likelihood function admits a unique stationary point at  $\tilde{\theta}^{\text{ML}}$ . Since fixed points of  $\mathcal{F}$  correspond precisely to stationary points of  $\mathbf{S}_\theta$ , the solution  $\Delta^*$  in Eq. (C59) must coincide with the MLE error,  $\Delta^* = \Delta^{\text{ML}}$ . As a result, we conclude that

$$\mathcal{E} \subseteq \{\|\Delta^{\text{ML}}\|_\infty \leq \epsilon\}. \quad (\text{C60})$$

To this end, let us set  $\delta_H = \delta_R = \delta/4$ .  $\tau_-$  is then given by

$$\tau_- = \tau_{0-} - DM^{-\frac{1}{2}}, \quad (\text{C63})$$

where

$$\tau_{0-} := \left(1 - \epsilon \frac{d\mu_R}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}}\right) \epsilon, \quad (\text{C64})$$

$$D := \left(\|\mathbf{F}_\theta^{-1}\|_{\text{op}} \sqrt{\frac{4V_H}{\delta}} \sqrt{d} + \frac{1}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \sqrt{\frac{4V_R}{\delta}} d \epsilon\right) \epsilon. \quad (\text{C65})$$

From Eqs. (C61) and (C62), we obtain

$$\Pr[\|\Delta^{\text{ML}}\|_\infty \geq \epsilon] \leq \sum_{a=1}^d \Pr[|\mathbf{e}_a^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| \geq \tau_-] + \frac{\delta}{2}. \quad (\text{C66})$$

### 5. Application of the Berry-Esseen theorem

To refine the upper bound in Eq. (C66), we invoke the Berry-Esseen theorem, which quantifies the rate of convergence in the central limit theorem. Specifically, it provides a uniform bound of order  $O(M^{-1/2})$  on the deviation between the cumulative distribution function of a normalized sum of independent random variables and that of the standard normal distribution.

By definition in Eq. (A12),  $\mathbf{S}_\theta$  is a sum of independent random variables. In addition, we have

$$\mathbb{E}[\mathbf{e}_a^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta] = 0, \quad \mathbb{V}[\mathbf{e}_a^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta] = \frac{[\mathbf{F}_\theta^{-1}]_{aa}}{M} =: \frac{\sigma_{\theta,a}^2}{M}, \quad (\text{C67})$$

where we have defined  $\sigma_{\theta,a} := \sqrt{[\mathbf{F}_\theta^{-1}]_{aa}}$ . Let us assume that the third absolute moment of the projected score is finite:

$$\rho := \sup_{\theta \in \Theta} \max_{a \in [d]} \mathbb{E}[|\mathbf{e}_a^T \mathbf{F}_\theta^{-1} \mathcal{L}_\theta^{(1)}(x_i)|^3] < \infty. \quad (\text{C68})$$

Consequently, for any  $\tau_- > 0$ , we obtain

$$\begin{aligned} & \Pr[|\mathbf{e}_a^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| \geq \tau_-] \\ &= \Pr\left[\left|\frac{\sqrt{M}}{\sigma_{\theta,a}} \mathbf{e}_a^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta\right| \geq \frac{\sqrt{M}\tau_-}{\sigma_{\theta,a}}\right] \\ &\leq 2 \int_{\frac{\sqrt{M}\tau_-}{\sigma_{\theta,a}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt + \frac{2C\rho}{\sigma_{\theta,a}^3 \sqrt{M}}. \end{aligned} \quad (\text{C69})$$

See Appendix B.

The above small-error restriction guarantees that  $\tau_{0-}$  is uniformly positive. Indeed, since  $\mu_R \leq \bar{\mu}_R$  and  $\|\mathbf{F}_\theta^{-1}\|_{\text{op}} \leq K_F$ , we have

$$\begin{aligned} \tau_{0-} &= \epsilon \left(1 - \frac{\epsilon d \mu_R}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}}\right) \\ &\geq \epsilon \left(1 - \frac{\epsilon d \bar{\mu}_R K_F}{2}\right) > \frac{\epsilon}{2} > 0, \end{aligned} \quad (\text{C70})$$

uniformly over  $\theta \in \Theta$ . Consequently,  $\tau_- = \tau_{0-} - DM^{-1/2} > 0$  whenever

$$M > \left(\frac{D}{\tau_{0-}}\right)^2. \quad (\text{C71})$$

We then further upper bound the integral term in Eq. (C69) by maximizing  $\sigma_{\theta,a}$  over  $\theta \in \Theta$  and  $a \in [d]$ , and then applying the Mills ratio inequality in Appendix B:

$$\begin{aligned} & \int_{\frac{\sqrt{M}\tau_-}{\sigma_{\theta,a}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &\leq \int_{\frac{\sqrt{M}\tau_-}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &\leq \frac{1}{\sqrt{2\pi}} \frac{\sigma}{\sqrt{M}\tau_-} e^{-\frac{M\tau_-^2}{2\sigma^2}}, \end{aligned} \quad (\text{C72})$$

where

$$\sigma := \sup_{\theta \in \Theta} \max_{a \in [d]} \sigma_{\theta,a} = \sqrt{\sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_\theta^{-1}]_{aa}}. \quad (\text{C73})$$

Combining Eqs. (C61), (C62), and (C72), we obtain

$$\begin{aligned} & \Pr[\|\Delta^{\text{ML}}\|_\infty \geq \epsilon] \\ &\leq \frac{2d}{\sqrt{2\pi}} \frac{\sigma}{\sqrt{M}\tau_-} e^{-\frac{M\tau_-^2}{2\sigma^2}} + \sum_{a=1}^d \frac{2C\rho}{\sigma_{\theta,a}^3 \sqrt{M}} + \frac{\delta}{2}. \end{aligned} \quad (\text{C74})$$

Therefore, it is sufficient to choose  $M$  such that

$$\frac{2d}{\sqrt{2\pi}} \frac{\sigma}{\sqrt{M}\tau_-} e^{-\frac{M\tau_-^2}{2\sigma^2}} + \sum_{a=1}^d \frac{2C\rho}{\sigma_{\theta,a}^3 \sqrt{M}} + \frac{\delta}{2} \leq \delta. \quad (\text{C75})$$

Equation (C75) then reduces to

$$\frac{2}{\sqrt{2\pi}} \frac{\sigma}{\sqrt{M}\tau_-} e^{-\frac{M\tau_-^2}{2\sigma^2}} \leq \delta', \quad (\text{C76})$$

where

$$\delta' := \frac{\delta}{2d} - \frac{\eta}{\sqrt{M}}, \quad \eta := \frac{1}{d} \sum_{a=1}^d \frac{2C\rho}{\sigma_{\theta,a}^3}. \quad (\text{C77})$$

Equation (C76) admits a solution only if  $\delta' \geq 0$ , equivalently

$$M \geq \left( \frac{2d\eta}{\delta} \right)^2. \quad (\text{C78})$$

Using the Lambert  $W$  function, Eq. (C76) can be re-expressed as

$$M \geq \sigma^2 \tau_-^{-2} W_0(2\delta'^{-2} \pi^{-1}), \quad (\text{C79})$$

where  $W_0$  denotes the principal branch of the Lambert  $W$  function. See Appendix B for details of  $W_0$ . Note that  $\delta'^{-2} \geq 4\delta^{-2}d^2$ . Since  $W_0(x)$  is increasing and concave for  $x > 0$  (see Appendix B), it follows that

$$W_0(\delta'^{-2}) \leq \frac{\delta^2}{4d^2\delta'^2} W_0(4\delta^{-2}d^2). \quad (\text{C80})$$

Consequently,

$$M \geq \sigma^2 \tau_-^{-2} \frac{\delta^2}{4d^2\delta'^2} W_0(8\pi^{-1}\delta^{-2}d^2) \quad (\text{C81})$$

guarantees

$$\Pr[\|\Delta^{\text{ML}}\|_\infty \geq \epsilon] \leq \delta. \quad (\text{C82})$$

Here, Eq. (C81) can be reduced to the quadratic inequality

$$\tau_{0-}M - \left( \frac{2d\eta}{\delta} \tau_{0-} + D + \sigma\sqrt{W_0} \right) \sqrt{M} + \frac{2d\eta}{\delta} D \geq 0. \quad (\text{C83})$$

Equation (C83) holds uniformly for  $M$  such that

$$M \geq \left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} + \sqrt{\left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} \right)^2 - \frac{2d\eta}{\delta} \frac{D}{\tau_{0-}}} \right)^2. \quad (\text{C84})$$

Considering Eqs. (C71), (C78), and (C84), we obtain

$$M \geq \max \left\{ \left( \frac{D}{\tau_{0-}} \right)^2, \left( \frac{2d\eta}{\delta} \right)^2, \left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} + \sqrt{\left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} \right)^2 - \frac{2d\eta}{\delta} \frac{D}{\tau_{0-}}} \right)^2 \right\}. \quad (\text{C85})$$

Since the upper bound must hold for all  $\theta \in \Theta$ , it suffices to take

$$M \geq M_U := \sup_{\theta \in \Theta} \max \left\{ \left( \frac{D}{\tau_{0-}} \right)^2, \left( \frac{2d\eta}{\delta} \right)^2, \left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} + \sqrt{\left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} \right)^2 - \frac{2d\eta}{\delta} \frac{D}{\tau_{0-}}} \right)^2 \right\}. \quad (\text{C86})$$

Since  $M_U$  is sufficient to guarantee the desired bound uniformly over  $\Theta$ , we have  $M_0 \leq M_U$ , which completes the proof.

## 6. Asymptotic limit $\epsilon \rightarrow 0$

We now derive the simplified small-error form stated in Theorem 1 in the main text. Fix  $0 < \delta \leq 1$  and  $d < \infty$ .

From the definitions of  $\tau_{0-}$  and  $D$ , for each  $\theta \in \Theta$  we have, as  $\epsilon \rightarrow 0$ ,

$$\tau_{0-} = \epsilon \left( 1 - \epsilon \frac{d\mu_R}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \right) = \epsilon(1 + O(\epsilon)), \quad (\text{C87})$$

$$D = \left( \sqrt{\frac{4V_H}{\delta}} \sqrt{d} + \frac{1}{2} \sqrt{\frac{4V_R}{\delta}} d\epsilon \right) \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \epsilon = O(\epsilon). \quad (\text{C88})$$

Here the  $O(\epsilon)$  terms are uniform over  $\theta \in \Theta$  under the boundedness assumptions of the theorem. Consequently,

$$\left( \frac{D}{\tau_{0-}} \right)^2 = O(1), \quad \left( \frac{2d\eta}{\delta} \right)^2 = O(1). \quad (\text{C89})$$

These two terms are therefore subleading compared with the  $\epsilon^{-2}$  term below.

It remains to expand the third term in the maximum. Define

$$Q_\epsilon(\boldsymbol{\theta}) := \frac{d\eta}{\delta} + \frac{D}{2\tau_{0,-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0,-}}, \quad R_\epsilon(\boldsymbol{\theta}) := \frac{2d\eta}{\delta} \frac{D}{\tau_{0,-}}. \quad (\text{C90})$$

we obtain

$$M_0(\epsilon, \delta, d) \leq W_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\boldsymbol{\theta} \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\boldsymbol{\theta}}^{-1}]_{aa} \epsilon^{-2} (1 + O(\epsilon)). \quad (\text{C96})$$

Then

$$y^* = (Q_\epsilon(\boldsymbol{\theta}) + \sqrt{Q_\epsilon(\boldsymbol{\theta})^2 - R_\epsilon(\boldsymbol{\theta})})^2. \quad (\text{C91})$$

Equivalently,

Since  $D/\tau_{0,-} = O(1)$  and  $\eta = O(1)$ , we have  $R_\epsilon(\boldsymbol{\theta}) = O(1)$ . On the other hand,

$$Q_\epsilon(\boldsymbol{\theta}) = \frac{\sigma\sqrt{W_0}}{2\epsilon} (1 + O(\epsilon)). \quad (\text{C92})$$

$$\lim_{\epsilon \rightarrow 0} \sup_{\boldsymbol{\theta} \in \Theta} \epsilon^2 M_0(\epsilon, \delta, d) \leq W_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\boldsymbol{\theta} \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\boldsymbol{\theta}}^{-1}]_{aa}. \quad (\text{C97})$$

Therefore,

This is precisely the simplified small-error upper bound stated in Theorem 1 in the main text.

$$\sqrt{Q_\epsilon(\boldsymbol{\theta})^2 - R_\epsilon(\boldsymbol{\theta})} = Q_\epsilon(\boldsymbol{\theta}) (1 + O(\epsilon^2)), \quad (\text{C93})$$

and hence

#### APPENDIX D: PROOF OF LOWER BOUND FOR $\ell_\infty$

$$y^* = \sigma^2 W_0 \epsilon^{-2} (1 + O(\epsilon)). \quad (\text{C94})$$

Assume that conditions (A1) and (A2) and the standard regularity conditions (R1)–(R5) stated in Appendix 3 B 4 hold. Then the following theorem applies.

Using

$$\sigma^2 = \sup_{\boldsymbol{\theta} \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\boldsymbol{\theta}}^{-1}]_{aa}, \quad (\text{C95})$$

*Theorem 2.* For  $0 < \epsilon$ ,  $0 < \delta < 1/\sqrt{8\pi\epsilon}$  and  $d < \infty$ , let  $M_0(\epsilon, \delta, d)$  denote the minimal number of repetitions such that

$$\forall M \geq M_0(\epsilon, \delta, d): \Pr[\|\tilde{\boldsymbol{\theta}}^{\text{ML}} - \boldsymbol{\theta}\|_\infty \leq \epsilon] \geq 1 - \delta \quad \text{for all } \boldsymbol{\theta} \in \Theta. \quad (\text{D1})$$

$M_0$  is then lower bounded as

$$M_0 \geq \max \left\{ M_2(\boldsymbol{\theta}, a), \left( \frac{\max\{\sigma_{\boldsymbol{\theta},a} - D, 0\}}{\tau_{0+}} \right)^2, \left( \frac{\eta}{\frac{1}{\sqrt{2\pi\epsilon}} - 2\delta} \right)^2 \right\}, \quad (\text{D2})$$

for any  $\boldsymbol{\theta} \in \Theta$  and  $a \in [d]$ . Here

$$M_2(\boldsymbol{\theta}, a) := \begin{cases} \left( \frac{B_{\boldsymbol{\theta},a} + \sqrt{\Delta_{\boldsymbol{\theta},a}}}{2} \right)^2, & B_{\boldsymbol{\theta},a} > 0 \text{ and } \Delta_{\boldsymbol{\theta},a} > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (\text{D3})$$

$$\Delta_{\boldsymbol{\theta},a} := B_{\boldsymbol{\theta},a}^2 - \frac{2\eta D}{\delta\tau_{0+}}, \quad (\text{D6})$$

and

$$\tau_{0,+} := \left( 1 + \epsilon \frac{d\mu_R}{2} \|\mathbf{F}_{\boldsymbol{\theta}}^{-1}\|_{\text{op}} \right) \epsilon, \quad (\text{D4})$$

$$D := \left( \sqrt{\frac{2V_H}{\delta}} \sqrt{d} + \frac{1}{2} \sqrt{\frac{2V_R}{\delta}} d\epsilon \right) \|\mathbf{F}_{\boldsymbol{\theta}}^{-1}\|_{\text{op}} \epsilon, \quad (\text{D7})$$

$$B_{\boldsymbol{\theta},a} := -\frac{D}{\tau_{0+}} - \frac{\eta}{2\delta} + \frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\boldsymbol{\theta},a}}}{\tau_{0+}}, \quad (\text{D5})$$

$$\eta := \frac{2C\rho}{\sigma_{\boldsymbol{\theta},a}^3}, \quad (\text{D8})$$

$$\rho := \sup_{\boldsymbol{\theta} \in \Theta} \max_{a \in [d]} \mathbb{E}[|\mathbf{e}_a^T \mathbf{F}_{\boldsymbol{\theta}}^{-1} \ell_{\boldsymbol{\theta}}^{(1)}(x)|^3], \quad (\text{D9})$$

$$\sigma_{\theta,a} := \sqrt{[\mathbf{F}_{\theta}^{-1}]_{aa}}. \quad (\text{D10})$$

In the small-error limit  $\epsilon \rightarrow 0$ , this lower bound reduces to

$$M_0(\epsilon, \delta, d) \gtrsim W_0(\delta^{-2}/8\pi) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\theta}^{-1}]_{aa} \epsilon^{-2}, \quad (\text{D11})$$

where the notation  $A_{\epsilon} \gtrsim B_{\epsilon}$  denotes

$$\liminf_{\epsilon \rightarrow 0} \frac{A_{\epsilon}}{B_{\epsilon}} \geq 1, \quad (\text{D12})$$

i.e.,  $A_{\epsilon} \geq (1 + o(1))B_{\epsilon}$  as  $\epsilon \rightarrow 0$  with fixed  $\delta$  and  $d$ .

### 1. Proof sketch

Our goal is to characterize the minimal number of repetitions  $M_0(\epsilon, \delta, d)$  required to guarantee that, for a prescribed accuracy  $\epsilon > 0$  and confidence level  $1 - \delta$ , with  $0 < \delta < 1/\sqrt{8\pi e}$ , the MLE  $\tilde{\theta}^{\text{ML}}$  satisfies

$$\forall M \geq M_0(\epsilon, \delta, d): \Pr[\|\tilde{\theta}^{\text{ML}} - \theta\|_{\infty} \leq \epsilon] \geq 1 - \delta. \quad (\text{D13})$$

In this section, we derive a lower bound on the required number of repetitions. The argument proceeds by comparing the accuracy event with a larger auxiliary event. Let us again consider the accuracy event defined in Eq. (C15):

$$A(\epsilon) := \{\|\tilde{\theta}^{\text{ML}} - \theta\|_{\infty} \leq \epsilon\}. \quad (\text{D14})$$

We first introduce an auxiliary event  $L(\epsilon)$  satisfying

$$A(\epsilon) \subseteq L(\epsilon), \quad (\text{D15})$$

which immediately yields

$$\Pr[L(\epsilon)] \geq \Pr[A(\epsilon)]. \quad (\text{D16})$$

Next, we characterize the minimal number of repetitions  $M_L(\epsilon, \delta, d)$  such that

$$\forall M \geq M_L(\epsilon, \delta, d): \Pr[L(\epsilon)] \geq 1 - \delta. \quad (\text{D17})$$

Because  $A(\epsilon) \subseteq L(\epsilon)$ , any repetition threshold  $M$  satisfying  $\Pr[A(\epsilon)] \geq 1 - \delta$  necessarily also satisfies  $\Pr[L(\epsilon)] \geq 1 - \delta$ . Accordingly, the corresponding minimal thresholds obey

$$M_L(\epsilon, \delta, d) \leq M_0(\epsilon, \delta, d). \quad (\text{D18})$$

Therefore, a necessary condition for guaranteeing  $\Pr[A(\epsilon)] \geq 1 - \delta$  is

$$M \geq M_L(\epsilon, \delta, d). \quad (\text{D19})$$

### 2. Taylor expansion of the score function

We begin with the expansion

$$\Delta^{\text{ML}} = \mathbf{F}_{\theta}^{-1} S_{\theta} + \mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta^{\text{ML}} + \mathbf{F}_{\theta}^{-1} \mathbf{r}_{\theta} (\Delta^{\text{ML}}). \quad (\text{D20})$$

Taking the inner product with the basis vector  $\mathbf{e}_a$ , we obtain

$$\begin{aligned} \mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} S_{\theta} &= \Delta_a^{\text{ML}} - \mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta^{\text{ML}} - \mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} \mathbf{r}_{\theta} (\Delta^{\text{ML}}). \end{aligned} \quad (\text{D21})$$

Applying the triangle inequality and standard operator-norm bounds yields

$$\begin{aligned} |\mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} S_{\theta}| &\leq |\Delta_a^{\text{ML}}| + |\mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} (\mathbf{H}_{\theta} + \mathbf{F}_{\theta}) \Delta^{\text{ML}}| + |\mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} \mathbf{r}_{\theta} (\Delta^{\text{ML}})| \\ &\leq |\Delta_a^{\text{ML}}| + \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} \|\Delta^{\text{ML}}\|_2 + \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \|\mathbf{r}_{\theta}(\Delta^{\text{ML}})\|_2. \end{aligned} \quad (\text{D22})$$

We define the *good* event

$$\mathcal{G} := \{\|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} \leq c_H\} \cap \{R(\mathbf{x}) \leq c_R\}. \quad (\text{D23})$$

On the accuracy event  $\{\|\Delta^{\text{ML}}\|_{\infty} \leq \epsilon\}$ , we have  $|\Delta_a^{\text{ML}}| \leq \epsilon$  and  $\|\Delta^{\text{ML}}\|_2 \leq \sqrt{d}\epsilon$ . Combining these bounds with Eq. (D22), and then decomposing the probability according to  $\mathcal{G}$ , gives

$$\begin{aligned} \Pr[\{\|\Delta^{\text{ML}}\|_{\infty} \leq \epsilon\}] &\leq \Pr[\{\|\Delta^{\text{ML}}\|_{\infty} \leq \epsilon\} \cap \mathcal{G}] + \Pr[\{\|\Delta^{\text{ML}}\|_{\infty} \leq \epsilon\} \cap \mathcal{G}^c] \\ &\leq \Pr[\{\|\Delta^{\text{ML}}\|_{\infty} \leq \epsilon\} \cap \mathcal{G}] + \Pr[\mathcal{G}^c] \\ &\leq \Pr[\{\|\Delta^{\text{ML}}\|_{\infty} \leq \epsilon\} \cap \mathcal{G}] + \delta_H + \delta_R. \end{aligned} \quad (\text{D24})$$

On the event  $\{\|\Delta^{\text{ML}}\|_{\infty} \leq \epsilon\} \cap \mathcal{G}$ , Eq. (D22) implies

$$\begin{aligned} |\mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} S_{\theta}| &\leq \epsilon + \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} c_H \sqrt{d} \epsilon + \frac{1}{2} \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} c_R d \epsilon^2 =: \tau_+. \end{aligned} \quad (\text{D25})$$

We now choose  $\delta_H = \delta_R = \delta/2$ . With this choice, the threshold  $\tau_+$  can be expressed as

$$\tau_+ = \tau_{0+} + DM^{-\frac{1}{2}}, \quad (\text{D26})$$

where

$$\tau_{0+} := \left(1 + \epsilon \frac{d\mu_R}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}}\right) \epsilon, \quad (\text{D27})$$

$$D := \left( \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \sqrt{\frac{2V_H}{\delta}} \sqrt{d} + \frac{1}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \sqrt{\frac{2V_R}{\delta}} d \epsilon \right) \epsilon. \quad (\text{D28})$$

Combining Eqs. (D24) and (D25), we obtain

$$\Pr[\{\|\Delta^{\text{ML}}\|_\infty \leq \epsilon\}] \leq \Pr[|e_d^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| \leq \tau_+] + \delta. \quad (\text{D29})$$

### 3. Application of the Berry-Esseen theorem

Applying the Berry-Esseen theorem from Appendix B, we obtain

$$\begin{aligned} \Pr \left[ \left| \frac{\sqrt{M} (e_d^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta)}{\sigma_{\theta,a}} \right| \leq x \right] \\ \leq 1 - 2 \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt + \frac{2C\rho}{\sigma_{\theta,a}^3 \sqrt{M}}, \end{aligned} \quad (\text{D30})$$

where  $x := \sqrt{M} \tau_+ / \sigma_{\theta,a}$ . We emphasize that, although  $x$  is introduced for notational convenience, it depends on  $M$  through  $\tau_+$ .

To lower bound the Gaussian tail integral, we use the Mills ratio inequality stated in Appendix B. For all  $x > 0$ ,

$$\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt > \frac{x}{x^2 + 1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \quad (\text{D31})$$

We next introduce the auxiliary function

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x e^{-x^2/2}, & 0 \leq x < 1, \\ \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2}, & x \geq 1. \end{cases} \quad (\text{D32})$$

By construction, this function satisfies, for all  $x \geq 0$ ,

$$f(x) \leq \frac{2}{\sqrt{2\pi}} \frac{x}{x^2 + 1} \exp\left(-\frac{x^2}{2}\right). \quad (\text{D33})$$

Combining Eqs. (D30)–(D33), we obtain

$$\Pr \left[ \left| \frac{\sqrt{M} (e_d^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta)}{\sigma_{\theta,a}} \right| \leq x \right] \leq 1 - f(x) + \frac{2C\rho}{\sigma_{\theta,a}^3 \sqrt{M}}. \quad (\text{D34})$$

Applying Eqs. (D34)–Eq. (D29), we obtain

$$\Pr[\{\|\Delta^{\text{ML}}\|_\infty \leq \epsilon\}] \leq 1 - f(x) + \frac{2C\rho}{\sigma_{\theta,a}^3 \sqrt{M}} + \delta. \quad (\text{D35})$$

Therefore, for the guarantee

$$\Pr[\{\|\Delta^{\text{ML}}\|_\infty \leq \epsilon\}] \geq 1 - \delta \quad (\text{D36})$$

to hold, Eq. (D35) necessarily requires

$$f(x) \leq 2\delta + \frac{\eta}{\sqrt{M}} =: \delta', \quad \eta := \frac{2C\rho}{\sigma_{\theta,a}^3}. \quad (\text{D37})$$

Following the argument in Appendix D, let  $M_1(\epsilon, \delta, d)$  denote the minimal number of repetitions such that

$$\forall M \geq M_1(\epsilon, \delta, d): f(x) \leq \delta'. \quad (\text{D38})$$

Because Eq. (D38) is a necessary condition for the desired success probability, the corresponding minimal thresholds satisfy

$$M_1(\epsilon, \delta, d) \leq M_0(\epsilon, \delta, d). \quad (\text{D39})$$

We now determine the relevant branch of  $f(x)$ . The function  $f(x)$  is increasing on  $0 \leq x \leq 1$ , decreasing on  $x \geq 1$ , and attains its maximum at

$$f(1) = \frac{1}{\sqrt{2\pi}e}. \quad (\text{D40})$$

Under the assumption

$$0 < \delta < \frac{1}{\sqrt{8\pi}e}, \quad (\text{D41})$$

we have  $2\delta < 1/\sqrt{2\pi}e$ . Therefore,  $\delta' \leq f(1)$  is guaranteed whenever

$$\sqrt{M} \geq \frac{\eta}{\frac{1}{\sqrt{2\pi}e} - 2\delta}. \quad (\text{D42})$$

We first show that the branch  $0 \leq x \leq 1$  cannot determine the eventual threshold  $M_1(\epsilon, \delta, d)$  in Eq. (D38). Importantly, our goal is not to identify an arbitrary solution of  $f(x) = \delta'$ . Rather, we seek the final crossing point beyond which the inequality

$$f(x) \leq \delta' \quad (\text{D43})$$

remains satisfied for all larger  $M$ . On this branch,

$$f(x) = \frac{1}{\sqrt{2\pi}} x e^{-x^2/2}, \quad (\text{D44})$$

so that

$$f'(x) = \frac{1}{\sqrt{2\pi}}(1-x^2)e^{-x^2/2} \geq 0, \quad 0 \leq x \leq 1. \quad (\text{D45})$$

Meanwhile,

$$\delta' = 2\delta + \frac{\eta}{\sqrt{M}} \quad (\text{D46})$$

is monotonically decreasing in  $M$ .

Using

$$\tau_+ = \tau_{0+} + \frac{D}{\sqrt{M}}, \quad x = \frac{\sqrt{M}\tau_+}{\sigma_{\theta,a}} = \frac{\tau_{0+}\sqrt{M} + D}{\sigma_{\theta,a}}, \quad (\text{D47})$$

we further obtain

$$\frac{dx}{dM} = \frac{\tau_{0+}}{2\sigma_{\theta,a}\sqrt{M}} > 0. \quad (\text{D48})$$

Therefore, as long as  $0 \leq x \leq 1$ ,

$$\frac{d}{dM}[f(x) - \delta'] = f'(x) \frac{dx}{dM} + \frac{\eta}{2M^{3/2}} > 0. \quad (\text{D49})$$

Hence, throughout the branch  $0 \leq x \leq 1$ , the quantity  $f(x) - \delta'$  increases with  $M$ . Consequently, if the equality  $f(x) = \delta'$  is attained within this branch, then for sufficiently small increases in  $M$  while remaining in the same branch, one necessarily has  $f(x) > \delta'$ . Such a crossing therefore cannot correspond to the threshold beyond which  $f(x) \leq \delta'$  holds for all larger  $M$ . Instead, it can only characterize the boundary of a small- $M$  regime.

Hence, the eventual threshold  $M_1(\epsilon, \delta, d)$  must be determined by the decreasing branch  $x \geq 1$ . Using Eq. (D47), the condition  $x \geq 1$  is equivalent to

$$x \geq 1 \iff \tau_{0+}\sqrt{M} + D \geq \sigma_{\theta,a}. \quad (\text{D50})$$

Rearranging this condition yields

$$M \geq \left( \frac{\max\{\sigma_{\theta,a} - D, 0\}}{\tau_{0+}} \right)^2. \quad (\text{D51})$$

In particular, when  $D \geq \sigma_{\theta,a}$ , the inequality  $x \geq 1$  holds for all  $M \geq 0$ .

#### 4. Lambert W-function

On the branch  $x \geq 1$ , the necessary condition  $f(x) \leq \delta'$  takes the form

$$\frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2} \leq \delta' \iff x^2 e^{x^2} \geq \frac{1}{2\pi\delta'^2}. \quad (\text{D52})$$

Applying the principal branch of the Lambert  $W$ -function to Eq. (D52), we obtain

$$x^2 \geq W_0(\delta'^{-2}/2\pi). \quad (\text{D53})$$

Using

$$x^2 = \frac{M\tau_+^2}{\sigma_{\theta,a}^2}, \quad (\text{D54})$$

this yields the implicit necessary condition

$$M \geq W_0(\delta'^{-2}/2\pi)\tau_+^{-2}\sigma_{\theta,a}^2. \quad (\text{D55})$$

The above condition is implicit because both  $\delta'$  and  $\tau_+$  depend on  $M$ . To obtain a tractable explicit lower bound, we now use the concavity of the Lambert  $W_0$ -function.

#### 5. Further lower bound using concavity of Lambert $W$ -function

For  $z \geq 0$ , the principal branch  $W_0(z)$  is concave and satisfies  $W_0(0) = 0$ . Since

$$\delta' = 2\delta + \frac{\eta}{\sqrt{M}} \geq 2\delta, \quad (\text{D56})$$

we have

$$W_0(\delta'^{-2}/2\pi) = W_0\left[\frac{4\delta^2}{\delta'^2} \frac{\delta^{-2}}{8\pi}\right] \geq \frac{4\delta^2}{\delta'^2} W_0(\delta^{-2}/8\pi). \quad (\text{D57})$$

Consequently, Eq. (D55) implies the weaker necessary condition

$$M \geq \frac{4\delta^2}{\delta'^2} W_0(\delta^{-2}/8\pi)\tau_+^{-2}\sigma_{\theta,a}^2. \quad (\text{D58})$$

Since all quantities are nonnegative, Eq. (D58) is equivalent to

$$\sqrt{M} \delta' \tau_+ \geq 2\delta \sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}^2}. \quad (\text{D59})$$

Substituting

$$\delta' = 2\delta + \frac{\eta}{\sqrt{M}}, \quad \tau_+ = \tau_{0+} + \frac{D}{\sqrt{M}}, \quad (\text{D60})$$

and multiplying both sides by  $\sqrt{M}$ , we obtain

$$(2\delta\sqrt{M} + \eta)(\tau_{0+}\sqrt{M} + D) \geq 2\delta\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}}\sqrt{M}. \quad (\text{D61})$$

Dividing Eq. (D61) by  $2\delta\tau_{0+} > 0$ , we obtain

$$M + \left( \frac{D}{\tau_{0+}} + \frac{\eta}{2\delta} - \frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}}}{\tau_{0+}} \right) \sqrt{M} + \frac{\eta D}{2\delta\tau_{0+}} \geq 0. \quad (\text{D62})$$

Let  $y = \sqrt{M}$ , and define

$$B := -\frac{D}{\tau_{0+}} - \frac{\eta}{2\delta} + \frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}}}{\tau_{0+}}, \quad \Delta := B^2 - \frac{2\eta D}{\delta\tau_{0+}}. \quad (\text{D63})$$

Equation (D62) is then equivalently expressed as

$$y^2 - By + \frac{\eta D}{2\delta\tau_{0+}} \geq 0. \quad (\text{D64})$$

If  $B \leq 0$  or  $\Delta \leq 0$ , this quadratic inequality does not generate any nontrivial forbidden interval for  $y \geq 0$ . If  $B > 0$  and  $\Delta > 0$ , then the quadratic is negative precisely on the interval

$$\frac{B - \sqrt{\Delta}}{2} < y < \frac{B + \sqrt{\Delta}}{2}. \quad (\text{D65})$$

Therefore, the necessary condition is violated throughout this interval. Accordingly, the relevant eventual lower-bound threshold is given by the square of the larger endpoint:

$$M_2 := \begin{cases} \left( \frac{B + \sqrt{\Delta}}{2} \right)^2, & B > 0 \text{ and } \Delta > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (\text{D66})$$

Combining with Eqs. (D51) and (D42), we obtain

$$M_L := \max \left\{ M_2, \left( \frac{\max\{\sigma_{\theta,a} - D, 0\}}{\tau_{0+}} \right)^2, \left( \frac{\eta}{\frac{1}{\sqrt{2\pi e}} - 2\delta} \right)^2 \right\}. \quad (\text{D67})$$

Hence, the sample-complexity threshold necessarily satisfies

$$M_0(\epsilon, \delta, d) \geq M_L. \quad (\text{D68})$$

This completes the derivation of the lower bound for a finite number of measurement outcomes.

## 6. Asymptotic limit $\epsilon \rightarrow 0$

We now derive the simplified small-error form stated in Theorem 2 in the main text. Fix  $0 < \delta < 1/\sqrt{8\pi e}$ ,  $d < \infty$ ,

and  $\theta$ , and consider the small-error limit  $\epsilon \rightarrow 0$ . From the definitions of  $\tau_{0+}$  and  $D$ , we have

$$\tau_{0+} = \left( 1 + \epsilon \frac{d\mu_R}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \right) \epsilon = \epsilon(1 + O(\epsilon)), \quad (\text{D69})$$

and

$$D = O(\epsilon). \quad (\text{D70})$$

In addition,  $\eta$  and  $\sigma_{\theta,a}$  are independent of  $\epsilon$  in this limit. First, Eq. (D42) satisfies

$$\left( \frac{\eta}{\frac{1}{\sqrt{2\pi e}} - 2\delta} \right)^2 = O(1) \quad (\text{D71})$$

and is therefore negligible compared with the leading  $\epsilon^{-2}$  scale. Second, using Eqs. (D69) and (D70), Eq. (D51) satisfies

$$\left( \frac{\max\{\sigma_{\theta,a} - D, 0\}}{\tau_{0+}} \right)^2 = \max \left\{ \frac{\sigma_{\theta,a}^2}{\epsilon^2} (1 + o(1)), 0 \right\}. \quad (\text{D72})$$

Lastly, it remains to evaluate  $M_2$ . From the definition of  $B$ ,

$$B = -\frac{D}{\tau_{0+}} - \frac{\eta}{2\delta} + \frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}}}{\tau_{0+}}. \quad (\text{D73})$$

Since  $D/\tau_{0+} = O(1)$  and  $\eta/(2\delta) = O(1)$ , whereas

$$\frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}}}{\tau_{0+}} = \frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}}}{\epsilon} (1 + o(1)), \quad (\text{D74})$$

we obtain

$$B = \frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}}}{\epsilon} (1 + o(1)). \quad (\text{D75})$$

In particular,  $B > 0$  for sufficiently small  $\epsilon$ . The discriminant is

$$\Delta = B^2 - \frac{2\eta D}{\delta\tau_{0+}}. \quad (\text{D76})$$

Since  $D/\tau_{0+} = O(1)$ , the second term is  $O(1)$ , whereas Eq. (D75) gives

$$B^2 = \frac{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}^2}{\epsilon^2} (1 + o(1)). \quad (\text{D77})$$

Therefore,

$$\Delta = B^2(1 + o(1)), \quad \sqrt{\Delta} = B(1 + o(1)). \quad (\text{D78})$$

It follows that

$$M_2 = \left( \frac{B + \sqrt{\Delta}}{2} \right)^2 = \frac{W_0(\delta^{-2}/8\pi)\sigma_{\theta,a}^2}{\epsilon^2} (1 + o(1)). \quad (\text{D79})$$

Finally, the condition

$$0 < \delta < \frac{1}{\sqrt{8\pi e}} \quad (\text{D80})$$

implies

$$\frac{\delta^{-2}}{8\pi} > e, \quad (\text{D81})$$

and therefore

$$W_0(\delta^{-2}/8\pi) > 1. \quad (\text{D82})$$

Consequently, the  $M_2$  contribution in Eq. (D79) asymptotically dominates the branch contribution in Eq. (D72). Thus, for fixed  $\theta$  and  $a$ ,

$$M_0(\epsilon, \delta, d) \gtrsim W_0(\delta^{-2}/8\pi)[\mathbf{F}_{\theta}^{-1}]_{aa}\epsilon^{-2}. \quad (\text{D83})$$

We now derive the simplified small-error form stated in Theorem 2 in the main text. The pointwise asymptotic lower bound obtained above can be equivalently expressed as

$$\liminf_{\epsilon \rightarrow 0} \epsilon^2 M_0(\epsilon, \delta, d) \geq W_0(\delta^{-2}/8\pi)[\mathbf{F}_{\theta}^{-1}]_{aa} \quad (\text{D84})$$

for every  $\theta \in \Theta$  and every  $a \in [d]$ , whenever  $0 < \delta < 1/\sqrt{8\pi e}$ .

Define

$$\mathfrak{F} := \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\theta}^{-1}]_{aa}. \quad (\text{D85})$$

If  $\mathfrak{F} < \infty$ , then for every  $\gamma > 0$  there exist  $\theta_{\gamma} \in \Theta$  and  $a_{\gamma} \in [d]$  such that

$$[\mathbf{F}_{\theta_{\gamma}}^{-1}]_{a_{\gamma}a_{\gamma}} \geq \mathfrak{F} - \gamma. \quad (\text{D86})$$

Applying Eq. (D84) to the pair  $(\theta_{\gamma}, a_{\gamma})$  yields

$$\liminf_{\epsilon \rightarrow 0} \epsilon^2 M_0(\epsilon, \delta, d) \geq W_0(\delta^{-2}/8\pi)(\mathfrak{F} - \gamma). \quad (\text{D87})$$

Since  $\gamma > 0$  is arbitrary, we conclude that

$$\liminf_{\epsilon \rightarrow 0} \epsilon^2 M_0(\epsilon, \delta, d) \geq W_0(\delta^{-2}/8\pi) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\theta}^{-1}]_{aa}. \quad (\text{D88})$$

This is precisely the simplified small-error lower bound stated in Theorem 2.

## APPENDIX E: PROOF OF THE UPPER BOUND FOR $\mathcal{L}_2$

Assume that conditions (A1) and (A2) and the standard regularity conditions (R1)–(R5) stated in Appendix 3 B 4 hold. Then the following theorem applies.

*Theorem 3.* Fix  $d < \infty$ . There exists a constant  $\epsilon_0 > 0$  such that, for every  $0 < \epsilon < \epsilon_0$  and  $0 < \delta \leq 1$ , the following holds. Let  $M_0(\epsilon, \delta, d)$  denote the minimal number of repetitions such that

$$\forall M \geq M_0(\epsilon, \delta, d): \Pr[\|\tilde{\theta}^{\text{ML}} - \theta\|_2 \leq \epsilon] \geq 1 - \delta \quad \text{for all } \theta \in \Theta. \quad (\text{E1})$$

$M_0$  is then upper bounded as

$$M_0 \leq \sup_{\theta \in \Theta} \max \left\{ \left( \frac{D}{\tau_{0-}} \right)^2, \left( \frac{2d\eta}{\delta} \right)^2, \left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} + \sqrt{\left( \frac{d\eta}{\delta} + \frac{D}{2\tau_{0-}} + \frac{\sigma\sqrt{W_0}}{2\tau_{0-}} \right)^2 - \frac{2d\eta}{\delta} \frac{D}{\tau_{0-}}} \right)^2 \right\}, \quad (\text{E2})$$

where

$$W_0 := W_0(8\pi^{-1}\delta^{-2}d^2), \quad (\text{E3})$$

$$\tau_{0-} := \left( 1 - \epsilon \frac{\sqrt{d}\mu_R}{2} \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \right) \frac{\epsilon}{\sqrt{d}}, \quad (\text{E4})$$

$$D := \left( \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \sqrt{\frac{4V_H}{\delta}} + \frac{1}{2} \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \sqrt{\frac{4V_R}{\delta}} \epsilon \right) \epsilon, \quad (\text{E5})$$

$$\eta := \frac{1}{d} \sum_{a=1}^d \frac{2C\rho}{\sigma_{\theta,a}^3}, \quad (\text{E6})$$

$$\rho := \sup_{\theta \in \Theta} \max_{a \in [d]} \mathbb{E}[|\mathbf{e}_a^T \mathbf{F}_{\theta}^{-1} \ell_{\theta}^{(1)}(x)|^3], \quad (\text{E7})$$

$$\sigma := \sqrt{\sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\theta}^{-1}]_{aa}}. \quad (\text{E8})$$

In the small-error limit  $\epsilon \rightarrow 0$ , Eq. (E2) reduces to

$$M_0 \lesssim dW_0(8\pi^{-1}\delta^{-2}d^2) \sup_{\theta \in \Theta} \max_{a \in [d]} [\mathbf{F}_{\theta}^{-1}]_{aa} \epsilon^{-2}. \quad (\text{E9})$$

We derive the upper bound for  $\ell_2$ -distance-based learning by reducing the problem to the previously established upper bound for  $\ell_{\infty}$ -distance-based learning. The key observation is the standard norm inequality: for any  $\mathbf{x} \in \mathbb{R}^d$ ,

$$\|\mathbf{x}\|_2 \leq \sqrt{d} \|\mathbf{x}\|_{\infty}. \quad (\text{E10})$$

Applying this inequality to the MLE estimation error vector  $\Delta^{\text{ML}}$ , we obtain

$$\begin{aligned} \Pr[\|\Delta^{\text{ML}}\|_2 \geq \epsilon] &\leq \Pr\left[\|\Delta^{\text{ML}}\|_{\infty} \geq \frac{\epsilon}{\sqrt{d}}\right] \leq \Pr[\mathcal{E}^c] \\ &= \Pr[\{\|\mathbf{F}_{\theta}^{-1} \mathbf{S}_{\theta}\|_{\infty} \geq \tau_{-}\} \cup \{\|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} \geq c_H\} \cup \{R \geq c_R\}] \\ &\leq \Pr[\{\|\mathbf{F}_{\theta}^{-1} \mathbf{S}_{\theta}\|_{\infty} \geq \tau_{-}\}] + \delta_H + \delta_R, \end{aligned} \quad (\text{E11})$$

where the inequalities follow directly from Eq. (C61). Compared with the  $\ell_{\infty}$  case, the definitions of the threshold  $\tau_{-}$  and the good event  $\mathcal{E}$  are modified to incorporate the norm-conversion factor  $d^{1/2}$ . Specifically, we define

$$\tau_{-} := \frac{\epsilon}{\sqrt{d}} - \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} c_H \epsilon - \frac{1}{2} \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} c_R \epsilon^2, \quad (\text{E12})$$

and

$$\begin{aligned} \mathcal{E} := \{ \|\mathbf{F}_{\theta}^{-1} \mathbf{S}_{\theta}\|_{\infty} \leq \tau_{-} \} \cap \{ \|\mathbf{H}_{\theta} + \mathbf{F}_{\theta}\|_{\text{op}} \leq c_H \} \\ \cap \{ R \leq c_R \}. \end{aligned} \quad (\text{E13})$$

The remainder of the proof proceeds identically to the  $\ell_{\infty}$ -based case.

## APPENDIX F: PROOF OF THE LOWER BOUND FOR $\ell_2$

Assume that conditions (A1) and (A2) and the standard regularity conditions (R1)–(R5) stated in Appendix 3 B 4 hold. Then the following theorem applies.

*Theorem 4.* For  $0 < \epsilon$ ,  $0 < \delta < 1/\sqrt{8\pi e}$  and  $d < \infty$ , let  $M_0(\epsilon, \delta, d)$  denote the minimal number of repetitions such that

$$\forall M \geq M_0(\epsilon, \delta, d): \Pr[\|\tilde{\theta}^{\text{ML}} - \theta\|_2 \leq \epsilon] \geq 1 - \delta \text{ for all } \theta \in \Theta. \quad (\text{F1})$$

$M_0$  is then lower bounded as

$$M_0 \geq \max \left\{ M_2(\theta), \left( \frac{\max\{\sigma_{\theta} - D, 0\}}{\tau_{0+}} \right)^2, \left( \frac{\eta}{\frac{1}{\sqrt{2\pi e}} - 2\delta} \right)^2 \right\}, \quad (\text{F2})$$

for any  $\theta \in \Theta$ . Here

$$M_2(\theta) := \begin{cases} \left( \frac{B_{\theta} + \sqrt{\Delta_{\theta}}}{2} \right)^2, & B_{\theta} > 0 \text{ and } \Delta_{\theta} > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (\text{F3})$$

and

$$B_{\theta} := -\frac{D}{\tau_{0+}} - \frac{\eta}{2\delta} + \frac{\sqrt{W_0(\delta^{-2}/8\pi)\sigma_{\theta}}}{\tau_{0+}}, \quad (\text{F4})$$

$$\Delta_{\theta} := B_{\theta}^2 - \frac{2\eta D}{\delta\tau_{0+}}, \quad (\text{F5})$$

$$\tau_{0,+} := \left( 1 + \epsilon \frac{\mu_R}{2} \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \right) \epsilon, \quad (\text{F6})$$

$$D := \left( \sqrt{\frac{2V_H}{\delta}} + \frac{1}{2} \sqrt{\frac{2V_R}{\delta}} \epsilon \right) \|\mathbf{F}_{\theta}^{-1}\|_{\text{op}} \epsilon, \quad (\text{F7})$$

$$\eta := \frac{2C\rho_{\theta}}{\sigma_{\theta}^3}, \quad (\text{F8})$$

$$\sigma_\theta := \sqrt{\mu_{\max}(\mathbf{F}_\theta^{-1})}. \quad (\text{F9})$$

Here

$$\rho_\theta := \mathbb{E}[|u_\theta^T \mathbf{F}_\theta^{-1} \ell_\theta^{(1)}(X)|^3], \quad (\text{F10})$$

where  $u_\theta$  is a unit eigenvector of  $\mathbf{F}_\theta^{-1}$  corresponding to  $\mu_{\max}(\mathbf{F}_\theta^{-1})$ . In the small-error limit  $\epsilon \rightarrow 0$ , this lower bound reduces to

$$M_0(\epsilon, \delta, d) \gtrsim W_0(\delta^{-2}/8\pi) \sup_{\theta \in \Theta} \mu_{\max}(\mathbf{F}_\theta^{-1}) \epsilon^{-2}. \quad (\text{F11})$$

We begin with the expansion

$$\Delta^{\text{ML}} = \mathbf{F}_\theta^{-1} \mathbf{S}_\theta + \mathbf{F}_\theta^{-1} (\mathbf{H}_\theta + \mathbf{F}_\theta) \Delta^{\text{ML}} + \mathbf{F}_\theta^{-1} \mathbf{r}_\theta (\Delta^{\text{ML}}). \quad (\text{F12})$$

Let  $\mu_{\max}(\mathbf{F}_\theta^{-1})$  denote the largest eigenvalue of  $\mathbf{F}_\theta^{-1}$ , and let  $u_\theta \in \mathbb{R}^d$  be a corresponding unit eigenvector, namely,

$$\mathbf{F}_\theta^{-1} u_\theta = \mu_{\max}(\mathbf{F}_\theta^{-1}) u_\theta, \quad \|u_\theta\|_2 = 1. \quad (\text{F13})$$

Taking the inner product of the above decomposition with  $u_\theta$  yields

$$\begin{aligned} u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta &= u_\theta^T \Delta^{\text{ML}} - u_\theta^T \mathbf{F}_\theta^{-1} (\mathbf{H}_\theta + \mathbf{F}_\theta) \Delta^{\text{ML}} - u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{r}_\theta (\Delta^{\text{ML}}). \end{aligned} \quad (\text{F14})$$

Applying the triangle inequality and standard operator-norm bounds, we obtain

$$\begin{aligned} |u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| &\leq |u_\theta^T \Delta^{\text{ML}}| + |u_\theta^T \mathbf{F}_\theta^{-1} (\mathbf{H}_\theta + \mathbf{F}_\theta) \Delta^{\text{ML}}| + |u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{r}_\theta (\Delta^{\text{ML}})| \\ &\leq \|\Delta^{\text{ML}}\|_2 + \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \|\mathbf{H}_\theta + \mathbf{F}_\theta\|_{\text{op}} \|\Delta^{\text{ML}}\|_2 + \|\mathbf{F}_\theta^{-1}\|_{\text{op}} \|\mathbf{r}_\theta(\Delta^{\text{ML}})\|_2. \end{aligned} \quad (\text{F15})$$

Consequently, we have

$$\Pr[\|\Delta^{\text{ML}}\|_2 \leq \epsilon] \leq \Pr[\{\|\Delta^{\text{ML}}\|_2 \leq \epsilon\} \cap \mathcal{G}] + \delta_H + \delta_R, \quad (\text{F16})$$

where the good event  $\mathcal{G}$  is defined as

$$\mathcal{G} := \{\|\mathbf{H}_\theta + \mathbf{F}_\theta\|_{\text{op}} \leq c_H\} \cap \{R \leq c_R\}. \quad (\text{F17})$$

On the event  $\{\|\Delta^{\text{ML}}\|_2 \leq \epsilon\} \cap \mathcal{G}$ , Eq. (F15) implies

$$\begin{aligned} |u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| &\leq \epsilon + \|\mathbf{F}_\theta^{-1}\|_{\text{op}} c_H \epsilon + \frac{1}{2} \|\mathbf{F}_\theta^{-1}\|_{\text{op}} c_R \epsilon^2 =: \epsilon'(\epsilon, d, M). \end{aligned} \quad (\text{F18})$$

Combining Eqs. (F16) and (F18), we obtain

$$\Pr[\|\Delta^{\text{ML}}\|_2 \leq \epsilon] \leq \Pr[|u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta| \leq \epsilon'] + \delta_H + \delta_R. \quad (\text{F19})$$

Finally, the projected score has mean zero and variance

$$\mathbb{E}[u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta] = 0, \quad \mathbb{V}[u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta] = \frac{\mu_{\max}(\mathbf{F}_\theta^{-1})}{M} = \frac{\sigma_\theta^2}{M}. \quad (\text{F20})$$

The remainder of the proof proceeds identically to the  $\ell_\infty$ -based case, with the one-dimensional projection

$u_\theta^T \mathbf{F}_\theta^{-1} \mathbf{S}_\theta$  playing the role of the coordinate projection used there.

## APPENDIX G: SINGULAR FIM

### 1. Asymptotic unbiased estimator

We establish asymptotic unbiasedness under mild concentration and moment assumptions.

*Theorem 5.* Assume that there exists  $\epsilon_0 > 0$  such that for any  $\delta \in (0, 1]$  and any  $\epsilon \in (0, \epsilon_0]$  there exists an integer  $M_0 = M_0(\delta, \epsilon)$  satisfying, for all  $M \geq M_0$ ,

$$\Pr[|\tilde{\theta}(\mathbf{x}) - \theta| \leq \epsilon] \geq 1 - \delta. \quad (\text{G1})$$

Moreover, assume there exists  $\eta > 0$  such that

$$\sup_{M \geq 1} \mathbb{E}[|\tilde{\theta} - \theta|^{1+\eta}] =: C < \infty. \quad (\text{G2})$$

Then  $\tilde{\theta}$  is asymptotically unbiased, i.e.,

$$\lim_{M \rightarrow \infty} \mathbb{E}[\tilde{\theta}] = \theta. \quad (\text{G3})$$

*Proof.* Define  $\Delta := \tilde{\theta} - \theta$ . For a given  $\epsilon \in (0, \epsilon_0)$ , decompose

$$\mathbb{E}[\Delta] = \mathbb{E}[\Delta \mathbf{1}\{|\Delta| \leq \epsilon\}] + \mathbb{E}[\Delta \mathbf{1}\{|\Delta| > \epsilon\}], \quad (\text{G4})$$

where  $\mathbf{1}\{\cdot\}$  denotes the indicator function. For the first term,

$$|\mathbb{E}[\Delta \mathbf{1}\{|\Delta| \leq \epsilon\}]| \leq \mathbb{E}[|\Delta| \mathbf{1}\{|\Delta| \leq \epsilon\}] \leq \epsilon. \quad (\text{G5})$$

For the second term, applying Hölder's inequality with conjugate exponents  $p = 1 + \eta$  and  $q = (1 + \eta)/\eta$  gives

$$\begin{aligned} |\mathbb{E}[\Delta \mathbf{1}\{|\Delta| > \epsilon\}]| &\leq \mathbb{E}[|\Delta|^{1+\eta}]^{\frac{1}{1+\eta}} \Pr(|\Delta| > \epsilon)^{\frac{\eta}{1+\eta}} \\ &\leq C^{\frac{1}{1+\eta}} \Pr(|\Delta| > \epsilon)^{\frac{\eta}{1+\eta}}, \end{aligned} \quad (\text{G6})$$

where the last inequality uses (G2).

For a given  $\delta \in (0, 1)$ , Eq. (G1) implies that, for all  $M \geq M_0(\delta, \epsilon)$ ,  $\Pr(|\Delta| > \epsilon) \leq \delta$ . Hence, combining Eqs. (G4), (G5), and (G6), we obtain

$$|\mathbb{E}[\Delta]| \leq \epsilon + C^{\frac{1}{1+\eta}} \delta^{\frac{\eta}{1+\eta}} \quad (\text{G7})$$

for all  $M \geq M_0(\delta, \epsilon)$ .

To conclude, let  $\alpha > 0$  be arbitrary. Choose  $\epsilon \in (0, \epsilon_0)$  such that  $\epsilon \leq \alpha/2$ , and choose  $\delta \in (0, 1)$  such that

$$C^{\frac{1}{1+\eta}} \delta^{\frac{\eta}{1+\eta}} \leq \alpha/2.$$

Then Eq. (G7) implies that, for all  $M \geq M_0(\delta, \epsilon)$ ,

$$|\mathbb{E}[\Delta]| \leq \alpha.$$

Therefore,  $\mathbb{E}[\Delta] \rightarrow 0$  as  $M \rightarrow \infty$ . Since  $\Delta = \tilde{\theta} - \theta$ , we obtain  $\mathbb{E}[\tilde{\theta}] \rightarrow \theta$ . ■

## 2. Biased estimator

By Theorem 5, for any scalar estimator  $\tilde{\theta}$  satisfying the moment condition (G2), failure of asymptotic unbiasedness necessarily implies failure of the concentration property (G1). Equivalently, there exist  $\delta \in (0, 1]$  and  $\epsilon \in (0, \epsilon_0]$  such that, for every integer  $M_0$ , one can find  $M \geq M_0$  satisfying

$$\Pr[|\tilde{\theta}(\mathbf{x}) - \theta| \leq \epsilon] < 1 - \delta. \quad (\text{G8})$$

In the singular-FIM setting, this observation motivates restricting attention to *unbiasedly estimable* directions.

### a. Estimable subspace

Let

$$\text{supp}(\mathbf{F}_\theta) := \text{Im}(\mathbf{F}_\theta) \quad (\text{G9})$$

denote the support subspace of the FIM. It is well known that a scalar functional  $\mathbf{a}^T \theta$  is unbiasedly estimable if and only if

$$\mathbf{F}_\theta \mathbf{F}_\theta^+ \mathbf{a} = \mathbf{a}, \quad (\text{G10})$$

or equivalently, if and only if  $\mathbf{a} \in \text{supp}(\mathbf{F}_\theta)$ , where  $\mathbf{F}_\theta^+$  denotes the Moore-Penrose pseudoinverse of  $\mathbf{F}_\theta$  [63].

In particular, the coordinate parameter  $\theta_a$  is unbiasedly estimable if and only if the corresponding basis vector  $\mathbf{e}_a$  satisfies

$$\mathbf{F}_\theta \mathbf{F}_\theta^+ \mathbf{e}_a = \mathbf{e}_a. \quad (\text{G11})$$

Therefore, in the singular case, the relevant object is not merely a subset of the original coordinate parameters, but rather the estimable subspace  $\text{supp}(\mathbf{F}_\theta)$  itself.

Let

$$r := \text{rank}(\mathbf{F}_\theta), \quad k := d - r. \quad (\text{G12})$$

Since  $\mathbf{F}_\theta$  is symmetric and positive semidefinite, we have the orthogonal decomposition

$$\mathbb{R}^d = \ker(\mathbf{F}_\theta) \oplus \text{supp}(\mathbf{F}_\theta). \quad (\text{G13})$$

## b. Regularity assumptions

We impose the following assumptions on the log-likelihood function  $\ell_\theta(\mathbf{x})$ :

- (B1) Profile maximizer and stationary point: For each fixed null-space coordinate, the log-likelihood, viewed as a function of the estimable coordinates, admits a unique maximizer in the interior of the parameter space  $\Theta$ . Moreover, this maximizer is the unique stationary point with respect to the estimable coordinates.
- (B2) Smoothness: The function  $\ell_\theta(\mathbf{x})$  is three times continuously differentiable on  $\Theta$ .
- (B3) Constant support subspace: The rank  $r$  is constant throughout  $\Theta$ , and the support subspace  $\text{supp}(\mathbf{F}_\theta)$  is independent of  $\theta$  over  $\Theta$ .

## c. Orthogonal reparametrization

Under Assumption (B3), we may choose an orthogonal matrix

$$\mathbf{U} := (\mathbf{U}^{(0)} \ \mathbf{U}^{(1)}) \in \mathbb{R}^{d \times d}, \quad (\text{G14})$$

where the columns of  $\mathbf{U}^{(0)} \in \mathbb{R}^{d \times k}$  form an orthonormal basis of  $\ker(\mathbf{F}_\theta)$ , and the columns of  $\mathbf{U}^{(1)} \in \mathbb{R}^{d \times r}$  form an orthonormal basis of  $\text{supp}(\mathbf{F}_\theta)$ . We then introduce the orthogonal reparametrization

$$\boldsymbol{\xi} := \mathbf{U}^T \theta = \begin{pmatrix} \boldsymbol{\xi}^{(0)} \\ \boldsymbol{\xi}^{(1)} \end{pmatrix}, \quad (\text{G15})$$

where  $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^k$  denotes the null-space coordinates, while  $\boldsymbol{\xi}^{(1)} \in \mathbb{R}^r$  denotes the estimable coordinates.

*Lemma 1.* (Null-space directions annihilate the score). Assume that the Fisher information matrix is defined by the score outer product

$$\mathbf{F}_\theta := \mathbb{E}_\theta[\mathbf{S}_\theta(X)\mathbf{S}_\theta(X)^T], \quad (\text{G16})$$

where  $\mathbf{S}_\theta(x) := \nabla_\theta \ell_\theta(x)$ . Then for every  $\mathbf{u} \in \ker(\mathbf{F}_\theta)$ ,

$$\mathbf{u}^T \mathbf{S}_\theta(X) = 0 \quad p_\theta\text{-a.s.} \quad (\text{G17})$$

Moreover, on the common support where  $p_\theta(x) > 0$ ,

$$\mathbf{u}^T \nabla_\theta p_\theta(x) = 0 \quad p_\theta\text{-a.s.} \quad (\text{G18})$$

*Proof.* Let  $\mathbf{u} \in \ker(\mathbf{F}_\theta)$ . Then

$$0 = \mathbf{u}^T \mathbf{F}_\theta \mathbf{u} = \mathbb{E}_\theta[\mathbf{u}^T \mathbf{S}_\theta(X) \mathbf{S}_\theta(X)^T \mathbf{u}] = \mathbb{E}_\theta[(\mathbf{u}^T \mathbf{S}_\theta(X))^2]. \quad (\text{G19})$$

Since the integrand is nonnegative, the expectation can vanish only if

$$(\mathbf{u}^T \mathbf{S}_\theta(X))^2 = 0$$

holds  $p_\theta$ -almost surely, which establishes (G17). On the common support where  $p_\theta(x) > 0$ , we have

$$\nabla_\theta p_\theta(x) = p_\theta(x) \nabla_\theta \log p_\theta(x) = p_\theta(x) \mathbf{S}_\theta(x). \quad (\text{G20})$$

Multiplying both sides by  $\mathbf{u}^T$  and using (G17) yields (G18).  $\blacksquare$

*Lemma 2.* (Block form of the score under an orthogonal decomposition). Assume (B3). Define  $\xi := \mathbf{U}^T \theta$  and write  $\xi = (\xi^{(0)}, \xi^{(1)})^T$  accordingly. Then the score in the  $\xi$ -coordinates satisfies

$$\mathbf{S}_\xi(X) := \mathbf{U}^T \nabla_\theta \ell_\theta(X) = \begin{pmatrix} \mathbf{0} \\ \mathbf{S}_\xi^{(1)}(X) \end{pmatrix} \quad p_\theta\text{-a.s.} \quad (\text{G21})$$

In particular, the score has no component along the null-space directions  $\xi^{(0)}$ .

*Proof.* Since  $\theta = \mathbf{U}\xi$ , the Jacobian is

$$\frac{\partial \theta}{\partial \xi} = \mathbf{U},$$

and therefore, by the chain rule,

$$\nabla_\xi \ell_\xi(X) = \left( \frac{\partial \theta}{\partial \xi} \right)^T \nabla_\theta \ell_\theta(X) = \mathbf{U}^T \nabla_\theta \ell_\theta(X). \quad (\text{G22})$$

Each column  $\mathbf{u}_i$  of  $\mathbf{U}^{(0)}$  belongs to  $\ker(\mathbf{F}_\theta)$ . Hence, by Lemma 1,

$$\mathbf{u}_i^T \nabla_\theta \ell_\theta(X) = 0$$

holds  $p_\theta$ -almost surely. Stacking these identities gives

$$(\mathbf{U}^{(0)})^T \nabla_\theta \ell_\theta(X) = \mathbf{0} \quad p_\theta\text{-a.s.} \quad (\text{G23})$$

Therefore, the first block of  $\mathbf{U}^T \nabla_\theta \ell_\theta(X)$  vanishes, establishing (G21).  $\blacksquare$

Equation (G21) shows that the likelihood is locally insensitive (in the sense of vanishing directional derivatives,  $p_\theta$ -a.s.) along the null-space directions  $\xi^{(0)}$ .

#### d. Profile maximum likelihood on the estimable subspace

For fixed  $\xi^{(0)}$ , define the profile MLE for the estimable coordinates by

$$\tilde{\xi}^{(1),\text{ML}} \in \arg \max_{\xi \in \mathbb{R}^r} \ell_{(\xi^{(0)}, \xi)}(X). \quad (\text{G24})$$

Let the corresponding displacement be

$$\tilde{\Delta}^{(1),\text{ML}} := \tilde{\xi}^{(1),\text{ML}} - \xi^{(1)}. \quad (\text{G25})$$

We define the reduced, or estimable, score and Hessian as

$$\mathbf{S}_\xi^{(1)}(X) := \nabla_{\xi^{(1)}} \ell_\xi(X), \quad (\text{G26})$$

$$\mathbf{H}_\xi^{(11)}(X) := \nabla_{\xi^{(1)}}^2 \ell_\xi(X). \quad (\text{G27})$$

By the first-order optimality condition for the profile MLE, together with Assumption (B1), we have

$$\mathbf{0} = \mathbf{S}_{(\xi^{(0)}, \tilde{\xi}^{(1),\text{ML}})}^{(1)}(X). \quad (\text{G28})$$

Applying Taylor's theorem to the reduced score around  $\xi$  yields

$$\mathbf{0} = \mathbf{S}_\xi^{(1)}(X) + \mathbf{H}_\xi^{(11)}(X) \tilde{\Delta}^{(1),\text{ML}} + \mathbf{r}_\xi^{(1)}(\tilde{\Delta}^{(1),\text{ML}}), \quad (\text{G29})$$

where  $\mathbf{r}_\xi^{(1)}(\cdot)$  denotes the corresponding Taylor remainder.

#### e. Pseudoinverse reduction

The Fisher information matrix in the  $\xi$ -coordinates is

$$\mathbf{F}_\xi = \mathbf{U}^T \mathbf{F}_\theta \mathbf{U} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{F}}_\xi \end{pmatrix}, \quad (\text{G30})$$

where  $\bar{\mathbf{F}}_\xi \in \mathbb{R}^{r \times r}$  is positive definite. Consequently, its Moore-Penrose inverse is

$$\mathbf{F}_\xi^+ = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{F}}_\xi^{-1} \end{pmatrix}. \quad (\text{G31})$$

Adding and subtracting  $\bar{\mathbf{F}}_\xi \tilde{\Delta}^{(1),\text{ML}}$  in Eq. (G29), and then rearranging, gives

$$\bar{\mathbf{F}}_{\xi} \tilde{\Delta}^{(1),\text{ML}} = \mathbf{S}_{\xi}^{(1)} + (\mathbf{H}_{\xi}^{(11)} + \bar{\mathbf{F}}_{\xi}) \tilde{\Delta}^{(1),\text{ML}} + \mathbf{r}_{\xi}^{(1)} (\tilde{\Delta}^{(1),\text{ML}}). \quad (\text{G32})$$

Thus,

$$\begin{aligned} \tilde{\Delta}^{(1),\text{ML}} &= \bar{\mathbf{F}}_{\xi}^{-1} \mathbf{S}_{\xi}^{(1)} + \bar{\mathbf{F}}_{\xi}^{-1} (\mathbf{H}_{\xi}^{(11)} + \bar{\mathbf{F}}_{\xi}) \tilde{\Delta}^{(1),\text{ML}} \\ &\quad + \bar{\mathbf{F}}_{\xi}^{-1} \mathbf{r}_{\xi}^{(1)} (\tilde{\Delta}^{(1),\text{ML}}). \end{aligned} \quad (\text{G33})$$

### f. Consequence for sample complexity bounds

We therefore conclude that the learning problem reduces to the non-singular case restricted to the  $r$ -dimensional estimable subspace  $\text{supp}(\mathbf{F}_{\theta})$ . As a result, sample-complexity bounds established for the non-singular setting extend directly to the estimable coordinates  $\xi^{(1)}$ , with the replacements

$$\mathbf{F}_{\theta}^{-1} \rightarrow \mathbf{F}_{\theta}^{+}, \quad d \rightarrow r = \text{rank}(\mathbf{F}_{\theta}), \quad (\text{G34})$$

provided the corresponding bound is formulated in an orthogonally invariant manner.

## APPENDIX H: VALIDITY OF (A1)–(A2) IN STATISTICAL MODELS

- (A1) Unique maximizer and stationary point:  $\ell_{\theta}(\mathbf{x})$  has a unique maximizer  $\tilde{\theta}^{\text{ML}}$  in the interior of the parameter domain  $\Theta$  and it is the unique stationary point.
- (A2) Smoothness:  $\ell_{\theta}(\mathbf{x})$  is three times continuously differentiable with respect to  $\theta$  on the parameter domain  $\Theta$ .

### 1. Bernoulli model

Let  $x_1, \dots, x_M \in \{0, 1\}$  be i.i.d. Bernoulli random variables with parameter  $\theta \in (0, 1)$ , and define

$$S := \sum_{i=1}^M x_i.$$

The log-likelihood function is

$$\begin{aligned} \ell_{\theta}(\mathbf{x}) &= S \log \theta + (M - S) \log(1 - \theta), \\ \theta \in \Theta &:= (0, 1). \end{aligned} \quad (\text{H1})$$

- (A1). The first derivative of the log-likelihood is

$$\ell'_{\theta}(\mathbf{x}) = \frac{S}{\theta} - \frac{M - S}{1 - \theta}. \quad (\text{H2})$$

Setting  $\ell'_{\theta}(\mathbf{x}) = 0$  yields the unique stationary point

$$\theta^{\text{ML}} = \frac{S}{M}, \quad (\text{H3})$$

provided  $0 < S < M$ . Moreover, since

$$\ell''_{\theta}(\mathbf{x}) = -\frac{S}{\theta^2} - \frac{M - S}{(1 - \theta)^2} < 0, \quad (\text{H4})$$

the log-likelihood is strictly concave on  $(0, 1)$ , and therefore  $\theta^{\text{ML}}$  is the unique maximizer.

- (A2). The function  $\ell_{\theta}(\mathbf{x})$  is infinitely differentiable on  $(0, 1)$ .

### 2. Gaussian model with known variance

Let  $\mathbf{y}_1, \dots, \mathbf{y}_M \in \mathbb{R}^d$  be i.i.d. Gaussian random vectors distributed according to  $\mathcal{N}(\theta, \Sigma)$ , where the covariance matrix  $\Sigma \succ 0$  is known and  $\theta \in \mathbb{R}^d$  is unknown. Up to an additive constant, the log-likelihood function is

$$\ell_{\theta}(\mathbf{x}) = -\frac{1}{2} \sum_{i=1}^M (\mathbf{y}_i - \theta)^{\text{T}} \Sigma^{-1} (\mathbf{y}_i - \theta), \quad \theta \in \Theta := \mathbb{R}^d. \quad (\text{H5})$$

- (A1). The gradient of the log-likelihood is

$$\nabla_{\theta} \ell_{\theta}(\mathbf{x}) = M \Sigma^{-1} (\bar{\mathbf{y}} - \theta), \quad \bar{\mathbf{y}} := \frac{1}{M} \sum_{i=1}^M \mathbf{y}_i. \quad (\text{H6})$$

Therefore,  $\nabla_{\theta} \ell_{\theta}(\mathbf{x}) = 0$  if and only if

$$\theta = \bar{\mathbf{y}} =: \theta^{\text{ML}},$$

which is thus the unique stationary point. Moreover, the Hessian is

$$\nabla_{\theta}^2 \ell_{\theta}(\mathbf{x}) = -M \Sigma^{-1}, \quad (\text{H7})$$

which is strictly negative definite because  $\Sigma \succ 0$ . Hence the log-likelihood is strictly concave, and  $\theta^{\text{ML}}$  is the unique global maximizer.

- (A2). The function  $\ell_{\theta}(\mathbf{x})$  is a quadratic polynomial in  $\theta$  and is therefore infinitely differentiable.

### 3. Multinomial model

Let  $(n_1, \dots, n_K)$  be multinomial counts with total count

$$M = \sum_{k=1}^K n_k,$$

and parameter vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ , where  $\theta_k > 0$  and  $\sum_{k=1}^K \theta_k = 1$ . The log-likelihood function is

$$\ell_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^K n_k \log \theta_k, \quad (\text{H8})$$

defined on the probability simplex.

(A1). The partial derivatives are

$$\frac{\partial \ell_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \theta_k} = \frac{n_k}{\theta_k}. \quad (\text{H9})$$

Because the parameters satisfy the normalization constraint

$$\sum_{k=1}^K \theta_k = 1,$$

the stationary point must be determined using a Lagrange multiplier. Define

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \sum_{k=1}^K n_k \log \theta_k + \lambda \left( \sum_{k=1}^K \theta_k - 1 \right). \quad (\text{H10})$$

The stationarity conditions give

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = \frac{n_k}{\theta_k} + \lambda = 0, \quad (\text{H11})$$

which implies

$$\theta_k = -\frac{n_k}{\lambda}. \quad (\text{H12})$$

Using the normalization condition,

$$1 = \sum_{k=1}^K \theta_k = -\frac{1}{\lambda} \sum_{k=1}^K n_k = -\frac{M}{\lambda}, \quad (\text{H13})$$

and therefore  $\lambda = -M$ . Hence the unique stationary point under the normalization constraint is

$$\theta_k^{\text{ML}} = \frac{n_k}{M}. \quad (\text{H14})$$

Moreover, since

$$\frac{\partial^2 \ell_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \theta_k^2} = -\frac{n_k}{\theta_k^2} < 0, \quad (\text{H15})$$

the log-likelihood is strictly concave on the interior of the simplex, implying that this stationary point is the unique global maximizer.

(A2). The function  $\ell_{\boldsymbol{\theta}}(\mathbf{x})$  is infinitely differentiable throughout the interior region  $\theta_k > 0$ .

#### 4. Poisson model

Let  $x_1, \dots, x_M$  be i.i.d. Poisson random variables with mean  $\theta > 0$ . The log-likelihood function is

$$\ell_{\theta}(\mathbf{x}) = \sum_{i=1}^M (x_i \log \theta - \theta) + \text{const}, \quad \theta \in \Theta := (0, \infty). \quad (\text{H16})$$

(A1). The derivative of the log-likelihood is

$$\ell'_{\theta}(\mathbf{x}) = \sum_{i=1}^M \frac{x_i}{\theta} - M. \quad (\text{H17})$$

Setting  $\ell'_{\theta}(\mathbf{x}) = 0$  yields the unique stationary point

$$\theta^{\text{ML}} = \frac{1}{M} \sum_{i=1}^M x_i. \quad (\text{H18})$$

Moreover,

$$\ell''_{\theta}(\mathbf{x}) = -\sum_{i=1}^M \frac{x_i}{\theta^2} \leq 0, \quad (\text{H19})$$

so the log-likelihood is concave on  $(0, \infty)$ . If  $\sum_{i=1}^M x_i > 0$ , the concavity is strict and  $\theta^{\text{ML}}$  is the unique global maximizer.

(A2). The function  $\ell_{\theta}(\mathbf{x})$  is infinitely differentiable on  $\Theta = (0, \infty)$ .

#### 5. General exponential family (canonical parametrization)

Let  $x_1, \dots, x_M$  be i.i.d. measurement outcomes drawn from a *regular* exponential family with canonical parameter  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$  and density

$$p_{\boldsymbol{\theta}}(x) = h(x) \exp(\boldsymbol{\theta}^T \mathbf{t}(x) - A(\boldsymbol{\theta})), \quad \boldsymbol{\theta} \in \Theta, \quad (\text{H20})$$

where  $\mathbf{t}(x) \in \mathbb{R}^d$  is the sufficient statistic and  $A(\boldsymbol{\theta})$  is the log-partition function. Define the aggregated sufficient statistic

$$\mathbf{T} := \sum_{i=1}^M \mathbf{t}(x_i) \in \mathbb{R}^d. \quad (\text{H21})$$

The log-likelihood function is

$$\begin{aligned} \ell_{\theta}(\mathbf{x}) &:= \sum_{i=1}^M \log p_{\theta}(x_i) \\ &= \sum_{i=1}^M \log h(x_i) + \boldsymbol{\theta}^T \sum_{i=1}^M \mathbf{t}(x_i) - M A(\boldsymbol{\theta}) \\ &= \sum_{i=1}^M \log h(x_i) + \boldsymbol{\theta}^T \mathbf{T} - M A(\boldsymbol{\theta}). \end{aligned} \quad (\text{H22})$$

(A1). The gradient of the log-likelihood is

$$\nabla_{\theta} \ell_{\theta}(\mathbf{x}) = \mathbf{T} - M \nabla A(\boldsymbol{\theta}). \quad (\text{H23})$$

Setting  $\nabla_{\theta} \ell_{\theta}(\mathbf{x}) = \mathbf{0}$  yields the maximum-likelihood equation

$$\nabla A(\boldsymbol{\theta}^{\text{ML}}) = \frac{1}{M} \mathbf{T}. \quad (\text{H24})$$

Moreover, the Hessian is

$$\nabla_{\theta}^2 \ell_{\theta}(\mathbf{x}) = -M \nabla^2 A(\boldsymbol{\theta}). \quad (\text{H25})$$

In a minimal regular exponential family,  $A(\boldsymbol{\theta})$  is strictly convex on  $\Theta$ , so  $\nabla^2 A(\boldsymbol{\theta}) \succ 0$ . Hence the log-likelihood is strictly concave, and  $\nabla A$  is injective. Therefore, whenever  $\frac{1}{M} \mathbf{T}$  lies in the range of  $\nabla A$  (equivalently, in the mean-parameter space), the solution  $\boldsymbol{\theta}^{\text{ML}}$  exists and is unique.

(A2). Because the family is regular, the log-partition function  $A(\boldsymbol{\theta})$  is finite on the open set  $\Theta$  and is smooth (in fact, real analytic) throughout  $\Theta$ . Consequently,  $\ell_{\theta}(\mathbf{x})$  is infinitely differentiable on  $\Theta$ .

## 6. Pauli eigenvalue estimation

In this subsection, we verify that assumptions (A1)–(A2) hold for Pauli eigenvalue estimation under the standard measurement model. Throughout, we impose the normalization condition  $\lambda_0 = 1$ . For simplicity, let us denote the probability distribution in Eq. (62) by

$$p_{\lambda}(x) := p_x. \quad (\text{H26})$$

Here,  $\{p_x\}_{x=0}^{4^n-1}$  are the measurement outcome probabilities. These probabilities depend linearly on the Pauli

eigenvalues  $\boldsymbol{\lambda} = (1, \lambda_1, \dots, \lambda_{4^n-1})$ , and satisfy  $p_x \geq 0$  together with  $\sum_{x=0}^{4^n-1} p_x = 1$ . We treat  $(p_1, \dots, p_{4^n-1})$  as free parameters and eliminate  $p_0$  through the relation

$$p_0 = 1 - \sum_{k=1}^{4^n-1} p_k. \quad (\text{H27})$$

Accordingly, the parameter domain is

$$\Theta := \left\{ (p_1, \dots, p_{4^n-1}) : p_k > 0 \ \forall k, \sum_{k=1}^{4^n-1} p_k < 1 \right\}, \quad (\text{H28})$$

which is an open convex subset of  $\mathbb{R}^{4^n-1}$ .

Given an observed dataset  $\mathbf{x}$ , summarized by the outcome counts  $\{n_x\}_{x=0}^{4^n-1}$ , the total number of measurement outcomes is

$$M = \sum_{x=0}^{4^n-1} n_x.$$

The multinomial log-likelihood, up to an additive constant independent of  $\boldsymbol{\theta}$ , is given by

$$\begin{aligned} \ell_{\theta}(\mathbf{x}) &= \sum_{x=0}^{4^n-1} n_x \log p_x \\ &= \left( M - \sum_{k=1}^{4^n-1} n_k \right) \log \left( 1 - \sum_{k=1}^{4^n-1} p_k \right) + \sum_{k=1}^{4^n-1} n_k \log p_k, \end{aligned} \quad (\text{H29})$$

where the parameter vector is

$$\boldsymbol{\theta} = (p_1, \dots, p_{4^n-1}) \in \Theta.$$

Assumption (A2) follows immediately from the definition of  $\Theta$ . Indeed, for every  $\boldsymbol{\theta} \in \Theta$ , we have  $p_k > 0$  for all  $k$ , together with

$$1 - \sum_{k=1}^{4^n-1} p_k > 0.$$

Hence all logarithmic arguments appearing in Eq. (H29) are strictly positive. Therefore,  $\ell_{\theta}(\mathbf{x})$  is infinitely differentiable with respect to  $\boldsymbol{\theta}$  on  $\Theta$ , and in particular

$$\ell_{\theta}(\mathbf{x}) \in C^3(\Theta).$$

To verify Assumption (A1), we note that the multinomial likelihood is maximized by the empirical frequencies. Specifically,

$$p_k^{\text{ML}} = \frac{n_k}{M}, \quad k = 1, \dots, 4^n - 1, \quad (\text{H30})$$

with

$$p_0^{\text{ML}} = 1 - \sum_{k=1}^{4^n-1} p_k^{\text{ML}} = \frac{n_0}{M}. \quad (\text{H31})$$

Whenever

$$\tilde{\theta}^{\text{ML}} = (p_1^{\text{ML}}, \dots, p_{4^n-1}^{\text{ML}})$$

lies in the interior of  $\Theta$ , it is the unique maximizer of  $\ell_{\theta}(\mathbf{x})$ , and also the unique stationary point. Equivalently,

$$\nabla_{\theta} \ell_{\theta}(\mathbf{x}) = 0 \iff \theta = \tilde{\theta}^{\text{ML}}. \quad (\text{H32})$$

Therefore, Pauli eigenvalue estimation satisfies assumptions (A1) and (A2) under the normalization constraint  $\lambda_0 = 1$ .

- 
- [1] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, *Characterizing large-scale quantum computers via cycle benchmarking*, *Nat. Commun.* **10**, 5347 (2019).
  - [2] A. Hashim, L. B. Nguyen, N. Goss, B. Marinelli, R. K. Naik, T. Chistolini, J. Hines, J. P. Marceaux, Y. Kim, P. Gokhale, T. Tomesh, S. Chen, L. Jiang, S. Ferracin, K. Rudinger, T. Proctor, K. C. Young, I. Siddiqi, and R. Blume-Kohout, *Practical introduction to benchmarking and characterization of quantum computers*, *PRX Quantum* **6**, 030202 (2025).
  - [3] R. Harper, S. T. Flammia, and J. J. Wallman, *Efficient learning of quantum noise*, *Nat. Phys.* **16**, 1184 (2020).
  - [4] S. Chen, Z. Zhang, L. Jiang, and S. T. Flammia, *Efficient self-consistent learning of gate set Pauli noise*, *PRX Quantum* **7**, 010305 (2026).
  - [5] D. K. Tuckett, S. D. Bartlett, and S. T. Flammia, *Ultra-high error threshold for surface codes with biased noise*, *Phys. Rev. Lett.* **120**, 050505 (2018).
  - [6] E. H. Chen, T. J. Yoder, Y. Kim, N. Sundaresan, S. Srinivasan, M. Li, A. D. Córcoles, A. W. Cross, and M. Takita, *Calibrated decoders for experimental quantum error correction*, *Phys. Rev. Lett.* **128**, 110504 (2022).
  - [7] E. van den Berg, Z. K. Mineev, A. Kandala, and K. Temme, *Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors*, *Nat. Phys.* **19**, 1116 (2023).
  - [8] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. van den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme, and A. Kandala, *Evidence for the utility of quantum computing before fault tolerance*, *Nature (London)* **618**, 500 (2023).
  - [9] *Quantum error correction below the surface code threshold*, *Nature (London)* **638**, 920 (2025).

- [10] K. Tsubouchi, H. Kwon, L. Jiang, and N. Yoshioka, *Quantum advantages for syndrome-aware noisy logical observable estimation*, [arXiv:2603.05145](#).
- [11] H.-Y. Huang, R. Kueng, and J. Preskill, *Predicting many properties of a quantum system from very few measurements*, *Nat. Phys.* **16**, 1050 (2020).
- [12] S. Chen, J. Cotler, H.-Y. Huang, and J. Li, *Exponential separations between learning with and without quantum memory*, in *Proceedings of the 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2022), pp. 574–585.
- [13] J. Haah, R. Kothari, R. O'Donnell, and E. Tang, *Query-optimal estimation of unitary channels in diamond distance*, in *Proceedings of the 2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2023), pp. 363–390.
- [14] H.-Y. Hu, M. Ma, W. Gong, Q. Ye, Y. Tong, S. T. Flammia, and S. F. Yelin, *Ansatz-free Hamiltonian learning with Heisenberg-limited scaling*, *PRX Quantum* **6**, 040315 (2025).
- [15] H.-Y. Huang, Y. Tong, D. Fang, and Y. Su, *Learning many-body Hamiltonians with Heisenberg-limited scaling*, *Phys. Rev. Lett.* **130**, 200403 (2023).
- [16] Y. Wu, Y. Zhang, C. Wang, and X. Yuan, *Hamiltonian dynamics learning: A scalable approach to quantum process characterization*, [arXiv:2503.24171](#).
- [17] H.-Y. Huang, R. Kueng, and J. Preskill, *Information-theoretic bounds on quantum advantage in machine learning*, *Phys. Rev. Lett.* **126**, 190505 (2021).
- [18] Z.-H. Liu, R. Brunel, E. E. B. Østergaard, O. Cordero, S. Chen, Y. Wong, J. A. H. Nielsen, A. B. Bregnsbo, S. Zhou, H.-Y. Huang, C. Oh, L. Jiang, J. Preskill, J. S. Neergaard-Nielsen, and U. L. Andersen, *Quantum learning advantage on a scalable photonic platform*, *Science* **389**, 1332 (2025).
- [19] S. Chen, S. Zhou, A. Seif, and L. Jiang, *Quantum advantages for Pauli channel estimation*, *Phys. Rev. A* **105**, 032435 (2022).
- [20] M. Subramanian, H. Kwon, and L. Jiang, *Quantum channel learning with limited parallel access*, [arXiv:2608.05307](#).
- [21] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, and J. R. McClean, *Quantum advantage in learning from experiments*, *Science* **376**, 1182 (2022).
- [22] S. Chen, W. Gong, and Q. Ye, *Optimal tradeoffs for estimating Pauli observables*, in *Proceedings of the 2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2024), pp. 1086–1105.
- [23] E. Coroi and C. Oh, *Exponential advantage in continuous-variable quantum state learning*, [arXiv:2501.17633](#).
- [24] M. C. Caro, *Learning quantum processes and Hamiltonians via the Pauli transfer matrix*, *ACM Trans. Quantum Comput.* **5**, 1 (2024).
- [25] S. Chen, C. Oh, S. Zhou, H.-Y. Huang, and L. Jiang, *Tight bounds on Pauli channel learning without entanglement*, *Phys. Rev. Lett.* **132**, 180805 (2024).
- [26] A. Seif, S. Chen, S. Majumder, H. Liao, D. S. Wang, M. Malekakhlagh, A. Javadi-Abhari, L. Jiang, and Z. K. Mineev, *Entanglement-enhanced learning of quantum processes at scale*, [arXiv:2408.03376](#).

- [27] M. Kim and C. Oh, *On the fundamental resource for exponential advantage in quantum channel learning*, [arXiv:2507.11089](#).
- [28] C. Oh, S. Chen, Y. Wong, S. Zhou, H.-Y. Huang, J. A. H. Nielsen, Z.-H. Liu, J. S. Neergaard-Nielsen, U. L. Andersen, L. Jiang, and J. Preskill, *Entanglement-enabled advantage for learning a bosonic random displacement channel*, *Phys. Rev. Lett.* **133**, 230604 (2024).
- [29] V. Giovannetti, S. Lloyd, and L. Maccone, *Quantum-enhanced measurements: Beating the standard quantum limit*, *Science* **306**, 1330 (2004).
- [30] V. Giovannetti, S. Lloyd, and L. Maccone, *Quantum metrology*, *Phys. Rev. Lett.* **96**, 010401 (2006).
- [31] V. Giovannetti, S. Lloyd, and L. Maccone, *Advances in quantum metrology*, *Nat. Photon.* **5**, 222 (2011).
- [32] G. Chiribella, G. M. D'Ariano, and P. Perinotti, *Quantum circuit architecture*, *Phys. Rev. Lett.* **101**, 060401 (2008).
- [33] G. Chiribella, G. M. D'Ariano, and P. Perinotti, *Theoretical framework for quantum networks*, *Phys. Rev. A* **80**, 022339 (2009).
- [34] G. Chiribella, *Optimal networks for quantum metrology: Semidefinite programs and product rules*, *New J. Phys.* **14**, 125008 (2012).
- [35] Y. Yang, *Memory effects in quantum metrology*, *Phys. Rev. Lett.* **123**, 110501 (2019).
- [36] A. Altherr and Y. Yang, *Quantum metrology for non-Markovian processes*, *Phys. Rev. Lett.* **127**, 060501 (2021).
- [37] S. Kurdziałek, P. Dulan, J. Majasak, S. Chakraborty, and R. Demkowicz-Dobrzański, *Quantum metrology using quantum combs and tensor network formalism*, *New J. Phys.* **27**, 013019 (2025).
- [38] T. J. Proctor, P. A. Knott, and J. A. Dunningham, *Multiparameter estimation in networked quantum sensors*, *Phys. Rev. Lett.* **120**, 080501 (2018).
- [39] M. Gessner, L. Pezzè, and A. Smerzi, *Sensitivity bounds for multiparameter quantum metrology*, *Phys. Rev. Lett.* **121**, 130503 (2018).
- [40] F. Albarelli, J. F. Friel, and A. Datta, *Evaluating the Holevo Cramér-Rao bound for multiparameter quantum metrology*, *Phys. Rev. Lett.* **123**, 200503 (2019).
- [41] C. Helstrom, *Quantum Detection and Estimation Theory* (Academic Press, New York, 1976).
- [42] S. L. Braunstein and C. M. Caves, *Statistical distance and the geometry of quantum states*, *Phys. Rev. Lett.* **72**, 3439 (1994).
- [43] M. G. Paris, *Quantum estimation for quantum technology*, *Int. J. Quantum Inf.* **7**, 125 (2009).
- [44] J. Liu, H. Yuan, X.-M. Lu, and X. Wang, *Quantum Fisher information matrix and multiparameter estimation*, *J. Phys. A* **53**, 023001 (2020).
- [45] S. Ragy, M. Jarzyna, and R. Demkowicz-Dobrzański, *Compatibility in multiparameter quantum metrology*, *Phys. Rev. A* **94**, 052108 (2016).
- [46] J. S. Sidhu, Y. Ouyang, E. T. Campbell, and P. Kok, *Tight bounds on the simultaneous estimation of incompatible parameters*, *Phys. Rev. X* **11**, 011028 (2021).
- [47] F. Albarelli and R. Demkowicz-Dobrzański, *Probe incompatibility in multiparameter noisy quantum metrology*, *Phys. Rev. X* **12**, 011039 (2022).
- [48] Y. Watanabe, T. Sagawa, and M. Ueda, *Optimal measurement on noisy quantum systems*, *Phys. Rev. Lett.* **104**, 020401 (2010).
- [49] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, *On the Lambert W function*, *Adv. Comput. Math.* **5**, 329 (1996).
- [50] F. W. Olver, *NIST Handbook of Mathematical Functions Hardback and CD-ROM* (Cambridge university press, Cambridge, England, 2010).
- [51] R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, England, 1985).
- [52] D.-J. Zhang and D. M. Tong, *Inferring physical properties of symmetric states from the fewest copies*, *Phys. Rev. Lett.* **133**, 040202 (2024).
- [53] S. Zhou, M. Zhang, J. Preskill, and L. Jiang, *Achieving the Heisenberg limit in quantum metrology using quantum error correction*, *Nat. Commun.* **9**, 78 (2018).
- [54] W. Dür, M. Skotiniotis, F. Fröwis, and B. Kraus, *Improved quantum metrology using quantum error correction*, *Phys. Rev. Lett.* **112**, 080801 (2014).
- [55] G. Arrad, Y. Vinkler, D. Aharonov, and A. Retzker, *Increasing sensing resolution with error correction*, *Phys. Rev. Lett.* **112**, 150801 (2014).
- [56] E. M. Kessler, I. Lovchinsky, A. O. Sushkov, and M. D. Lukin, *Quantum error correction for metrology*, *Phys. Rev. Lett.* **112**, 150802 (2014).
- [57] D. Layden, S. Zhou, P. Cappelaro, and L. Jiang, *Ancilla-free quantum error correction codes for quantum metrology*, *Phys. Rev. Lett.* **122**, 040502 (2019).
- [58] R. Demkowicz-Dobrzański, J. Czajkowski, and P. Sekatski, *Adaptive quantum metrology under general Markovian noise*, *Phys. Rev. X* **7**, 041009 (2017).
- [59] S. Zhou and L. Jiang, *Asymptotic theory of quantum channel estimation*, *PRX Quantum* **2**, 010343 (2021).
- [60] I. Rojko, D. Layden, P. Cappelaro, J. Home, and F. Reiter, *Bias in error-corrected quantum sensing*, *Phys. Rev. Lett.* **128**, 140503 (2022).
- [61] Q. Zhuang, J. Preskill, and L. Jiang, *Distributed quantum sensing enhanced by continuous-variable error correction*, *New J. Phys.* **22**, 022001 (2020).
- [62] H. Kwon, U. R. Fischer, S.-W. Lee, and L. Jiang, *Restoring Heisenberg scaling in time via autonomous quantum error correction*, [arXiv:2504.13168](#).
- [63] H. Kwon, K. Tsubouchi, C.-T. Chu, and L. Jiang, *Criteria for unbiased estimation: Applications to noise-agnostic sensing and learnability of quantum channel*, [arXiv:2503.17362](#).
- [64] M. Tsang, F. Albarelli, and A. Datta, *Quantum semiparametric estimation*, *Phys. Rev. X* **10**, 031023 (2020).
- [65] S. Chen, W. Gong, and S. Zhou, *Instance-optimal high-precision shadow tomography with few-copy measurements: A metrological approach*, [arXiv:2602.04952](#).
- [66] A. C. Berry, *The accuracy of the Gaussian approximation to the sum of independent variates*, *Trans. Am. Math. Soc.* **49**, 122 (1941).
- [67] R. Durrett, *Probability: Theory and Examples* (Cambridge university press, Cambridge, England, 2019), Vol. 49.
- [68] W. Feller, *An Introduction to Probability Theory and its Applications* (John Wiley & Sons, New York, 1991), Vol. 2.

- 
- [69] Z. W. Birnbaum, *An inequality for mill's ratio*, *Ann. Math. Stat.* **13**, 245 (1942).
- [70] R. D. Gordon, *Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument*, *Ann. Math. Stat.* **12**, 364 (1941).
- [71] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, 3rd ed. (Oxford University Press, Oxford, 2001).
- [72] J. P. Mills, *Table of the ratio: Area to bounding ordinate, for any portion of normal curve*, *Biometrika* **18**, 395 (1926).
- [73] V. Spokoiny, *Parametric estimation. Finite sample theory*, *Ann. Stat.* **40**, 2877 (2012).
- [74] A. W. Van der Vaart, *Asymptotic Statistics* (Cambridge university press, Cambridge, England, 2000), Vol. 3.
- [75] I. A. Ibragimov and R. Z. Has' Minskii, *Statistical Estimation: Asymptotic Theory* (Springer Science & Business Media, New York, 2013).