How much can we learn from quantum random circuit sampling?

Tudor Manole,[1,*] Daniel K. Mark,[2,*] Wenjie Gong,[2] Bingtian Ye,[2] Yury Polyanskiy,[1,3,†] and Soonwon Choi[2,‡]

[1] *Statistics and Data Science Center, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *Center for Theoretical Physics—a Leinweber Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[3] *Department of EECS, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

(Dated: October 14, 2025)

Benchmarking quantum devices is a foundational task for the sustained development of quantum technologies. However, accurate *in situ* characterization of large-scale quantum devices remains a formidable challenge: such systems experience many different sources of errors, and cannot be simulated on classical computers. Here, we introduce new benchmarking methods based on random circuit sampling (RCS), that substantially extend the scope of conventional approaches. Unlike existing benchmarks that report only a single quantity—the circuit fidelity—our framework extracts rich diagnostic information, including spatiotemporal error profiles, correlated and contextual errors, and biased readout errors, without requiring any modifications of the experiment. Furthermore, we develop techniques that achieve this task without classically intractable simulations of the quantum circuit, by leveraging *side information*, in the form of bitstring samples obtained from reference quantum devices. Our approach is based on advanced high-dimensional statistical modeling of RCS data. We sharply characterize the information-theoretic limits of error estimation, deriving matching upper and lower bounds on the sample complexity across all regimes of side information. We identify surprising phase transitions in learnability as the amount of side information varies. We demonstrate our methods using publicly available RCS data from a state-of-the-art superconducting processor, obtaining *in situ* characterizations that are qualitatively consistent yet quantitatively distinct from component-level calibrations. Our results establish both practical benchmarking protocols for current and future quantum computers and fundamental information-theoretic limits on how much can be learned from RCS data.

I. INTRODUCTION

Quantum information processing has made remarkable progress in recent years, reaching major milestones such as the demonstration of beyond-classical computational tasks [1–4], as well as the realization of quantum error correction and early fault-tolerant operations [3, 5–10]. As quantum devices become increasingly advanced, we also need improved methods to characterize and benchmark them.

Quantum processors experience a variety of errors, including coherent errors [11–13], errors that vary over space and time [1], correlated multi-qubit errors [14, 15], leakage errors [16–18], and *contextual* errors whose presence depend on the choice of earlier operations. This diversity of error types reflects the diversity of physical mechanisms for error in quantum devices. Examples include excitation into higher transmon levels [16], spontaneous emission of photons from atoms [17, 18], and slow ringdown [19] or fluctuations in control pulses [11]. These various types of errors must be identified and quantified in the effort to improve quantum hardware components, their interconnects, and systems architecture [19]. In this regard, traditional methods often fall short in accurately characterizing complex devices. In most common practices, a quantum system is characterized component-by-component and operation-by-operation, separately, rather than when they work together as in a full quantum circuit [1]. In such approaches, certain types of errors may be missed or incorrectly estimated [20].

Cross-entropy benchmarking (XEB) is the state-of-the art approach to characterize large scale quantum devices [21]. The XEB fidelity is a proxy for the quantum fidelity and provides a single-number summary of circuit performance. It is remarkably versatile, being applicable across different hardware platforms, both for physical and logical circuits, and beyond the ideal setting of deep random unitary circuits [2, 3, 22, 23]; hence it is one of a few industry-standard approaches for benchmarking quantum systems [24–27], and has seen wide adoption [1–3, 23, 28–31]. However, cross-entropy benchmarking has two major limitations. First, it aggregates all types of errors into a single metric—the state infidelity—thereby obscuring detailed information needed to guide improvements in quantum hardware. Second, it relies on classical simulation of the ideal circuit output, which limits its applicability to relatively small system sizes.

In this work, we present new benchmarking methods based on random circuit sampling (RCS). Our approach requires no modification in experiments, as it relies on exactly the same data needed for XEB. However, our methods significantly extend the scope of the XEB fidelity, largely addressing the aforementioned limitations and providing detailed information about noise. Explicitly, starting from the bitstring data obtained from noisy RCS, our method produces an elaborate report that contains a rich set of information such as the estimated circuit fidelity, spatiotemporal profiles of single- or two-qubit errors, correlated or contextual errors, bi-

* These authors contributed equally to this work.
† yp@mit.edu
‡ soonwon@mit.edu

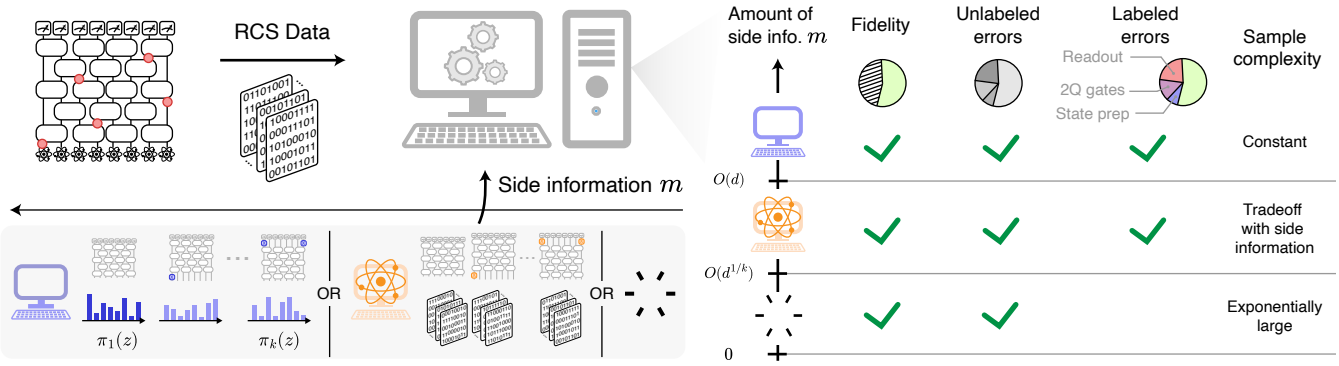

Figure 1. Overview of this work. We develop methods to learn many noise parameters from random circuit sampling (RCS) bitstring data. Our protocol makes use of side information, which can be the classically computed bitstring distribution, or samples of these distributions obtained from a reference quantum computer. Our analysis includes the special case where no side information is available — even in this case, error rates can be learned given enough RCS data. Our method allows extracting more information than previously known benchmarking methods: in addition to the state fidelity, we can estimate the error rates for many types of errors, including state preparation errors, correlated multi-qubit errors, contextual errors which depend on previous gates applied, and readout errors. We find a phase diagram that dictate the hardness, and types of information that can be learned from RCS data, as a function of the amount of side information available. The sample complexity in each regime is analyzed.

ased readout errors, and, in certain cases identifies the dominant physical processes behind the errors. This report can be made in a computationally efficient manner, provided sufficient amount of *side information* that describe the expected sampling distributions in the absence of unwanted errors. We consider three different regimes of side information: A) maximal side information where the expected output distribution is available from explicit classical computation, B) partial side information, where the expected distributions can be inferred from $m < \infty$ bitstring samples generated from a clean reference quantum computer, and C) no side information where there is no reference data (Fig. 1). Formally, A) and C) correspond to $m = \infty$ and $m = 0$, respectively. The regimes B) and C) are increasingly relevant for demonstrations of beyond-classical circuit sampling.

Our approach to benchmarking is based on improved modeling of noisy random circuit output, building upon a model introduced by [32]. We show that this modeling approach adequately resolves the different types and spacetime positions of errors, using advanced statistical data-processing algorithms of the high-dimensional measurement data. In statistical terms, we describe RCS data as arising from a mixture of random high-dimensional probability distributions, representing the various error channels present in the random quantum circuit. We provide a detailed description of our statistical setup, which is accessible to statisticians, in Appendix B.

Our algorithms require more RCS samples when there is less side-information available. Interestingly, we discover *phase transitions* in the sample complexity of error learning, which allow us to precisely characterize the regimes in which side information is beneficial. In all settings, we prove that our methods achieve optimal sample complexity, saturating information-theoretic lower bounds. These sample complexity bounds provide valuable insight into the scope of possible applications of our benchmarking methods. As an example, our results imply that one can benchmark the full execution of an RCS experiment on a quantum processor of approximately 50 qubits in an efficient manner, both in memory and computation, given sufficient side information from a reference quantum computer.

Finally, we apply our methods to existing, publicly available RCS data [1] to characterize the diverse sources of error that arise in a state-of-the-art quantum processor. Unlike earlier error characterizations of the quantum processor, our report (Fig. 4) provides *in-situ* information about the errors experienced in full operation of the circuit. The extracted error rates and breakdown into different sources are consistent with the expected behavior.

II. OVERVIEW

Our method builds upon the operating principle of cross-entropy benchmarking (XEB) [21]: that the output of a random circuit is a highly entangled wavefunction $|\psi\rangle$ which is a superposition over all N-bit strings. Measuring this state in the computational basis amounts to sampling from the probability distribution $\pi_1(z) \equiv |\langle z|\psi\rangle|^2$ over $d = 2^N$ possible bitstrings. In a typical circuit, this probability distribution is highly complex: its individual entries $\pi_1(z)$ fluctuate strongly among bitstrings z, and form a many-body *speckle pattern* that is essentially unique to the state $|\psi\rangle$ (Fig. 1) in practical

settings[1]. Moreover, for a sufficiently deep circuit, $|\psi\rangle$ is well-approximated as a Haar-random state [33] and hence $\pi_1(z)$ satisfies universal statistical properties on the distribution of $\{\pi_1(z)\}$ values, known as the Porter-Thomas distribution [1, 34]. This universality provides the foundation for the XEB fidelity.

The XEB fidelity is an approximation of the quantum fidelity $\langle\psi|\rho|\psi\rangle$ between the ideal target state $|\psi\rangle$ produced by a programmed quantum circuit and the experimentally prepared mixed state ρ. It compares experimental bitstrings against the ideal distribution π_1 to estimate the fraction of samples drawn from π_1. The *white noise model*, often evoked to explain the XEB fidelity, makes this notion of the "fraction" concrete. It posits that the experimental distribution $p(z) \equiv \langle z|\rho|z\rangle$ is of the form

$$\text{White noise model: } p(z) = F\pi_1(z) + (1-F)/d. \quad (1)$$

Experimental samples are therefore drawn from $\pi_1(z)$ with probability F and from the uniform distribution $\pi_{\text{wh}}(z) \equiv 1/d$ with probability $1-F$. This relation arises, for example, when $\rho = F|\psi\rangle\langle\psi| + (1-F)I_d/d$ is the globally depolarized state (where I_d/d is the maximally mixed state), but can also arise from the combined effect of many errors throughout the circuit [35]. The operational meaning of F is the probability that an experiment successfully executes the circuit with no error, hence estimating the quantum fidelity.

In this work, we adopt a more refined *k-component model* [32]

$$k\text{-component model: } p(z) = \sum_{i=1}^{k} c_i \pi_i(z)\ . \quad (2)$$

Each term $\pi_i(z)$ is the distribution associated with each physical error pattern i. For example, during one execution of a deep circuit, a particular qubit at site a may experience a Pauli X error at a specific time t. Alternatively, a pair of qubits at site a and b experience Pauli X and Y errors at times t_1 and t_2, respectively. We imagine all such patterns of errors ("events") that may reasonably occur in the circuit and enumerate them by the index i. We identify the special case $i = 1$ with the perfect execution of the circuit.

Formally, we can understand that the index i enumerates over the ensemble of *quantum trajectories* obtained from an unraveling of error channels [36]. Each trajectory is associated with a pure wavefunction evolving under the programmed unitary circuit interspersed by error (Kraus) operators at specific spacetime locations. Rapid scrambling of random unitary circuits [37, 38] ensure that such trajectory states are, with high probability, also Haar-random and hence their measurement distribution $\pi_i(z)$ can be approximated as vectors independently sampled from the Porter-Thomas distribution [11, 32].

In the simplest version of our protocols, we essentially perform cross-entropy benchmarking on each distribution $\pi_i(z)$ to estimate its coefficient c_i, representing the probability of the particular error event i. We consider a total of k error events. Advanced statistical algorithms allow us to utilize the information provided by bitstring measurements and simultaneously and efficiently estimate all c_i. This enables *in-situ* characterization of errors in a quantum circuit, complementing prevailing approaches of error characterization by single- and two-qubit experiments. Error rates may differ between these isolated experiments and full operation of a quantum circuit when all components are simultaneously in operation. This also enables the characterization of complex error types such as correlated and contextual errors.

For large quantum systems, obtaining exact knowledge of $\pi_i(z)$ for many different error patterns is computationally intractable. To this end, we study the error estimation task when $\pi_i(z)$ is known only partially. Specifically, we envision that information about $\pi_i(z)$ is obtained by sampling from a *reference* quantum computer that prepares the ideal state $|\psi\rangle$ and any of the noise trajectory states. Such states can be obtained, for example, by stringent quantum error detection [3, 17, 18]. Drawing inspiration from a related line of work in the statistical literature [39–41], we dub this *side information*, which enables our use of the experimental data for parameter estimation. We quantify the amount of side information as the number of measurements m of each trajectory state, yielding bitstrings W_{im} sampled from each π_i. By comparing W_{im} and bitstring data from noisy RCS, we produce the same kind of benchmark report.

We find phase transitions in sample complexity as a function of the amount of side information (Fig. 1): in the full information, $m = \infty$ phase (which includes the case where π_i are classically simulated), the c_i's can be estimated with a number of samples independent of system size. As m decreases and crosses the phase boundary at $m = O(d)$ (*i.e.*, sublinear in d), we enter the partial side-information phase which features a sample complexity tradeoff: the sample complexity is set by the *product* $n \times m$ of the numbers of experimental and side-information samples. Estimation is feasible as long as $nm \geq d \log k$. Surprisingly, even in the $m = 0$ limit with no side information, estimation is still possible by making use of the universal Porter-Thomas properties of each π_i. This transition occurs at $m = O(d^{1/k})$ which depends on the number of errors k considered in our model. In this phase, only the *unlabeled* $\{c_i\}$ can be estimated, i.e. we can determine the values of the c_i's, such as the largest c_i, but cannot assign the indices i to each value. Detailed expressions for the sample complexities are summarized in Table I.

Across all regimes, the key properties we utilize are the universal properties of high-dimensional Porter-Thomas (or Dirichlet random) distributions π_i. These proper-

[1] We note that states of the form $\exp(i\theta P)|\psi\rangle$ (for some Pauli operator P consisting of Z operators) displays the same speckle pattern $\pi_1(z)$. We are indifferent to those cases since they do not change computational outcomes.

ties *concentrate*: random instances are close to the average value with (exponentially) small fluctuations. For example, two random distributions have fixed overlap $\sum_z \pi_i(z)\pi_j(z) = (1+\delta_{ij})/d + O(d^{-3/2})$ and hence are in some sense approximately orthonormal. This bilinear structure is tremendously helpful: the product $\sum_z \pi_i(z)p(z)$ can be estimated with only a few samples from a distribution $p(z)$, far fewer than the (exponentially large) number needed to accurately estimate each entry $p(z)$. As long as the number of signals k is less than the dimension d, k different products $\sum_z \pi_i(z)p(z)$ can be straightforwardly distinguished with minimal overhead.

In the partial side information phase, we utilize a *collision estimator* which counts the number of bitstrings z seen in both the RCS and reference data. The Porter-Thomas nature of the distributions π_i ensures that they fluctuate around the uniform distribution, and in this regime, the classical birthday paradox ensures that collisions begin to occur once the *product* of sample sizes $n \times m$ exceeds the dimension d [42]. This is precisely the threshold at which our methods can reliably estimate c_i, up to a logarithmic correction in k.

Even in the absence of *any* side information ($m = 0$), the typicality of high-dimensional Dirichlet random distributions means that information about c is present in the measurement data, independent of the distributions π_i, and hence not requiring knowledge of them. As a simple example, under Eq. (4), the sum-of-squares $\sum_i c_i^2$ is well estimated by the collision probability $d\sum_z p(z)^2 - 1$ [11, 23]. This, and higher moments of $p(z)$, can be estimated with enough RCS data.

The high-dimensional nature of RCS data—in which the Hilbert space dimension d significantly exceeds the sample size n—is both a blessing and a curse throughout our analysis. This high-dimensionality enables the universal behavior of random Porter-Thomas distributions, leading to particularly simple and practical algorithms that generalize XEB, while on the other hand placing lower bounds on the sample complexity of error characterization which sometimes grow exponentially in system size. From a technical lens, our analysis leverages tools from high-dimensional statistical theory [43], though in the setting of count-based data which is heteroscedastic in nature, unlike the more classical setting of Gaussian additive noise.

We show that our rates of estimation are optimal. We establish information-theoretic lower bounds on the sample complexity of the inference task. We establish matching upper bounds by explicitly presenting optimal statistical estimators, sharply resolving the question of parameter estimation with RCS data.

We proceed to apply our methods to synthetic data obtained from numerical simulations as well as publicly available data produced in a quantum experiment [1]. We confirm that we can identify time-varying error rates and non-local correlated errors. From the real-world data, we successfully estimate state preparation, rates of errors affecting single- and two-qubits, as well as biased readout errors, resolved on each qubit. Our results are qualitatively consistent with anticipated values, but, crucially, our approach estimates those error rates *in-situ* whereas the previously reported values rely on data obtained from separate experiments.

Our results imply the possibility of benchmarking quantum devices in the *beyond-classical regime*, consisting of up to 50 qubits, using only modest classical resources. This follows from our results by choosing the sample sizes to scale as $n = m = \sqrt{d} = 2^{25}$, an amount which can be handled by our classical algorithms without overwhelming memory or computational overhead. On the other hand, our lower bound implies that we cannot do better than this: there is a fundamental information-theoretic limit on how much we can learn from RCS data. In order to circumvent our lower bound, one must use quantum circuits with additional structure, such as those in mirror benchmarking protocols [26, 44, 45] or random Clifford circuits [46, 47].

Related Work. As mentioned previously, the work of [32] was the first to propose model (4), and statistical estimators for c when $m = \infty$. When $m = 0$, Ref. [1] introduced a method called speckle XEB, for estimating the XEB fidelity under the white noise model (1) without knowledge of π_1, using the empirical second moment of the bitstring data (subsequently termed "self-XEB" in Ref. [23]). Their procedure was generalized by the work of [11] to model (4), who showed that higher-order empirical moments of the bitstring data can be used to estimate the moments of the unordered vector c. They also derived an upper bound on the sample complexity of estimating the second moment of the vector c, which can be viewed as a precursor to our sample complexity bounds to come, for the special case $k = 2$.

Estimating the overlap fidelity between two quantum states prepared on separate quantum computers has been studied as "cross-platform verification" in Refs. [48, 49], and specifically in the context of randomized measurements in Ref. [50]. However, not much is understood about its sample complexity, and the task of learning multiple parameters in a cross-platform approach has not been explored before our work.

III. LEARNING FROM BITSTRINGS

A. Random circuit sampling as a statistical mixture model

Having provided a high-level overview of our results, in this section we begin our technical discussion. Our benchmarking methods are developed for RCS data [21], in which a quantum processor executes a circuit obtained by composing randomly-sampled single- and two-qubit gates, a popular benchmark in the field with several demonstrations [1, 3, 28–31, 51, 52].

Random circuit sampling has the advantage of being an unbiased measure of quality of a quantum device:

its underlying gates are randomized and hence the measurement outcomes are not biased towards one particular type of error, nor are they tailored to a particular circuit, which may have highly structured outcomes. In this respect, it is similar to randomized benchmarking [44, 46, 53–56] as well as the suite of tools known as the *randomized measurement toolbox* [57].

Originally motivated as a quantum advantage demonstration, RCS has since become a general-purpose and widely-used tool for benchmarking quantum hardware [58]. Furthermore, random quantum circuits are good approximations for Haar-random unitaries [33], which makes them useful for quantum information tasks [59] including state learning [57, 60] and random-number generation, as well as for the study of questions in basic science, such as quantum chaos and thermalization [37, 61, 62].

In an RCS experiment, an N-qubit random circuit produces an output state $|\psi\rangle \equiv U|\psi_0\rangle$, which is then measured in the computational basis. Each experiment yields a random N-bitstring $z \in \{0,1\}^N$, whose probability distribution is given by $\pi_1(z) \equiv |\langle z|\psi\rangle|^2$ in the noiseless case. In practice, there are uncontrolled errors, and a noisy physical quantum device instead transforms $|\psi_0\rangle$ into a mixed state ρ, with a corresponding measurement distribution $p(z) = \langle z|\rho|z\rangle$.

The most widely-used technique for analyzing RCS data is the (linear) cross-entropy benchmark (XEB) [1, 21]. The XEB estimates the many-body fidelity $F \equiv \langle\psi|\rho|\psi\rangle$ by comparing the experimental distribution $p(z)$ with the ideal one $\pi_1(z)$:

$$F_{\mathrm{XEB}} = d \sum_{z\in\{0,1\}^N} p(z)\pi_1(z) - 1. \tag{3}$$

Experimental samples provide an *empirical* estimate of $p(z)$ which furnishes an unbiased estimator $\widehat{F}_{\mathrm{XEB}}$ (Appendix B). In turn, the XEB furnishes an accurate approximation of the quantum fidelity, $F_{\mathrm{XEB}} \approx F$ in sufficiently deep random circuits with local noise [63, 64].

The above relation is justified by two properties. First, random unitary circuit dynamics results in distributions π_1 with highly typical properties. This is referred to by the *Porter-Thomas* distribution [21, 34] which governs the distribution of values $\pi_1(z)$. Mathematically, this is equivalent to assuming that π_1 is a random probability vector sampled from the *Dirichlet distribution*, i.e. uniformly random on the probability simplex (see condition **(PT)** below for a formal definition). Second, one needs to make an assumption about the bitstring output of the "noisy part" of the state. The simplest such model is the white noise model [Eq. (1)]. However, the XEB remains an accurate estimate of the fidelity even in the more general situation, repeated here: we assume that the quantum device may experience $k-1$ different incoherent error patterns, and thus that the random state ρ outputs bitstrings according to the bitstring distribution:

$$p_c(z|\Pi) = \sum_{i=1}^{k} c_i \pi_i(z), \quad z \in \{0,1\}^N, \tag{4}$$

where π_1 is the ideal *random* probability distribution, $\pi_2, \ldots, \pi_k$ are the *random* probability distributions of k different incoherent error sources in the circuit, and $c = (c_1, \ldots, c_k)$ is the corresponding vector of probabilities ("error weights"). See Refs. [11, 32] for similar models. We collect the distributions π_i into a single matrix $\Pi \in \mathbb{R}^{k\times d}$ with entries $\Pi_{ij} = \pi_i(z_j)$, where the index i denotes the error type, and j denotes the bitstring index, and explicitly highlighted the dependence of p_c on Π. The matrix Π is random and depends on the choice of random circuit and on the errors in the model (see Eq. (6) below). In statistical language, we recognize model (4) as a mixture model consisting of k components, the first of which corresponds to the ideal bitstring distribution π_1, which occurs with a probability c_1 that can be viewed as an analogue of the XEB fidelity F. The remaining terms of the mixture model correspond to $k-1$ noise sources π_i, each occurring with probability c_i.

Eq. (4) is not only more physically realistic, it also allows for learning beyond the single-number summary of the circuit infidelity provided by XEB. Learning the coefficients c_i in our model provides detailed information about the error processes that contribute to this infidelity.

Physically, Eq. (4) arises from the following model of the noisy state

$$\rho = \mathcal{R}_J \circ \mathcal{U}_J \circ \cdots \circ \mathcal{U}_1 \circ \mathcal{R}_0[|\psi_0\rangle\langle\psi_0|] \tag{5}$$

where $\mathcal{R}_i$ denotes the error channel at circuit layer i and $\mathcal{U}_i[\,\cdot\,] \equiv U_i[\,\cdot\,]U_i^\dagger$ denotes the ideal quantum unitary acting on layer i. Each error channel consists of a number of physical errors, denoted by Kraus operators K_ℓ [36]:

$$\mathcal{R}_i[\rho] = \sum_\ell \Gamma_\ell^{(i)} K_\ell \rho K_\ell^\dagger, \tag{6}$$

with physical error rates $\Gamma_\ell^{(i)}$ that can depend both on error type, location, and time (layer).

Eq. (4) is related to (6) by the following: Each distribution π_i is associated with a particular *trajectory* $(K_{\ell_0}, K_{\ell_1}, \cdots, K_{\ell_J})$ that denotes a sequence of Kraus operators. For instance, the ideal distribution π_1 corresponds to the trajectory where all $K_{\ell_i} = \mathbb{I}$, and c_1 is the probability that no error occurred. Other trajectories include those where one K_{ℓ_i} is non-trivial, indicating the occurrence of an error at layer i, of type ℓ_i. Its corresponding distribution is given by

$$\pi_{(\ell_i,i)}(z) = |\langle z|U_J \cdots K_{\ell_i} U_i \cdots U_1|\psi_0\rangle|^2. \tag{7}$$

When the operators K_{ℓ_i} are unitary, e.g. for a Pauli error channel, equation (7) is a probability distribution (non-negative and summing to 1), and due to the operator

spreading [38] in the random circuit, each $\pi_{(\ell_i,i)}(z)$ is an independent Dirichlet-random distribution (see condition **(PT)** below).

This condition is necessary for the statistical model (4) and our theoretical analysis, but will not be necessary for the analysis of real data in Section V: our estimators are robust to deviations from Assumption **(PT)**.

Current XEB approaches typically require complete knowledge of the ideal distribution π_1, which is extremely challenging to classically simulate when the system size N is approximately greater than 30. In practice, a typical workaround is to estimate the XEB fidelity based on patches of disconnected circuits, or with specially structured circuits that can be simulated [1]. However, such methods require dedicated experiments and are not guaranteed to provide an accurate estimate of the global circuit fidelity. Relaxing the assumption that π_1 is classically computable not only addresses practical needs, it also defines a theoretically rich statistical problem. Fixing notation we will use in the rest of this work, we denote the bitstring measurements ("RCS data") as i.i.d. samples

$$Z_1, \ldots, Z_n \,\big|\, \Pi \sim p_c(\cdot|\Pi), \tag{8}$$

where we have explicitly highlighted the dependence of p_c on the random circuit realization. This determines the matrix Π, a deterministic function of the random choice of circuit, which we equivalently treat as a random variable in itself **(PT)**. We will frequently summarize these measurements in terms of the empirical counts $Y_j = \sum_{i=1}^n \delta_{Z_i,z_j}$, indexed by $j = 1, \ldots, d$. For example, the statistical estimator for the XEB fidelity is simply $\widehat{F}_n = (d/n)\sum_{j=1}^d Y_j \pi_1(z_j) - 1$, arising from the approximation $p(z_j) \approx Y_j/n$.

We additionally assume that the practitioner has access to *side information* in the form of m bitstring samples drawn from a reference quantum computer which perfectly implements the ideal circuit π_1, and any of the noisy circuits π_i:

$$W_{i1}, \ldots, W_{im} \,\big|\, \Pi \sim \pi_i, \quad i = 1, \ldots, k. \tag{9}$$

The parameter m allows us to systematically analyze different classes of protocols. On one extreme, when $m = \infty$, we interpret the matrix Π as being perfectly known. This corresponds to the conventional situation in which the bitstring distributions can be classically simulated, as in the XEB setup. Meanwhile, the $m = 0$ limit indicates that no side information is given (see Refs. [11, 23] for earlier work in this limit). This regime is particularly relevant for large system sizes and deep circuits, where no classical simulation or reference quantum computation is viable. Even in this situation, our benchmarking algorithms provide nontrivial information about the characteristics of noise. Our information-theoretic phase transitions indicate that the boundaries of these two regimes occur at $m \geq d$ and $m \leq d^{1/k}$, respectively. When m lies between these two extremes, a reference quantum computer provides partial *side information* about Π in the form of samples from the ideal distribution π_1, and all noisy distributions π_i, and we develop algorithms which efficiently leverage this side information.

B. Estimators

Given n samples from a distribution of the form Eq. (4), our task is to estimate the error weights $c = (c_1, \ldots, c_k)$, with the aid of m samples of side information that give us knowledge about Π.

How might it be possible to estimate a large number k of parameters from a single realization of a random circuit? The key is that our data is high-dimensional: they are samples drawn from a $d = 2^N$ dimensional probability distribution $p(z)$. For a collection of k different circuits (here representing the original circuit with injected errors), it is highly likely that their output distributions are linearly independent. In other words, in a high-dimensional space, different errors distort the output distribution in different ways and hence can be distinguished in the measurement data.

We develop several estimators to estimate the parameter vector $c = (c_1, \ldots, c_k)$, suitable suitable for various regimes of side information m. We discuss several of these estimators in what follows, deferring a more complete discussion to Appendix B, including further discussion of related statistical literature.

1. *Regime A: Classical Simulation ($m = \infty$)*

In the simplest case with classically-computed side information, the matrix Π is known to the practitioner. In this regime, we estimate the parameter c_i in two steps. We first observe that the products $\zeta_i = \sum_z \pi_i(z) p_c(z|\Pi)$ satisfy

$$\zeta_i = \sum_\ell c_\ell \sum_z \pi_i(z)\pi_\ell(z) = \frac{1 + c_i}{d} + O(d^{-3/2}), \tag{10}$$

as a result of the concentration of the Porter-Thomas rows of Π. One can form unbiased estimators $(1/n)\sum_{j=1}^d Y_j \pi_i(z_j)$ of these products, which leads to a first simple estimator of c:

$$\widehat{c}_i^{\,\mathrm{XEB}} = \frac{d}{n}\sum_{j=1}^d Y_j \pi_i(z_j) - 1, \quad i = 1, \ldots, k. \tag{11}$$

This estimator was first proposed by [32, Eq. (5.1)]. We refer to the vector $\widehat{c}^{\,\mathrm{XEB}}$ as the (generalized) XEB estimator. Much like the XEB fidelity estimator $\widehat{F}_n$, this generalized XEB estimator achieves a sample complexity which does not depend on the Hilbert space dimension d. It does, however, depend linearly on the number of errors k, and this dependence can be mitigated by appropriate regularization. For example, given an appropriate tuning parameter $\lambda > 0$, we will show that the

hard-thresholded [65] XEB estimator

$$\widehat{c}_i^{\mathrm{HT}} = \begin{cases} \widehat{c}_i^{\mathrm{XEB}}, & \widehat{c}_i^{\mathrm{XEB}} > \lambda, \\ 0, & \text{otherwise}, \end{cases} \quad i = 1, \ldots, k, \tag{12}$$

has sample complexity which merely degrades logarithmically with the number of errors k. Roughly speaking, this favorable dependence on k arises from the fact that the vector c has bounded ℓ_1 norm, and is therefore approximately sparse, a fact which is leveraged by estimator (12) [66–68].

As we discuss in Appendix B, the XEB estimator can be viewed as an approximation of the ordinary least squares estimator for performing a linear regression of the histogram Y onto $\Pi^\top$, whereas our model is a multinomial regression model for which the canonical estimator is the maximum likelihood estimator (MLE), defined by

$$\widehat{c}^{\mathrm{MLE}} = \operatorname*{argmax}_{\gamma \in \Delta_k} \sum_{j=1}^{d} Y_j \log\bigl(\Pi_{\cdot j}^\top \gamma\bigr). \tag{13}$$

This estimator was also noted by [32, Eq. (5.2)]. Much like in the RCS literature, where linearization of the XEB fidelity is typically adopted, we do not find that the MLE and XEB estimators differ appreciably due to the small magnitude of Π, and hence the mild heteroscedasticity of the histogram Y (cf. Appendix B). Indeed, we have found that the XEB and MLE estimators perform similarly when the independent Porter-Thomas assumption on the rows of Π holds. The XEB estimator has the disadvantage of not being robust to deviations from this modeling assumption, but it has the advantage of being easily computable even for large system sizes $d = 2^N$, whereas the program (13) can be somewhat more costly to optimize despite its convexity. Another advantage of the MLE is the fact that it is free of tuning parameters, yet still provides accurate estimates when k is large [69]. Roughly speaking, this behavior is due to the restriction of the optimization problem (13) to the simplex, which significantly reduces the volume of the search space despite the potentially large magnitude of k.

2. *Regime B: Partial Side Information ($1 < m < \infty$)*

When the amount of side information m is finite but nonzero, we recommend the following adaptation of the XEB estimator:

$$\widehat{c}_i^{\mathrm{coll}} = \frac{d}{nm} \sum_{\ell=1}^{n} \sum_{r=1}^{m} \delta_{Z_\ell, W_{ir}} - 1, \quad i = 1, \ldots, k, \tag{14}$$

Up to centering and scaling, this estimator consists of counting the number of collisions between the primary bitstring samples $\{Z_\ell\}$ and each of the side samples $\{W_{ir}\}$. Once again, $\widehat{c}^{\mathrm{coll}}$ can be regularized using hard-thresholding, and we will show that the resulting estimator achieves the optimal sample complexity of estimating c.

An important practical benefit of the collision estimator is the fact that its computational complexity scales as $O(k(n+m))$, while its memory complexity scales as $O(k \cdot \min\{n, m\})$, neither of which depend on the Hilbert space dimension d. This estimator can therefore be used in the beyond-classical regime where objects of dimension d cannot easily be stored in the memory of a classical processor. Another practical benefit of this estimator is its linear structure, which allows it to be updated as more bitstring data becomes available, without needing to be recomputed.

As before, it is also natural to consider the maximum likelihood estimator, which is now given by

$$\operatorname*{argmax}_{\gamma \in \Delta_k} \log \int_{\Delta_d^k} \prod_{j=1}^{d} \left((\Pi_{\cdot j}^\top \gamma)^{Y_j} \prod_{i=1}^{k} \pi_{ij}^{V_{ij}} \right) d\Pi, \tag{15}$$

where the integral is taken over the set of $k \times d$ matrices whose rows are constrained to the d-dimensional simplex. Unlike equation (13), this optimization problem is nonconvex. In Appendix B, we develop a heuristic optimization algorithm for this problem by interpreting equation (15) as a partition function which integrates over states Π. Using a mean-field approximation, we derive an algorithm that maximizes the corresponding variational Gibbs free entropy, and consists of iteratively solving the following fixed-point equation with respect to $\gamma \in \Delta_k$:

$$n = \sum_{j=1}^{d} \frac{Y_j S_{ij}}{\sum_{r=1}^{k} S_{rj} \gamma_r}, \quad \text{where } S_{ij} = \exp\{\psi(1 + V_{ij})\}, \tag{16}$$

for $i = 1, \ldots, k$, where ψ denotes the di-gamma function. This iteration is, once again, computable with time and memory complexity that are independent of the Hilbert space dimension d, since the histogram Y is supported on at most n entries. A close analogue of the fixed point equation (16) arises in a statistical method for text analysis known as latent Dirichlet allocation [70]. In that context, it has been argued [71–73] that the mean-field approximation can be significantly improved by working with an analogue of the Thouless-Anderson-Palmer (TAP) free entropy, and we believe it is an interesting avenue of future work to adapt such ideas to our model. We defer further discussion to Appendix B.

3. *Regime C: Blind Source Separation ($m = 0$)*

In the most difficult regime where $m = 0$, model (4) is invariant to relabeling the mixture components. Remarkably, even in this regime we are able to estimate c up to reordering its elements, thus allowing us to identify error patterns without any prior knowledge of how different errors affects the measurement outcome probabilities.

Our strategy in this setting is to leverage the fact that the first k moments $m_j(c) = \sum_{i=1}^k c_i^j$ uniquely characterize c up to ordering [74]. Indeed, this characterization is a consequence of *Newton's identities*, which assert that the coefficients of the polynomial $f(z) = \prod_{i=1}^k (z - c_i)$ can be written solely in terms of $m_1(c), \ldots, m_k(c)$ (cf. Appendix K 2 a). The concetration of high-dimensional Porter-Thomas distributions allows us to identify and estimate the moments $m_j(c)$ directly from the bitstring data. These moment estimators can, in turn, be used to construct an estimator $\widehat{f}$ of the k-degree polynomial f. Our estimator of c, denoted $\widehat{c}^{\mathrm{mom}}$, is then given by the collection of k roots of $\widehat{f}$. We defer a rigorous description of this estimator to Appendix B 3 a.

C. Sample Complexity of Error Learning

We now state our main results regarding the sample complexity of error estimation under model (4). It will be convenient to state our sample complexity bounds in terms of the *minimax estimation risk*, a standard statistical benchmark for quantifying the best possible error that can be achieved by a statistical estimator uniformly over the space Δ_k. Concretely, the minimax risk $\mathcal{M}(n, d, k, m)$ is defined as the smallest achievable upper bound epsilon for the average ℓ_2 distance between the estimated values $\widehat{c}$ and the worst case true value c: $\max_c \mathbb{E}\|\widehat{c} - c\|_2 < \epsilon$. Here, the averaging is taken over randomness of the measured samples from an experiment, reference computers, as well as the Porter Thomas distributions Π (arising from random circuit choices).

Our results are stated under two conditions. First, we impose the following assumptions on the problem parameters n, d, k, m.

(S) Let $\rho = \min\{m/d, 1\}$. Then, there exists an arbitrarily small constant $\gamma > 0$ such that the following assertions hold.

(i) $n^{1+\gamma} \leq d$.

(ii) Either $nm \leq d$ or $nm > d^{1+\gamma}$.

(iii) Either $k \leq \sqrt{n\rho}$, or $k > (\sqrt{n\rho})^{1+\gamma}$.

(iv) Either $k \leq d \leq m$, or $d > m^{1+\gamma}$ and $k^{1+\gamma} < \frac{d}{m}$.

Condition (i) requires the sample size to be smaller than the Hilbert space dimension d, which is the most practical regime for RCS experiments. Once the sample size exceeds d, a number of different approaches based on (approximate) quantum state tomography become available [75]. Conditions (ii) and (iii) are mild assumptions made for ease of exposition; they preclude the problem parameters from falling in narrow regimes where logarithmic corrections appear in our sample complexity bounds, which we do not bother to characterize sharply. Condition (iv) is not needed for our upper bounds, but is used in our lower bounds; this condition allows the number of errors k to be on the same order as d when $m \geq d$, but somewhat limits the magnitude of k when m is smaller than d.

Second, as discussed in Section III, we assume the random unitary circuit is sufficiently deep for the following Porter-Thomas assumption to be met.

(PT) The random matrix $\Pi \in \mathbb{R}^{k \times d}$ has mutually independent rows $\Pi_{i\cdot}$ which follow the flat Dirichlet distribution on the $(d-1)$-dimensional simplex. That is, for each $i = 1, \ldots, k$, one can write $\Pi_{i\cdot} = (X_{i1}, \ldots, X_{id}) / \sum_j X_{ij}$, where the random variables X_{ij} are independent, and follow a Porter-Thomas distribution:

$$\mathbb{P}(X_{ij} > x) = e^{-dx}, \quad \text{for all } x \geq 0.$$

Although Assumption **(PT)** is needed for much of our theory, it is not needed for several of our estimators, as we explore in Appendix B.

In what follows, for any two nonnegative-valued functions f, g, we write $f(x) \asymp_a g(x)$ if there exist constants $C_1, C_2 > 0$, possibly depending on a quantity a, such that $C_1 f(x) \leq g(x) \leq C_2 f(x)$ for all x. Our first main result is stated as follows.

Theorem 1. *Under conditions **(PT)** and **(S)**, we have*

$$\mathcal{M}(n, d, k, m) \asymp_\gamma \min\left\{ \left(\frac{k}{n\rho}\right)^{\frac{1}{2}}, \left(\frac{\log k}{n\rho}\right)^{\frac{1}{4}}, 1 \right\},$$

where $\rho = \min\{m/d, 1\}$.

Theorem 1 reveals several distinct regimes in the sample complexity of error learning. When k is held fixed, and the amount of side information m exceeds d, the

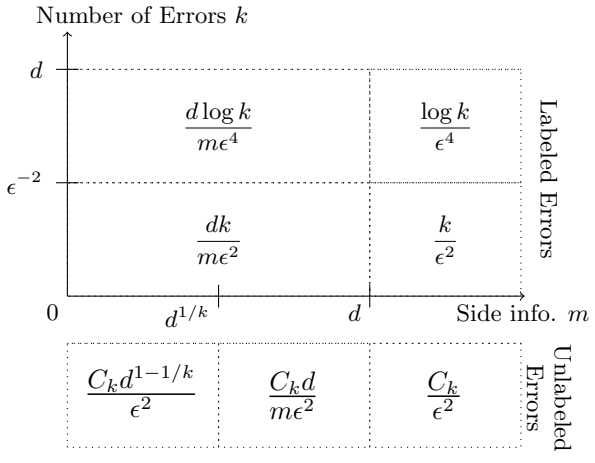

Table I. Heuristic summary of the sample complexities for noise learning as a function of number of errors k and amount of side information m (Theorems 1–2). The sample complexity for labeled errors (top) represents the smallest sample size n for which the vector c can be estimated to accuracy ϵ under the ℓ_2 norm, whereas the sample complexity for unlabeled errors (bottom) is measured for the unordered collection of error rates. C_k denotes a generic constant depending on k.

sample complexity scales as ϵ^{-2}, which is dimension-independent and coincides with the rate of decay of the traditional central limit theorem. The same sample complexity is achievable for estimating the XEB under the white noise model (1) [1], and is enabled by the fact that the matrix Π can be accurately estimated when m is so large. On the other hand, when $m < d$, although the matrix Π is not consistently estimable, the parameter vector c can be consistently estimated so long as the *product* nm exceeds the Hilbert space dimension d. As described above, this product scaling can be interpreted via the birthday paradox.

Theorem 1 also sharply characterizes the dependence of the sample complexity on the number of errors k. When k is large, we find that the sample complexity merely degrades logarithmically in k, at the price of a quadratically slower dependence on the accuracy parameter ϵ. In particular, Theorem 1 implies that when $nm > d \log k$, one can estimate a number of errors which is comparable to the Hilbert space dimension.

Theorem 1 also indicates that the minimax risk approaches a nondecreasing rate of convergence when the amount of side information approaches zero. This is perhaps unsurprising since the parameters c_i are only uniquely defined up to ordering in the absence of side information. Remarkably, however, our next result shows that it is still possible to estimate the *unlabeled* entries of c with a number of samples that scales sublinearly in d. In what follows, we denote by $\mathcal{M}_<(n, d, k, m)$ the *unlabeled* minimax risk, namely the smallest real number $\epsilon \in (0, 1)$ for which there exists an estimator $\widehat{c}$ such that for any $c \in \Delta$, one has

$$\min_{\sigma \in \mathcal{S}_k} \left(\sum_{i=1}^{k} |\widehat{c}_{\sigma(i)} - c_i|^2 \right)^{\frac{1}{2}} \leq \epsilon,$$

where $\mathcal{S}_k$ is the set of permutations on $[k]$. The following result sharply characterizes the unlabeled minimax risk when k is held fixed.

Theorem 2. *Under conditions* ***(PT)*** *and* ***(S)****, we have*

$$\mathcal{M}_<(n, d, k, m) \asymp_{k,\gamma} \frac{1}{\sqrt{n}} \cdot \begin{cases} \sqrt{d^{1-\frac{1}{k}}}, & 0 \leq m < d^{1/k}, \\ \sqrt{d/m}, & d^{1/k} \leq m < d, \\ 1, & d \leq m < \infty. \end{cases}$$

Theorem 2 shows that, even in the absence of side information, the ordered vector c can be consistently recovered when $n \geq d^{\frac{k-1}{k}}$. Although this rate degrades exponentially in the system size, its exponent is sublinear, contrary to tomographic methods which typically suffer from superlinear exponents for recovery of the full underlying quantum state ρ [75]. This gap can make a significant difference in practice; for instance, if one adopts a two-component model with $k = 2$, akin to the white noise model (1), then Theorem 2 shows that the fidelity can be recovered with only $\sqrt{d}$ samples, without any information about the bitstring distributions π_1 and π_2. This observation is consistent with the past work of [11], which indicated that the second moment of c can be recovered with $\sqrt{d}$ samples.

It would be natural to expect that any amount of side information m would improve the sample complexity beyond the $m = 0$ case, however Theorem 2 surprisingly shows that this is not the case: the sample complexity remains constant for all $m \leq d^{1/k}$. Beyond this point, however, a phase transition occurs, and the sample complexity improves *linearly* with the amount of side information, scaling analogously as in the case of Theorem 1.

In the regime $m < d^{1/k}$, the lower bound of Theorem 2 is achieved by error vectors c which are close to being uniform. Remarkably, it turns out that faster rates of convergence are achievable when some of the entries of c are separated from each other. We make this fact precise in Appendix B 3 a, where we show that the error of estimating c improves as a function of the separation between its entries. We highlight here an implication of this result for fidelity estimation. In what follows, we denote by $c_{(1)} \geq \cdots \geq c_{(k)}$ the sorted entries of c, and we interpret $F := c_{(1)}$ as the fidelity.

Proposition 1. *Let conditions* ***(PT)*** *and* ***(S)*** *hold with $m = 0$, and fix $\delta > 0$. Then, there exists an estimator $\widehat{F}$ such that for any $c \in \Delta_k$ with $c_{(1)} > c_{(2)} + \delta$, we have*

$$\mathbb{E}_c \left| \widehat{F} - F \right| \leq C \sqrt{\frac{d^{k-1}}{n^k}},$$

for a constant $C > 0$ depending only on δ, k, γ.

This result is achieved by taking $\widehat{F}$ to be the largest entry of the moment estimator described in Section III B 3, and does not rely on knowledge of δ. This highlights an important property of the moment estimator: it can estimate the fidelity more accurately than the other entries of c. This estimator does so *adaptively*, without requiring assumptions of c or modification of the algorithm itself. To see this, if we heuristically set $\gamma = 0$ and take n to be on the same order as d (i.e. where estimation in regime C is feasible), then, absent any side information, the whole vector c is estimable at the rate $n^{-1/2k}$—which degrades exponentially in k—whereas its largest entry is estimable at the faster rate $n^{-1/2}$, whenever it is δ-separated from the remaining entries. While the required exponential samples with system size currently limits this to a theoretical result, it hints at the possibility of practical estimators with similar properties.

IV. SIMULATION STUDY

To demonstrate the utility of our methods, we analyze synthetic data in two distinct scenarios. On the one hand, we consider Regime A where Π is classically computed, and estimate c with the maximum likelihood estimator (13). On the other hand, we consider Regime B where Π is only available through side-information, in which case we use the variational estimator (16). We additionally report simulations for Regime C in Appendix B.

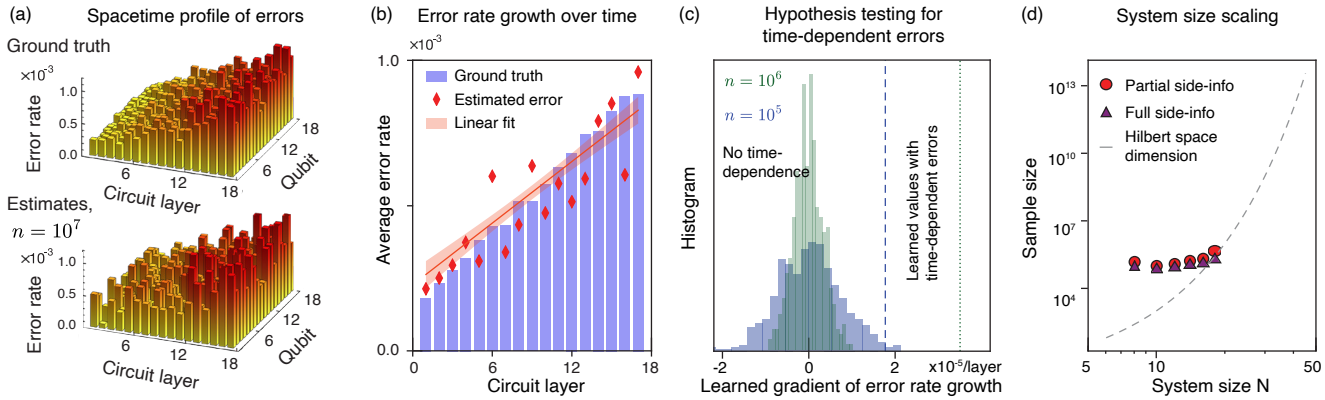

Figure 2. Learning time-dependent error rates. With synthetic data, we demonstrate the use of our protocol to learn about errors that grow over time. (a)(b)(c) We simulate a $N = 18$ one-dimensional brickwork random circuit subject to single-site X, Y, Z Pauli errors, whose error rates (per qubit, per layer) grow from 2.5×10^{-4} in the first layer to 10^{-3} in the last layer. In order to simulate utility-scale circuits of different system sizes, we also set circuit depth equal to the system size while reducing the per-layer error rates such that the total fidelity is fixed at $F \approx 0.5$. (a) Upper panel: the ground truth values of the error rates at each spacetime location. Lower panel: estimated error rates with $n = 10^7$ RCS samples and perfect side information (i.e., $m = \infty$). (b) By averaging over qubits, only 10^6 samples are required to learn the increasing rate with high precision. The blue bars indicate the ground truth; red diamonds mark the estimated error rate, and the shaded red line indicates an extracted rate of error growth (a linear fit to the red diamonds). A non-zero linear fit gradient indicates increasing error rates. (c) Model validation between time-dependent and time-independent error models. To ensure that the learned time-dependence is statistically significant, we compare the extracted gradient (vertical dashed line) against the distribution of gradients learned under the null hypothesis of time-independent error rates, obtained via parametric bootstrap (Appendix H). The histogram of such gradients provides a confidence interval and p-values for time-dependent errors: 10^5 and 10^6 samples (indicated in green and blue respectively) are sufficient to learn the error rate growth with statistical significance. (d) System-size dependence of the sample complexity for model validation. As the system size increases, although the Hilbert space dimension increases exponentially (dashed line), the required sample size for model validation grows only polynomially with system size (orange circles). This sample complexity is defined as the number of RCS samples required to discriminate between a fixed rate of error growth and no error growth with 5σ significance. With increasing system size, classical simulation will not be feasible. In addition, we simulate the case of incomplete side-information (purple triangles) where $m = n$, i.e. the number of side-information samples (per error component) is the same as the number of RCS samples. The sample complexity does not differ significantly between the two cases.

A. Learning time-dependent errors

We first showcase the use of our technique to detect the presence of time-dependent error rates. Such time-dependence can exist in various quantum platforms due to distinct physical reasons, including non-Markovian noise [76], burst errors [77], and atomic heating in an optical tweezer [78]. To this end, we numerically simulate a depth-16 circuit of a one-dimensional $N = 18$ qubit chain. We perform a circuit-level noise simulation with a random quantum circuit. At every layer and every qubit, we inject Pauli X, Y and Z single qubit errors with space- and time-dependent probabilities c, corresponding to single-qubit Pauli channels with varying rates.

In this simulation, we set the average single-qubit error rates to grow by a factor of 4 over the course of the entire circuit (see Appendix H). Our estimators successfully extract the individual time-dependent error rates (Fig. 2a) with 10^7 samples, within the ability of the existing state-of-the-art quantum platforms. Since our aim is to study whether the error rate increases over time, we also perform a statistical test for the null hypothesis that the error rate is constant across layers, which in principle should require fewer samples. This indeed turns out to be the case: the null hypothesis can be rejected with approximately 10^5 samples at level 0.95 (Fig. 2c), and with overwhelming significance when the sample size is of order 10^6.

For this system size, 10^5 samples is comparable to the Hilbert space dimension $2^{18} = 262\,144$. However, we find that the number of samples required for this hypothesis test grows slowly with system size, and we expect it to be far below the Hilbert space dimension for systems of sizes $N > 20$ (Fig. 2d).

B. Learning correlated errors

We also demonstrate the detection of weak correlated errors, such as two-qubit correlated X errors (Fig. 3a) and multi-qubit $XX \cdots X$ errors along a row or column of qubits in a 2D geometry (Fig. 3b). Such errors may occur, for example, due to qubit crosstalk [79, 80] or control-line or readout multiplexing [81] and are fundamentally inaccessible to calibration experiments on isolated subsets of qubits. Despite these errors being weak (respectively 0.2% and 0.1%), 10^7 samples suffice to de-

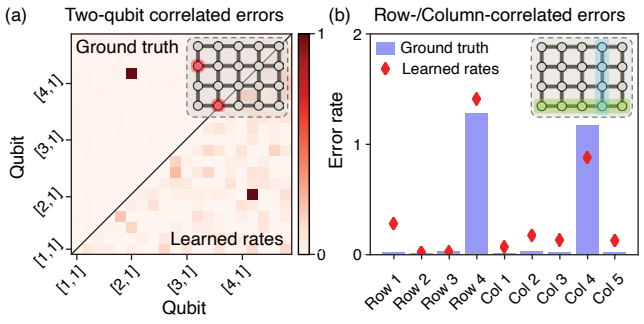

Figure 3. Reconstructing correlated errors from synthetic data simulated for a 4×5, depth-5 circuit. We consider two types of correlated error: (a) two-qubit XX errors, or (b) multi-qubit $XX \cdots X$ errors along one row or one column to simulate errors induced along a shared control line. In both settings, we also include time-independent single-site Pauli errors at every qubit with a rate 2×10^{-3}, chosen such that the total many-body fidelity is $F \approx 0.5$. (a) We learn the rates (averaged over layers) of two-qubit errors $X_u X_v$ for all pairs of qubits simultaneously and represent them on a 2D plot: specifically the correlated error rates $c_{u.v} - c_u c_v$, which subtracts the expected two-qubit error rates from independent single-qubit errors on qubits u and v. We refer to this difference as the correlated error rates. Upper left half: ground truth: two qubits, highlighted in the inset, experience correlated XX errors at a rate of 10^{-3} per layer. Lower right half: extracted error rates from 10^7 samples correctly identify the correlated pair. (b) We also learn the rates of correlated errors on all the qubits in the same row or column. Blue bars: ground truth where one row and one column (inset) experience correlated errors. Red diamonds: extracted error rates. Again, 10^7 samples are sufficient to reliably learn about correlated errors.

tect them. Note also that our method does not depend on geometric locality, and is able to detect correlations between spatially separated qubits.

V. ANALYSIS OF EXPERIMENTAL RCS DATA

Finally, we apply our method to the Google Quantum AI random circuit sampling (RCS) data from Ref. [1]. In this experiment, random quantum circuits with sizes ranging from $N = 12$ to $N = 53$ were executed on a 2D grid. The dataset, which is publicly available, contains ten random circuit realizations per system size N and 500,000 measurement outcomes per circuit.

We perform exact simulations of random circuits up to system size $N = 18$, incorporating a variety of error mechanisms at each spacetime location. We specifically consider several sources of error: state-preparation errors, single-qubit dephasing errors, two-qubit controlled-Z dephasing and flip-flop ($|01\rangle\langle 10| + |10\rangle\langle 01|$) errors which may arise due to dressing by higher transmon levels, and readout errors which may be biased, i.e. have unequal error rates between $1 \to 0$ and $0 \to 1$ processes [1], for a total of $k = 461$ distinct errors. Errors near the beginning and the end of the circuit are correlated not only with other errors, but with the ideal state. Therefore to simplify our first analysis, we only simulate errors beyond the first and last three layers, i.e. the middle 8 circuit layers, see Appendix I for details.

Using our estimator (13) on publicly-available RCS data from Ref. [1], we extract physical error rates associated with each of the above sources across spacetime locations in the circuit. The results are summarized in Fig. 4: the data is highly rich and can be examined along multiple axes, including its behavior over space, time, and its magnitudes resolved by error type and location. Since $1 \to 0$ readout errors are a dominant process and their calibration values were reported in Fig. S24 of Ref. [1], we compare our learned values against reported values. We find similar average values of readout error, but slightly different qubit-to-qubit values. This could be due to the fact that readout error rates differ when the qubits were individually read out as opposed to simultaneously read out [1]. Meanwhile, the data sets we consider lie between both extremes: approximately half the qubits are simultaneously read out.

Physically-realistic error sources introduce systematic deviations from the i.i.d. Porter-Thomas assumption **(PT)** in several ways. As a result, one cannot directly equate the many-body fidelity to the coefficient c_1. We develop a theory for converting the learned rates c into physically meaningful quantities such as the many-body fidelity and the physical per-qubit, per-layer error rate, detailed in Appendix I 2. We summarize the dependence of the error rates over time and over qubits: error rates exhibit inhomogeneity over qubits but remain approximately constant across time, with larger state-preparation and readout errors at the initial and final circuit layers, respectively. Our estimates are consistent across the ten random circuit realizations analyzed, with reproducible trends being observed.

The orders of magnitude of the learned error rates in the bulk of the circuit agree with reported values of two-qubit gate errors, with a physical error rate (combined over 1q dephasing, 2q dephasing and 2q flip-flop) of 0.010(2) per qubit per layer, c.f. the reported mean two-qubit cycle benchmarking error rate of 0.0093(4) (Fig. 2 of Ref. [1]). Note that our '1q' and '2q' error sources refer to errors that can be expressed in terms of single- and two-qubit operators, and are not able to distinguish whether these come from single or two-qubit gates: this may be addressed with more careful positioning of errors in the circuit sequence [Fig. 4(d)].

We also investigate correlated readout errors by learning the rates of readout errors occurring on two (potentially distant) qubits simultaneously. While we see a large number of correlations which may be due to statistical noise, we also observe a certain directional-dependence even of non-local correlated readout errors. These may arise, for example from multiplexed qubit readout lines [1].

Finally, we find that our learning procedure is remarkably robust. Even if our k-component model does not specify all sources of error, the error rates for the com-

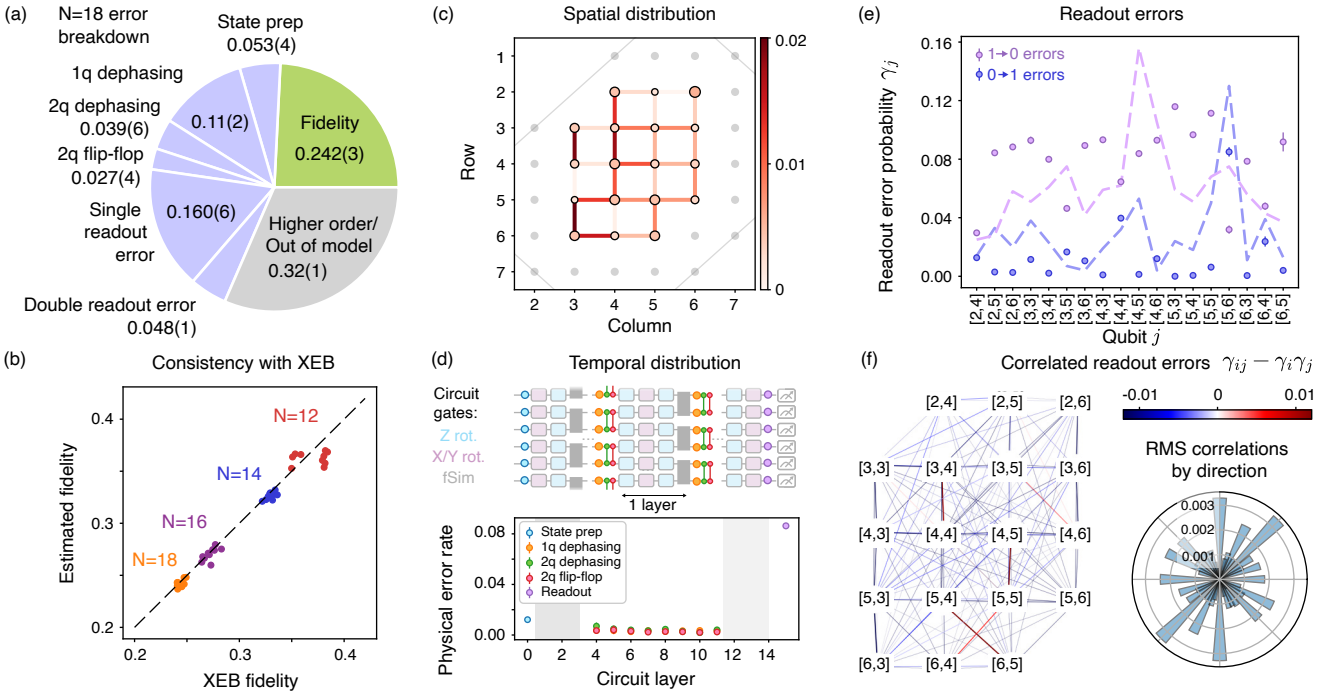

Figure 4. Analysis of experimental RCS data. We apply our methods to study the publicly-available RCS data from Ref. [1], results here shown for $N = 18$. Using the MLE, we resolve different types of errors in many spacetime locations. We simulate state-preparation, single-qubit (1q) dephasing errors, two-qubit (2q) gate dephasing and flip-flop errors, and single and double readout errors (full details in App. I), for a total of $k = 461$ total errors. (a) We summarize the combined contributions of each error type. Quoted values and error bars indicate the sample mean and its standard error over 10 random circuits. Modeled errors account for 68% of the total weight: a remaining 32% weight is fitted to the *white noise* term representing errors outside our model such as multiple errors, consistent with expectations for this fidelity value (App. I). Note that the rate of 1q dephasing errors we learn here are the rates of errors that can be *described* as single-qubit Z_j operators: these errors may also arise from two-qubit *gates*, and hence the proportion of 1q and 2q errors here are comparable, even though we expect them to primarily arise from two-qubit gates. (b) Converting the results of our benchmarking report into a many-body fidelity (App. I 2) yields results in close quantitative agreement with the XEB fidelity. (c) Learned error rates show considerable variation among qubits, in a consistent fashion over random circuit realizations. We plot the total rates of the 2q dephasing and flip-flop errors on nearest neighbors, indicated by the color of the red links. We also plot the single-qubit dephasing error rates, indicated by the size and color of each qubit. Qubits are arranged according to their physical layout on the device (borders and unused qubits for the $N = 18$ dataset in gray). The magnitude of learned error rates is consistent between system size, random circuit agreement, and their sum over error channels is consistent with Ref. [1] (main text). (d) Our procedure also yields time-resolved error rates, revealing approximately time-independent errors in the middle of the circuit. $1 \to 0$ readout errors were found to be the largest type of error. Above, we depict the positioning of our modeled errors in the circuit: in the ideal circuit, a single "layer" consists of four gates applied to each qubit, and we insert errors at layers in the circuit. Errors inserted near the start and end of the circuit have unusual properties, and we omit errors in the first and last three layers (gray regions) to avoid additional complications (see App. I 2). (e) We explicitly compare our estimated rates (points) of readout errors with those reported in Ref. [1] (dashed lines). The average rates of readout errors are quantitatively similar, with deviations on certain qubits: these may arise from the fact that only a subset of qubits are simultaneously measured, which may hence experience error rates different from when all qubits are simultaneously read out (Fig. S24 of Ref. [1]). Error bars indicate standard error over 10 random circuits. (f) Learning correlated readout errors: We estimate the physical error rate $\widehat{\gamma}_{ij}$ of double readout errors on qubits i and j, and compare it to the rates of independent errors $\widehat{\gamma}_i, \widehat{\gamma}_j$: their difference $\widehat{\gamma}_{ij} - \widehat{\gamma}_i\widehat{\gamma}_j$ quantify correlated readout errors. We indicate these correlations with the thickness and colors of lines between all pairs of qubits i and j. These correlations can be as large as a 1% rate, although typical values are closer to 0.2%. We see correlations between many pairs of qubits, with stronger correlations (surprisingly *negative*) between nearest neighbors as well as along the diagonals. We summarize these with a polar plot of the root-mean-squared (RMS) correlations averaged along each direction. Note that this is an *average* over qubit pairs with a given orientation, ignoring their separation, and not simply a sum, which would weight certain directions over others because of the different number of qubit pairs for each orientation.

ponents that are modeled have estimated values that are stable to the presence or absence of other components in the model. This is illustrated in the close agreement between the XEB fidelity and the fidelity estimated by our protocol in Fig. 4(b), which holds true for other quantities as well, such as the fractions of each error component in Fig. 4(a). To a large extent, this is because of the approximate orthogonality of the components π_i,

which implies that estimators such as $\widehat{c}_i^{\mathrm{XEB}}$ are accurate. Such estimators estimate each component without reference to the other components of the model and hence are inherently robust to this type of misspecification. However, in our error model in this section, components are non-orthogonal (App. I 2 a) yet our protocol retains this stability. Indeed, such stability is the operating principle behind the XEB: knowledge of the error processes are not required to estimate the many-body fidelity. This enables refining our error models hierarchically by systematically adding error sources according to expected significance, up to a desired level of precision.

VI. OUTLOOK

Our work demonstrates the utility of novel data processing methods to extract detailed information from random unitary circuits. These have become an industry standard for benchmarking quantum devices, for which many existing datasets are publicly available. Our methods pave the way to a more accurate understanding of the errors that quantum computers experience, which come in many forms, often unexpected. We anticipate further extensions of our methods not only to learn about errors, but also to sense unknown, complicated signals [45]. We also anticipate possible applications of our methods to basic science experiments, to learn about the properties of possibly exotic states prepared on a quantum computer [82].

From a statistical lens, the quantum information setting poses a unique set of new challenges for statisticians: its discrete data in the form of bitstring counts differs from traditional physics experiments involving continuous-variable data, and its high dimensionality means that each individual measurement reveals little about the underlying distribution. In our setting, we also encountered a synergistic dual role of randomness in the quantum circuit: random quantum circuits are hard to classically simulate and serves as a task that separates classical from quantum computers. However, they also have many typical properties which we exploit, and which led us to study a new family of high-dimensional latent variable models.

We anticipate that our methods developed in the setting of partial side information generalize to cases when the reference quantum computer is noisy. In the meantime, however, our methods are still applicable when clean side information samples can be obtained using quantum error detection [3] or error correction methods [6, 9].

Our results point to the broader relevance of high-dimensional statistical methods in quantum computing. The data produced by quantum devices are inherently high-dimensional, and the number of accessible samples is often far smaller than the dimension of the underlying Hilbert space. This imbalance makes quantum data analysis a natural arena for ideas from high-dimensional inference. We anticipate not only the fruitful application of existing tools—such as those introduced in early pioneering work on compressed-sensing-based quantum state tomography [83]—but also the development of new statistical frameworks tailored to the distinct structure and constraints of quantum computing.

ACKNOWLEDGEMENTS

TM would like to thank Florentina Bunea and Marten Wegkamp for discussions related to this work, and for bringing his attention to Ref. [69]. We thank Trond Andersen, Dmitry Abanin, Bryce Kobrin, Elizabeth Bennewitz, Nikita Astrakhantsev, Weijie Wu, Kostyantyn Kechedzhi, Dvir Kafri and Joonhee Choi for insightful discussions. We acknowledge support by the NSF QLCI Award OMA-2016245, the Center for Ultracold Atoms, an NSF Physics Frontiers Center (NSF Grant PHY-1734011), and the NSF CAREER award 2237244. TM gratefully acknowledges the support of a Norbert Wiener fellowship. WG is supported by the Hertz Foundation Fellowship.

[1] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandão, D. A. Buell, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature (London) **574**, 505 (2019).

[2] A. L. Shaw, Z. Chen, J. Choi, D. K. Mark, P. Scholl, R. Finkelstein, A. Elben, S. Choi, and M. Endres, Benchmarking highly entangled states on a 60-atom analog quantum simulator, Nature (London) **628**, 71 (2023).

[3] D. Bluvstein, S. J. Evered, A. A. Geim, S. H. Li, H. Zhou, T. Manovitz, S. Ebadi, M. Cain, M. Kalinowski, D. Hangleiter, *et al.*, Logical quantum processor based on reconfigurable atom arrays, Nature (London) **626**, 58 (2024).

[4] R. Haghshenas, E. Chertkov, M. Mills, W. Kadow, S.-H. Lin, Y.-H. Chen, C. Cade, I. Niesen, T. Begušić, M. S. Rudolph, *et al.*, Digital quantum magnetism at the frontier of classical simulations, arXiv:2503.20870 (2025).

[5] L. Egan, D. M. Debroy, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Newman, M. Li, K. R. Brown, M. Cetina, *et al.*, Fault-tolerant control of an error-corrected qubit, Nature (London) **598**, 281 (2021).

[6] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush, *et al.*, Suppressing quantum errors by scaling a surface code logical qubit, Nature (London) **614**, 676 (2023).

[7] V. V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. Brock, A. Ding, L. Frun-

zio, *et al.*, Real-time quantum error correction beyond break-even, Nature (London) **616**, 50 (2023).
[8] A. Paetznick, M. P. da Silva, C. Ryan-Anderson, J. M. Bello-Rivas, J. P. C. III, A. Chernoguzov, J. M. Dreiling, C. Foltz, F. Frachon, J. P. Gaebler, *et al.*, Demonstration of logical qubits and repeated error correction with better-than-physical error rates, arXiv:2404.02280 (2024).
[9] Quantum error correction below the surface code threshold, Nature (London) **638**, 920 (2025).
[10] D. Bluvstein, A. A. Geim, S. H. Li, S. J. Evered, J. P. B. Ataides, G. Baranes, A. Gu, T. Manovitz, M. Xu, M. Kalinowski, *et al.*, Architectural mechanisms of a universal fault-tolerant quantum computer, arXiv:2506.20661 (2025).
[11] A. L. Shaw, D. K. Mark, J. Choi, R. Finkelstein, P. Scholl, S. Choi, and M. Endres, Experimental signatures of Hilbert-space ergodicity: Universal bitstring distributions and applications in noise learning, Phys. Rev. X **15**, 031001 (2025).
[12] N. Kaufmann, I. Rojkov, and F. Reiter, Characterization of coherent errors in gate layers with robustness to pauli noise, Phys. Rev. Appl. **23**, 034014 (2025).
[13] T. Steckmann, D. Luo, Y.-X. Wang, S. R. Muleady, A. Seif, C. Monroe, M. J. Gullans, A. V. Gorshkov, O. Katz, and A. Schuckert, Error mitigation of shot-to-shot fluctuations in analog quantum simulators, arXiv:2506.16509 (2025).
[14] M. McEwen, L. Faoro, K. Arya, A. Dunsworth, T. Huang, S. Kim, B. Burkett, A. Fowler, F. Arute, J. C. Bardin, *et al.*, Resolving catastrophic error bursts from cosmic rays in large arrays of superconducting qubits, Nature Phys. **18**, 107 (2022).
[15] C. D. Wilen, S. Abdullah, N. Kurinsky, C. Stanford, L. Cardani, G. d'Imperio, C. Tomei, L. Faoro, L. Ioffe, C. Liu, *et al.*, Correlated charge noise and relaxation errors in superconducting qubits, Nature (London) **594**, 369 (2021).
[16] M. McEwen, D. Kafri, Z. Chen, J. Atalaya, K. J. Satzinger, C. Quintana, P. V. Klimov, D. Sank, C. Gidney, A. G. Fowler, *et al.*, Removing leakage-induced correlated errors in superconducting quantum error correction, Nature Commun. **12**, 1761 (2021).
[17] P. Scholl, A. L. Shaw, R. B.-S. Tsai, R. Finkelstein, J. Choi, and M. Endres, Erasure conversion in a high-fidelity Rydberg quantum simulator, Nature (London) **622**, 273 (2023).
[18] S. Ma, G. Liu, P. Peng, B. Zhang, S. Jandura, J. Claes, A. P. Burgers, G. Pupillo, S. Puri, and J. D. Thompson, High-fidelity gates and mid-circuit erasure conversion in an atomic qubit, Nature (London) **622**, 279 (2023).
[19] K. X. Wei, E. Pritchett, D. M. Zajac, D. C. McKay, and S. Merkel, Characterizing non-Markovian off-resonant errors in quantum gates, Phys. Rev. Appl. **21**, 024018 (2024).
[20] Y.-H. Chen and C. H. Baldwin, Randomized benchmarking with leakage errors, arXiv:2502.00154 (2025).
[21] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Characterizing quantum supremacy in near-term devices, Nature Phys. **14**, 595 (2018).
[22] D. K. Mark, J. Choi, A. L. Shaw, M. Endres, and S. Choi, Benchmarking quantum simulators using ergodic quantum dynamics, Phys. Rev. Lett. **131**, 110601 (2023).
[23] T. I. Andersen, N. Astrakhantsev, A. H. Karamlou, J. Berndtsson, J. Motruk, A. Szasz, J. A. Gross, A. Schuckert, T. Westerhout, Y. Zhang, E. Forati, *et al.*, Thermalization and criticality on an analogue–digital quantum simulator, Nature (London) **638**, 79 (2025).
[24] C. H. Baldwin, K. Mayer, N. C. Brown, C. Ryan-Anderson, and D. Hayes, Re-examining the quantum volume test: Ideal distributions, compiler optimizations, confidence intervals, and scalable resource estimations, Quantum **6**, 707 (2022).
[25] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, Validating quantum computers using randomized model circuits, Phys. Rev. A **100**, 032328 (2019).
[26] K. Mayer, A. Hall, T. Gatterman, S. K. Halit, K. Lee, J. Bohnet, D. Gresh, A. Hankin, K. Gilmore, J. Gerber, *et al.*, Theory of mirror benchmarking and demonstration on a quantum computer, arXiv:2108.10431 (2021).
[27] T. Proctor, K. Rudinger, K. Young, E. Nielsen, and R. Blume-Kohout, Measuring the capabilities of quantum computers, Nature Phys. **18**, 75 (2022).
[28] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, *et al.*, Strong quantum computational advantage using a superconducting quantum processor, Phys. Rev. Lett. **127**, 180501 (2021).
[29] Y. Liu, M. Otten, R. Bassirianjahromi, L. Jiang, and B. Fefferman, Benchmarking near-term quantum computers via random circuit sampling, arXiv:2105.05232 (2021).
[30] M. DeCross, R. Haghshenas, M. Liu, E. Rinaldi, J. Gray, Y. Alexeev, C. H. Baldwin, J. P. Bartolotta, M. Bohn, E. Chertkov, *et al.*, The computational power of random quantum circuits in arbitrary geometries, arXiv:2406.02501 (2024).
[31] D. Gao, D. Fan, C. Zha, J. Bei, G. Cai, J. Cai, S. Cao, F. Chen, J. Chen, K. Chen, *et al.*, Establishing a new benchmark in quantum computational advantage with 105-qubit Zuchongzhi 3.0 processor, Phys. Rev. Lett. **134**, 090601 (2025).
[32] Y. Rinott, T. Shoham, and G. Kalai, Statistical aspects of the quantum supremacy demonstration, Statistical Science **37**, 322 (2022).
[33] A. W. Harrow and R. A. Low, Random quantum circuits are approximate 2-designs, Commun. Math. Phys. **291**, 257 (2009).
[34] C. E. Porter and R. G. Thomas, Fluctuations of nuclear reaction widths, Phys. Rev. **104**, 483 (1956).
[35] A. M. Dalzell, N. Hunter-Jones, and F. G. Brandão, Random quantum circuits transform local noise into global white noise, Commun. Math. Phys. **405**, 78 (2024).
[36] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information* (Cambridge University Press, 2010).
[37] A. Nahum, S. Vijay, and J. Haah, Operator spreading in random unitary circuits, Phys. Rev. X **8**, 021014 (2018).
[38] M. P. A. Fisher, V. Khemani, A. Nahum, and S. Vijay, Random quantum circuits, Annu. Rev. Cond. Mat. **14**, 335 (2023).
[39] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic, Prediction-powered inference, Science **382**, 669 (2023).

[40] E. Xia and M. J. Wainwright, Prediction aided by surrogate training, arXiv:2412.09364 (2024).

[41] P. R. Gerber and Y. Polyanskiy, Likelihood-free hypothesis testing, IEEE Trans. Inf. Th. **70**, 7971 (2024).

[42] M. C. Wendl, Collision probability between sets of random variables, Stat. & Prob. Lett. **64**, 249 (2003).

[43] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Vol. 48 (Cambridge University Press, 2019).

[44] T. Proctor, S. Seritan, K. Rudinger, E. Nielsen, R. Blume-Kohout, and K. Young, Scalable randomized benchmarking of quantum computers using mirror circuits, Phys. Rev. Lett. **129**, 150502 (2022).

[45] W. Gong, B. Ye, D. K. Mark, and S. Choi, Robust multiparameter estimation using quantum scrambling, to appear on arXiv.

[46] E. Magesan, J. M. Gambetta, and J. Emerson, Scalable and robust randomized benchmarking of quantum processes, Phys. Rev. Lett. **106**, 180504 (2011).

[47] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, Phys. Rev. A **85**, 042311 (2012).

[48] C. Greganti, T. F. Demarie, M. Ringbauer, J. A. Jones, V. Saggio, I. A. Calafell, L. A. Rozema, A. Erhard, M. Meth, L. Postler, R. Stricker, P. Schindler, R. Blatt, T. Monz, P. Walther, and J. F. Fitzsimons, Cross-verification of independent quantum devices, Phys. Rev. X **11**, 031049 (2021).

[49] D. Zhu, Z. P. Cian, C. Noel, A. Risinger, D. Biswas, L. Egan, Y. Zhu, A. M. Green, C. H. Alderete, N. H. Nguyen, *et al.*, Cross-platform comparison of arbitrary quantum states, Nature Comm. **13**, 6620 (2022).

[50] A. Elben, B. Vermersch, R. van Bijnen, C. Kokail, T. Brydges, C. Maier, M. K. Joshi, R. Blatt, C. F. Roos, and P. Zoller, Cross-platform verification of intermediate scale quantum devices, Phys. Rev. Lett. **124**, 10504 (2020).

[51] Q. Zhu, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, M. Gong, *et al.*, Quantum computational advantage via 60-qubit 24-cycle random circuit sampling, Sci. Bull. (Beijing) **67**, 240 (2022).

[52] S. A. Moses, C. H. Baldwin, M. S. Allman, R. Ancona, L. Ascarrunz, C. Barnes, J. Bartolotta, B. Bjork, P. Blanchard, M. Bohn, *et al.*, A race-track trapped-ion quantum processor, Phys. Rev. X **13**, 041052 (2023).

[53] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, Phys. Rev. A **77**, 012307 (2008).

[54] J. J. Wallman and S. T. Flammia, Randomized benchmarking with confidence, New J. Phys. **16**, 103032 (2014).

[55] R. Harper, I. Hincks, C. Ferrie, S. T. Flammia, and J. J. Wallman, Statistical analysis of randomized benchmarking, Phys. Rev. A **99**, 052350 (2019).

[56] 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, Nature Commun. **10**, 5347 (2019).

[57] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, The randomized measurement toolbox, Nature Rev. Phys. **5**, 9 (2023).

[58] D. Lall, A. Agarwal, W. Zhang, L. Lindoy, T. Lindström, S. Webster, S. Hall, N. Chancellor, P. Wallden, R. Garcia-Patron, E. Kashefi, V. Kendon, J. Pritchard, A. Rossi, A. Datta, T. Kapourniotis, K. Georgopoulos, and I. Rungger, A review and collection of metrics and benchmarks for quantum computers: Definitions, methodologies and software, arXiv:2502.06717 (2025).

[59] A. Ambainis and J. Emerson, Quantum t-designs: t-wise independence in the quantum world, 22nd Annu IEEE Conf. Comp. Compl. **CCC'07**, 129 (2007).

[60] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nature Phys. **16**, 1050 (2020).

[61] T. Zhou and A. Nahum, Emergent statistical mechanics of entanglement in random unitary circuits, Phys. Rev. B **99**, 174205 (2019).

[62] P. Hayden and J. Preskill, Black holes as mirrors: quantum information in random subsystems, J. High Energy Phys. **2007** (09), 120.

[63] B. Ware, A. Deshpande, D. Hangleiter, P. Niroula, B. Fefferman, A. V. Gorshkov, and M. J. Gullans, A sharp phase transition in linear cross-entropy benchmarking, arXiv:2305.04954 (2023).

[64] A. Morvan, B. Villalonga, X. Mi, S. Mandra, A. Bengtsson, P. Klimov, Z. Chen, S. Hong, C. Erickson, I. Drozdov, *et al.*, Phase transitions in random circuit sampling, Nature (London) **634**, 328 (2024).

[65] D. L. Donoho and I. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika **81**, 425 (1994).

[66] D. L. Donoho and I. M. Johnstone, Minimax risk over ℓ_p-balls for ℓ_q-error, Probability Theory and Related Fields **99**, 277 (1994).

[67] G. Raskutti, M. J. Wainwright, and B. Yu, Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls, IEEE Trans. Inf. Theory **57**, 6976 (2011).

[68] Y. Li and G. Raskutti, Minimax optimal convex methods for Poisson inverse problems under ℓ_q-ball sparsity, IEEE Trans. Inf. Theory **64**, 5498 (2018).

[69] X. Bing, F. Bunea, S. Strimas-Mackey, and M. Wegkamp, Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations, The Annals of Statistics **50**, 3307 (2022).

[70] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, J. Machine Learning Res. **3**, 993 (2003).

[71] B. Ghorbani, H. Javadi, and A. Montanari, An instability in variational inference for topic models, Proc. 36th Int. Conf. Machine Learning , 2221 (2019).

[72] M. Celentano, Z. Fan, and S. Mei, Local convexity of the TAP free energy and AMP convergence for Z2-synchronization, The Annals of Statistics **51**, 519 (2023).

[73] M. Celentano, Z. Fan, L. Lin, and S. Mei, Mean-field variational inference with the TAP free energy: Geometric and statistical properties in linear models, arXiv:2311.08442 (2023).

[74] S. Hundrieser, T. Manole, D. Litskevich, and A. Munk, Local Poisson deconvolution for discrete signals, arXiv:2508.00842 (2025).

[75] J. Wright, *How to learn a quantum state*, Ph.D. thesis, Carnegie Mellon University (2016).

[76] D. A. Rower, L. Ateshian, L. H. Li, M. Hays, D. Bluvstein, L. Ding, B. Kannan, A. Almanakly, J. Braumüller, D. K. Kim, *et al.*, Evolution of $1/f$ flux

noise in superconducting qubits with weak magnetic fields, Phys. Rev. Lett. **130**, 220602 (2023).
[77] Y. Hirasaki, S. Daimon, T. Itoko, N. Kanazawa, and E. Saitoh, Detection of temporal fluctuation in superconducting qubits for quantum error mitigation, Appl. Phys. Lett. **123**, 184002 (2023).
[78] S. De Léséleuc, D. Barredo, V. Lienhard, A. Browaeys, and T. Lahaye, Analysis of imperfections in the coherent optical excitation of single atoms to Rydberg states, Phys. Rev. A **97**, 053803 (2018).
[79] D. M. Abrams, N. Didier, S. A. Caldwell, B. R. Johnson, and C. A. Ryan, Methods for measuring magnetic flux crosstalk between tunable transmons, Phys. Rev. Appl. **12**, 064022 (2019).
[80] C. N. Barrett, A. H. Karamlou, S. E. Muschinske, I. T. Rosen, J. Braumüller, R. Das, D. K. Kim, B. M. Niedzielski, M. Schuldt, K. Serniak, *et al.*, Learning-based calibration of flux crosstalk in transmon qubit arrays, Phys. Rev. Appl. **20**, 024070 (2023).
[81] J. Heinsoo, C. K. Andersen, A. Remm, S. Krinner, T. Walter, Y. Salathé, S. Gasparinetti, J.-C. Besse, A. Potočnik, A. Wallraff, and C. Eichler, Rapid high-fidelity multiplexed readout of superconducting qubits, Phys. Rev. Appl. **10**, 034040 (2018).
[82] E. Altman, K. R. Brown, G. Carleo, L. D. Carr, E. Demler, C. Chin, B. DeMarco, S. E. Economou, M. A. Eriksson, K.-M. C. Fu, *et al.*, Quantum simulators: Architectures and opportunities, PRX Quantum **2**, 017003 (2021).
[83] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators, New J. Phys. **14**, 095022 (2012).
[84] C. R. Rao, Maximum likelihood estimation for the multinomial distribution, Sankhyā: The Indian Journal of Statistics (1933-1960) **18**, 139 (1957).
[85] M. Birch, A new proof of the Pearson-Fisher theorem, The Annals of Mathematical Statistics , 817 (1964).
[86] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis Theory and Practice* (Springer, New York, NY, 2007).
[87] T. Manole and A. Khalili, Estimating the number of components in finite mixture models via the group-sort-fuse procedure, The Annals of Statistics **49**, 3043 (2021).
[88] I. M. Johnstone, Gaussian estimation: Sequence and wavelet models (2019), unpublished manuscript.
[89] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, Compressed sensing performance bounds under Poisson noise, IEEE Trans. Sig. Proc. **58**, 3990 (2010).
[90] S. Arlot and A. Celisse, A survey of cross-validation procedures for model selection, Statistics Surveys **4**, 40 (2010).
[91] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective* (Chapman and Hall/CRC, 2006).
[92] P.-L. Loh and M. J. Wainwright, High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity, The Annals of Statistics **40**, 1637 (2012).
[93] F. Jiang and Y. Ma, Poisson regression with error corrupted high dimensional features, Statistica Sinica **32**, 2023 (2022).
[94] F. Jiang, Y. Zhou, J. Liu, and Y. Ma, On high-dimensional Poisson models with measurement error: Hypothesis testing for nonlinear nonconvex optimization, The Annals of Statistics **51**, 233 (2023).
[95] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B **39**, 1 (1977).
[96] R. M. Neal and G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in *Learning in graphical models* (Springer, 1998) pp. 355–368.
[97] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, An introduction to variational methods for graphical models, Machine Learning **37**, 183 (1999).
[98] C. Zhong, S. Mukherjee, and B. Sen, Variational inference for latent variable models in high dimensions, arXiv:2506.01893 (2025).
[99] D. Donoho and V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts?, Adv. Neur. Inf. Proc. Sys. **16**, 1141 (2003).
[100] A. Hyvärinen, J. Hurri, and P. O. Hoyer, Independent component analysis, in *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision* (Springer) pp. 151–175.
[101] L. A. Goodman, Exploratory latent structure analysis using both identifiable and unidentifiable models, Biometrika **61**, 215 (1974).
[102] A. Moitra, *Algorithmic aspects of machine learning* (Cambridge University Press, 2018).
[103] P. Heinrich and J. Kahn, Strong identifiability and optimal minimax rates for finite mixture estimation, The Annals of Statistics **46**, 2844 (2018).
[104] N. Ho and X. Nguyen, Convergence rates of parameter estimation for some weakly identifiable finite mixtures, The Annals of Statistics **44**, 2726 (2016).
[105] N. Ho and X. Nguyen, Singularity structures and impacts on parameter estimation in finite mixtures of distributions, SIAM J. Math. Data Science **1**, 730 (2019).
[106] Y. Wu and P. Yang, Optimal estimation of Gaussian mixtures via denoised method of moments, The Annals of Statistics **48**, 1987 (2020).
[107] Y. Wei, S. Mukherjee, and X. Nguyen, Minimum Φ-distance estimators for finite mixing measures, arXiv:2304.10052 (2023).
[108] I. Ohn and L. Lin, Optimal Bayesian estimation of Gaussian mixtures with growing number of components, Bernoulli **29**, 1195 (2023).
[109] T. Manole and N. Ho, Refined convergence rates for maximum likelihood estimation under finite mixture models, Proc. 39th Int. Conf. Machine Learning **PMLR 162**, 14979 (2022).
[110] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Science **41**, 391 (1990).
[111] T. Hofmann, Probabilistic latent semantic indexing, *Proc. 22nd Annu. Int. ACM SIGIR Conf. Inf. Retrieval*, **SIGIR '99**, 50 (1999).
[112] S. Arora, R. Ge, and A. Moitra, Learning topic models–going beyond SVD, in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science* (IEEE, 2012) pp. 1–10.
[113] A. Anandkumar, D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu, A spectral algorithm for latent Dirichlet allocation, Adv. Neur. Inf. Proc. Sys. **25** (2012).

[114] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, A practical algorithm for topic modeling with provable guarantees, Proc. 30th Intl. Conf. Machine Learning **PMLR 28**, 280 (2013).

[115] S. Arora, R. Ge, R. Kannan, and A. Moitra, Computing a nonnegative matrix factorization–provably, Proc. 44th Annu. ACM Symposium Theory Comp. , 145 (2012).

[116] X. Bing, F. Bunea, and M. Wegkamp, Optimal estimation of sparse topic models, J. Machine Learning Research **21**, 1 (2020).

[117] Z. T. Ke and M. Wang, Using SVD for topic modeling, J. Am. Stat. Assoc. **119**, 434 (2024).

[118] H. Tran, Y. Liu, and C. Donnat, Sparse topic modeling via spectral decomposition and thresholding, arXiv:2310.06730 (2023).

[119] D. Do, S. Chakraborty, J. Terhorst, and X. Nguyen, Moment tensors of Dirichlet distributions and learning latent Dirichlet allocation, arXiv:2509.25441 (2025).

[120] X. Nguyen, Posterior contraction of the population polytope in finite admixture models, Bernoulli **21**, 618 (2015).

[121] Y. Wang, Convergence rates of latent topic models under relaxed identifiability conditions, E. J. Stat. **13**, 37 (2019).

[122] C. L. Canonne *et al.*, Topics and techniques in distribution testing: A biased but representative sample, Foundations and Trends Commun. Inf. Theory **19**, 1032 (2022).

[123] J. Jiao, K. Venkat, Y. Han, and T. Weissman, Minimax estimation of functionals of discrete distributions, IEEE Trans. Inf. Theory **61**, 2835 (2015).

[124] H. Teicher, Identifiability of mixtures of product measures, The Annals of Mathematical Statistics **38**, 1300 (1967).

[125] O. Barndorff-Nielsen, Identifiability of mixtures of exponential families, Journal of Mathematical Analysis and Applications **12**, 115 (1965).

[126] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning* (Cambridge University Press, 2024).

[127] O. Lepski, A. Nemirovski, and V. Spokoiny, On estimation of the L_r norm of a regression function, Probability Theory and Related Fields **113**, 221 (1999).

[128] Y. I. Ingster, On testing a hypothesis which is close to a simple hypothesis, Theory of Probability & Its Applications **45**, 310 (2001).

[129] Y. Han, J. Jiao, and T. Weissman, Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance, Proc. 31st Conf. Learning Theory **PMLR 75**, 3189 (2018).

[130] Y. Wu and P. Yang, Chebyshev polynomials, moment matching, and optimal estimation of the unseen, The Annals of Statistics **47**, 857 (2019).

[131] T. Schramm and A. S. Wein, Computational barriers to estimation from low-degree polynomials, The Annals of Statistics **50**, 1833 (2022).

[132] Y. Han and J. Niles-Weed, Approximate independence of permutation mixtures, arXiv:2408.09341 (2024).

[133] H. P. Rosenthal, On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables, Israel Journal of Mathematics **8**, 273 (1970).

[134] H. Rosenthal, On the span in L^p of sequences of independent random variables (II), Berkeley Symp. on Math. Statist. and Prob. **1972**, 149 (1972).

[135] A. K. Kuchibhotla and A. Chakrabortty, Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression, Information and Inference: A Journal of the IMA **11**, 1389 (2022).

[136] L. Leskelä and I. Välimaa, Sub-poisson distributions: Concentration inequalities, optimal variance proxies, and closure properties, arXiv:2508.12103 (2025).

[137] A. J. Lee, *U-Statistics: Theory and Practice* (CRC Press, 1990).

[138] O. Marchal and J. Arbel, On the sub-Gaussianity of the beta and Dirichlet distributions, Electronic Communications in Probability **22**, 1 (2017).

[139] B. Collins, S. Matsumoto, and J. Novak, The weingarten calculus, Notices of the American Mathematical Society **69**, 1 (2022).

[140] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer Science & Business Media, 2013).

[141] B. Collins and P. Śniady, Integration with respect to the haar measure on unitary, orthogonal and symplectic group, Communications in Mathematical Physics **264**, 773 (2006).

[142] Y. Bao, S. Choi, and E. Altman, Theory of the phase transition in random unitary circuits with measurements, Phys. Rev. B **101**, 104301 (2020).

[143] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, Estimating Rényi entropy of discrete distributions, IEEE Trans. Inf. Theory **63**, 38 (2016).

[144] A. Ostrowski, Recherches sur la méthode de Graeffe et les zéros des polynomes et des séries de Laurent, Acta Mathematica **72**, 99 (1940).

[145] A. M. Ostrowski, A theorem on clusters of roots of polynomial equations, SIAM J. Numer. Analysis **7**, 567 (1970).

[146] G. Szego, *Orthogonal polynomials*, Vol. 23 (American Mathematical Society, 1939).

SUPPLEMENTARY MATERIAL

Appendix A: Notation

Throughout the manuscript, we adopt the following notation.

- Given a vector $x \in \mathbb{R}^d$, $\|x\|_p$ denotes the ℓ_p norm of x, and $\|x\|$ denotes its ℓ_2 norm when no subscript is specified.

- Given $a, b \in \mathbb{R}$, we write $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$, and $(a)_+ = \max\{0, a\}$.

- For any matrix Π, $\|\Pi\|$ denotes its Frobenius norm, and $\Pi_{\cdot j}$ and $\Pi_{i\cdot}$ denote its j-th column and i-th row, respectively.

- For any two vectors $c, c' \in \mathbb{C}^k$, we write

$$W(c, c') = \inf_{\sigma \in \mathcal{S}_k} \left(\sum_{i=1}^{k} |c_{\sigma(i)} - c'_i|^2 \right)^{|frac.12} ,$$

  where the infimum is over the permutation group $\mathcal{S}_k$ on $[k]$.

- Δ_k denotes the $(k-1)$-dimensional simplex, namely the set of vectors $x \in \mathbb{R}^k$ with nonnegative entries such that $\|x\|_1 = 1$. Furthermore, given $1 \le k_0 \le k$, Δ_{k,k_0} denotes the set of all elements $c \in \Delta_k$ which have exactly k_0 distinct entries.

- Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, \dots\}$. For integers $m, r \in \mathbb{N}$, we denote the r-th falling factorial of m by $(m)_r = m(m-1)\dots(m-r+1)$.

- Given a random variable X, its r-th moment and cumulant (when they exist) are denoted by $m_r(X)$ and $\kappa_r(X)$ respectively. By abuse of notation, also abbreviate these quantities by $m_r(f)$ and $\kappa_r(f)$ when f is the distribution or density of X. We abuse notation by writing $m_\alpha(x) = \sum_{i=1}^k x_i^\alpha$ for any $x \in \mathbb{R}^k$.

- We denote by $I(A)$ the indicator function of a set A, and by $\delta_{i,j} = I(i = j)$ the Kronecker delta function.

- We denote by $\mathcal{E}_d$ the Porter-Thomas or exponential distribution with parameter d, and we use the abbreviation $\mathcal{E}_d^k \equiv \mathcal{E}_d^{\otimes k}$ for its k-fold product distribution. That is,

$$\mathrm{d}\mathcal{E}_d^k(\varpi) := d^k \exp\left(d\|\varpi\|_1\right) \mathrm{d}\varpi, \quad \varpi \in \mathbb{R}_+^k.$$

  We also abbreviate by $\mathcal{D}_d$ the flat Dirichlet (i.e. uniform) distribution over Δ_d. We abbreviate by $\mathrm{Mult}(n; p_1, \ldots, p_d)$ the multinomial distribution with n trials, d categories, and success probabilities $(p_1, \ldots, p_d) \in \Delta_d$. Furthermore, $\mathrm{Poi}(\lambda)$ denotes the Poisson distribution with intensity parameter $\lambda > 0$, and $\mathrm{Gamma}(\alpha, \lambda)$ denoted the Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$ (its mean is α/λ).

- For a random variable V and $\alpha \in (0, 2)$, we define the Orlicz norm of V by

$$\|V\|_{\psi_\alpha} = \inf\left\{\eta > 0 : \mathbb{E}\left[e^{(|V|/\eta)^\alpha}\right] \le 2\right\}. \tag{A1}$$

- Given two sequences of nonnegative real numbers $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \le C b_n$ for all $n \ge 1$, and we write $a_n \asymp b_n$ if $a_n \lesssim b_n \lesssim a_n$. Throughout the manuscript, the constant C may always depend on the parameter γ arising in condition **(S)**. In some cases, when it is clear from context, the constant C may also depend on k, or other problem parameters.

- We define the hard- and soft-thresholding functions, with a parameter $\lambda > 0$, for all $x \in \mathbb{R}^d$, by

$$\mathcal{H}_\lambda(x) = \begin{cases} x, & |x| \ge \lambda, \\ 0, & |x| \le \lambda, \end{cases} \qquad \mathcal{S}_\lambda(x) = \mathrm{sign}(x) \cdot \max\{|x| - \lambda, 0\}. \tag{A2}$$

- Given two probability measures P and Q, admitting densities f and g with respect to a σ-finite dominating measure ν, we make use of the standard statistical divergences: the total variation distance $\mathrm{TV}(P, Q) = \frac{1}{2}\int |f - g| d\nu$, the Hellinger distance $H^2(P, Q) = \int (\sqrt{f} - \sqrt{g})^2 d\nu$, the Kullback-Leibler divergence $\mathrm{KL}(P\|Q) = \int \log(f/g) f d\nu$, and the χ^2-divergence $\chi^2(P\|Q) = \int \frac{(f-g)^2}{g} d\nu$. If J is a joint distribution of P and Q, then its mutual information is denoted by $I(X; Y) = \mathrm{KL}(P \otimes Q\|J)$ for $(X, Y) \sim J$. If X is discrete, then $H(X) = -\mathbb{E}[\log f(X)]$ denotes the Shannon entropy of P.

Appendix B: Comparison of Estimators

In this Appendix, we expand on our discussion from Section III B, and discuss various estimators for the parameter vector $c = (c_1, \ldots, c_k)$ in Regimes A–C.

Let us begin by recalling our statistical model. Let $\Pi = (\pi_{ij}) \in \mathbb{R}^{k \times d}$ be a random matrix whose rows belong to the simplex Δ_d. Under condition **(PT)**, the rows of Π are i.i.d., and distributed according to the uniform distribution over Δ_d, also known as the flat Dirichlet law $\mathcal{D}_d$. Although Π is assumed to satisfy condition **(PT)** for our theoretical results, this condition is not required for all estimators below. Under this known marginal distribution of Π, we adopt the sampling model put forth in Section III: Given an unknown error vector $c \in \Delta_k$, one draws conditionally independent observations of the form

$$Z_1, \ldots, Z_n \,\big|\, \Pi \sim \sum_{j=1}^d (\Pi_{\cdot j}^\top c) \delta_{z_j}$$

$$W_{i1}, \ldots, W_{im} \,\big|\, \Pi \sim \sum_{j=1}^{d} \pi_{ij} \delta_{z_j}, \quad i = 1, \ldots, k,$$

where $z_j \in \{0,1\}^N$ denotes the binary enumeration of the integer j, and $d = 2^N$. Furthermore, let $Y_j := \sum_{\ell=1}^{n} I(Z_\ell = z_j)$ and $V_{ij} := \sum_{\ell=1}^{m} I(W_{i\ell} = z_j)$ denote the induced histograms, for all $i = 1, \ldots, k$ and $j = 1 \ldots, d$. Recall that we consider three regimes:

1. **Regime A** ($m = \infty$): The user observes $Z_1, \ldots, Z_n$ and the matrix Π.
2. **Regime B** ($0 < m < \infty$): The user observes $Z_1, \ldots, Z_n$ and $W_{11}, \ldots, W_{km}$.
3. **Regime C** ($m = 0$): The user merely observes $Z_1, \ldots, Z_n$.

In the following subsections, we discuss several practical estimators for c in these three settings, and connect them to existing statistical literature. We also report simulation studies comparing the numerical performance of these estimators. Whenever possible, we state upper bounds on their sample complexity. We begin with the simplest case of Regime A.

1. Regime A

In the simplest case $m = \infty$ where the matrix Π is known to the practitioner, our model reduces to a multinomial generalized linear model with identity link function, in the sense that the histogram Y satisfies the relation

$$Y \sim \mathrm{Mult}(n; \Pi^\top c). \tag{B1}$$

Since the rows of $\Pi \in \mathbb{R}^{k \times d}$ define probability mass functions with support size d, model (B1) can also be interpreted as a mixture of known multinomial distributions with unknown mixing weights. This is a well-studied model [69, 84–87] for which the most natural estimator is, perhaps, the maximum likelihood estimator, which we describe below. Nevertheless, we begin by detailing how the simple XEB estimator (11) arises in this model.

a. *The XEB Estimator*

Under a Poisson approximation of the multinomial distribution—which we will justify in the next section—one can think of the histogram entries Y_j as being approximately independent, and distributed as

$$Y_j \sim \mathrm{Poi}(n \Pi_{\cdot j}^\top c), \quad j = 1, \ldots, d. \tag{B2}$$

Under this approximation, one has

$$Y = n \Pi^\top c + \epsilon, \tag{B3}$$

where, conditionally on Π, ϵ is a vector with independent and mean-zero entries satisfying $\mathrm{Cov}[\epsilon | \Pi] = \mathrm{diag}(n \Pi^\top c)$. This defines a linear regression model with heteroscedastic errors. Under condition **(PT)**, the marginal covariance of ϵ is simply $\mathrm{Cov}[\epsilon] = n I_d / d$, which is homoscedastic, and is of the same order of magnitude as the entries of Π. Therefore, if one's goal is to obtain an estimator which performs well *unconditionally*—on average over Π—one possible approach to estimating c is to fit the ordinary least-squares estimator of c, which ignores the heteroscedastic nature of the problem,

$$\widehat{c}^{\,\mathrm{OLS}} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \|Y - n \Pi^\top x\|_2^2 = (\Pi \Pi^\top)^{-1} \Pi Y / n. \tag{B4}$$

One can also restrict the minimization to the natural parameter space Δ_k of c, and such a restriction has regularization advantages which we describe below, but no longer admits a closed form. Under condition **(PT)** and for large values of d, the matrix $\Pi\Pi^\top$ concentrates rapidly around its mean value, which is well-approximated by the matrix

$$A = (I_k + \mathbb{1}_k \mathbb{1}_k^\top)/d.$$

Thus, an even simpler estimator of c is given by

$$A^{-1}\Pi Y/n = \frac{d}{n}\left(\Pi Y - \frac{\mathbb{1}_k^\top \Pi Y}{k+1}\mathbb{1}_k\right). \tag{B5}$$

Up to centering, this approximation of the least-squares estimator is similar to that arising in linear regression with orthogonal design matrices [88]. The XEB estimator (11),

$$\widehat{c}^{\,\mathrm{XEB}} = (d/n)\Pi Y - \mathbb{1}_k \tag{B6}$$

can now be understood as a variant of equation (B5) in which the second term is simply approximated by $\mathbb{1}_k$—an approximation which can be justified when d is large using the fact that the mean of Y is $n\Pi^\top c$ with $\mathbb{1}_k^\top c = 1$.

As we have already indicated, the XEB estimator can be improved when k is large via its hard- or soft-thresholded counterparts:

$$\widehat{c}^{\,\mathrm{XHT}} = \mathcal{H}_\lambda(\widehat{c}^{\,\mathrm{XEB}}), \quad \widehat{c}^{\,\mathrm{XST}} = \mathcal{S}_\lambda(\widehat{c}^{\,\mathrm{XEB}}). \tag{B7}$$

Here, $\lambda > 0$ is a tuning parameter, and $\mathcal{H}_\lambda, \mathcal{S}_\lambda$ denote the thresholding functions defined in equation (A2). Once again, these estimators are similar in spirit to hard- and soft-thresholding estimators in linear regression models with orthogonal design matrix [65, 88]. In our context, they satisfy the following upper bound.

Proposition 2. *Under condition **(PT)**, there exists a universal constant $C > 0$ such that for all $1 \le k, n \le d$, there exists $\lambda \ge 0$ such that*

$$\sup_{c \in \Delta_k} \mathbb{E}_c \|\widehat{c}^{\,\mathrm{XHT}} - c\|_2 \le C \cdot \min\left\{(k/n)^{1/2}, (\log k/n)^{1/4}\right\}.$$

Proposition 2 is a special case of upper bounds for Regime B which we will develop in the next section, thus we leave it without explicit proof. It is worth emphasizing that Proposition 2 exhibits the same convergence rate as typically seen in ℓ_1-sparse linear regression problems [66, 67]. This is perhaps surprising, since the heteroscedastic nature of multinomial regression problems can alter the minimax estimation rate, as we shall see in our discussion of likelihood estimators below. However, one of the implicit observations in our upper and lower bounds is the fact that, under condition **(PT)**, the heterosedasticity of our model is sufficiently mild for it to behave like a homoscedastic model. This observation is what allows us to establish the minimax optimality of simple estimators, like the (regularized) XEB estimator, which have the advantage of being computable in closed-form—an important benefit for large-scale quantum computing problems where d grows exponentially with system size. Let us emphasize that the work of [68] also found regularized least squares estimators to be minimax-optimal in ℓ_1-constrained Poisson regression problems.

b. The Maximum Likelihood Estimator

Although the XEB estimator is minimax optimal and simple to compute, it relies heavily on the Porter-Thomas assumption **(PT)**. We have observed some deviations from this assumption in our real data analysis. An alternative estimator for Regime A which does not

rely on this assumption is the maximum likelihood estimator (MLE), defined by

$$\widehat{c}^{\mathrm{MLE}} = \operatorname*{argmax}_{x \in \Delta_k} \sum_{j=1}^{d} Y_j \log\bigl(\Pi_{\cdot j}^{\top} x\bigr). \tag{B8}$$

Unlike the XEB estimator, this optimization problem does not enjoy a closed form, but is nevertheless concave and can be computed using standard solvers. It also has the practical advantage of being free of tuning parameters, unlike our thresholded XEB estimators.

Several theoretical properties of the MLE have been investigated by Bing *et al.* [69] in the context of topic modeling (a framework which we discuss further in Appendix B 3 a). One of their remarkable findings is the fact that the MLE can identify the *sparsity pattern* of the error vector c without any explicit regularization. Concretely, under some conditions, they show that with probability tending to one,

$$\operatorname{supp}(\widehat{c}^{\mathrm{MLE}}) \subseteq \operatorname{supp}(c),$$

cf. [69, Theorem 5]. In our context, this property implies that if one specifies a conservative number of candidate errors k, many of which may not be present in a quantum device at hand, then the MLE is unlikely to assign a positive error rate to any of these non-existent errors. Although this result relies on conditions which are only met in our setting in the unrealistic case $n \gg d$, it nevertheless hints at an important practical property of the MLE.

Bing *et al.* [69] also derive ℓ_1 sample complexity upper bounds for the MLE. Staying again with the condition $n \gg d$, a special case of their results can be informally stated as

$$\mathbb{E}\|\widehat{c}^{\mathrm{MLE}} - c\|_1 \lesssim \min\left\{\kappa^{-2}\sqrt{\frac{\rho \log k}{n}}, \kappa^{-1}\sqrt{\frac{k}{n}}\right\}, \tag{B9}$$

where[2]

$$\kappa = \min_{v \in \mathbb{R}^k} \frac{\|\Pi^{\top} v\|_1}{\|v\|_1}, \quad \rho = \max_{1 \le j \le d} \frac{\|\Pi_{j\cdot}\|_\infty}{\Pi_{\cdot j}^{\top} c}. \tag{B10}$$

The quantity κ is an ℓ_1 analogue of the minimal singular value of the matrix Π. Under condition **(PT)** and $k = o(d)$, a simple derivation shows that κ scales as $k^{-1/2}$ up to logarithmic factors, with high probability. Furthermore, ρ is typically of constant order in our setting. Therefore, equation (B9) shows that the MLE achieves the ℓ_1 convergence rate $k/\sqrt{n}$. This rate is consistent with that of Proposition 2 when translated from the ℓ_2 to ℓ_1 norm. We expect that the MLE can also be shown to achieve the optimal convergence rate in the realistic regime where d is arbitrarily large—and related results can already be deduced for instance from the work of [89]—but we leave a careful analysis of this problem to future work.

c. *Numerical Comparison in Regime A*

We provide a brief numerical comparison of the five estimators discussed in the preceding subsections: the maximum likelihood estimator (B8), the least squares estimator (B4), the simplex-constrained least squares estimator, the XEB estimator (B6), and the hard-thresholded XEB estimator (B7). For the latter estimator, we choose the tuning parameter λ via two-fold cross-validation [90]. We apply these estimators to two different models:

---

[2] We state the results of [69] in the special case where, in their notation, $\underline{J} = [d]$ and $\overline{J} = [d]$. These assumptions hold in our setting with high probability when $n \gg d$.

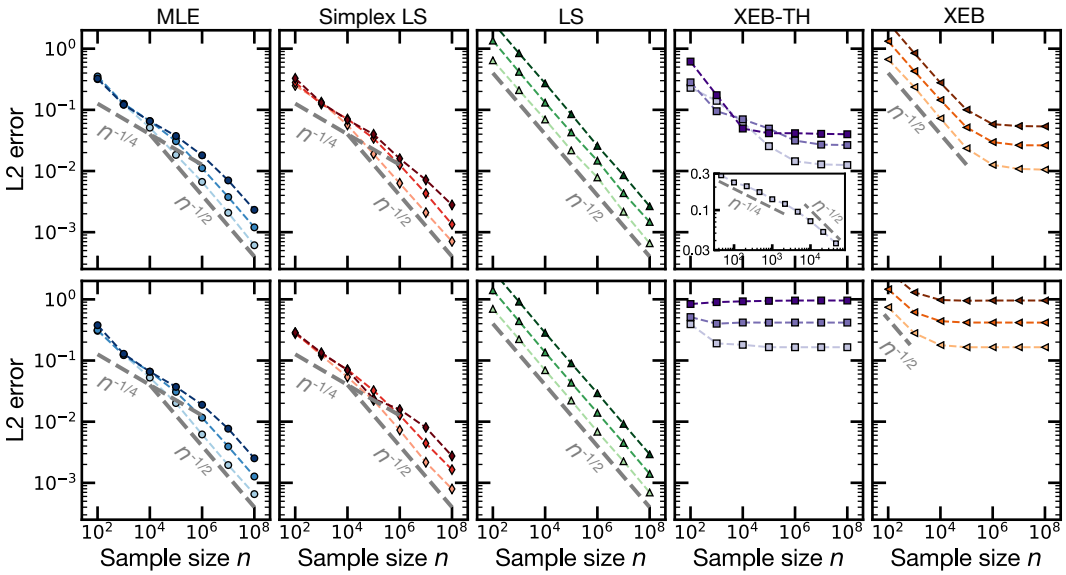

Fig. S1. Sample complexity of various estimators in regime A. For each panel, from the lightest to the darkest, k is chosen from $\{46, 181, 721\}$. For the top (bottom) row, the matrix Π is taken from the Dirichlet distribution (simulating random unitary circuits). Inset for XEB-TH: zoom-in of the regime with small sample sizes to highlight the $n^{-1/4}$ scaling. Each data point is obtained by averaging over at least 10 repetitions of simulation.

1. The rows of Π are drawn independently and uniformly from Δ_d (i.e. assumption **(PT)** holds).

2. The rows of Π are obtained from numerical simulation of a one-dimensional brick-layer quantum random circuit. In particular, the first row of Π corresponds to perfect simulation of one specific instance of the random circuit. Each of the other rows corresponds to one single Pauli error occurring in the circuit. Different rows correspond to different Pauli errors (either occurring on different qubits or at different layers of the circuit).

For both models, we choose the Hilbert space dimension as $d = 2^{16} = 65,536$, and the number of errors k from $\{46, 181, 721\}$. Recall that the first entry of c should be thought of as the fidelity of the device—representing the probability of noiseless execution. We always set this entry to 0.5; the remaining $k-1$ entries correspond to the probabilities of different errors occurring in the circuit, and we sample them from a uniform distribution on Δ_{k-1}.

The ℓ_2 errors of these estimators are summarized in Fig. S1 as a function of the sample size. A few important remarks are in order. First, the MLE, regularized XEB, and simplex-constrained least squares estimators all exhibit two regimes of estimation error. When the sample size n is relatively small, the ℓ_2 error appears to be almost independent of k, and improves roughly as $n^{-1/4}$, whereas for large n, the ℓ_2 error becomes linear in k and scales as $n^{-1/2}$; these scalings are consistent with Proposition 3. Second, the performance of the XEB estimators plateaus in the unrealistic regime $n > d$, which is to be expected from its derivation. This effect is more severe in the brick-layer model, which suffers from slight dependence among the row vectors in the Π matrix. Third, for the unregularized estimators (XEB and least squares), the ℓ_2 error always increases linearly in k. Finally, we find that even in realistic circuits, the MLE does not outperform the other estimators by an appreciable margin, even though it accounts for the mild heteroscedasticity of the model.

2. Regime B

When the amount of side information m is finite but nonzero, our problem can be interpreted as an *errors-in-variables* [91] multinomial regression problem, in which the matrix Π is unknown, but noisy samples W from its rows are given. We consider three estimators for this problem, which we discuss in turn.

a. The Collision Estimator

One of the benefits of the XEB estimator from the previous section is the fact that it is a *linear* functional of Π. Therefore, this estimator can be adapted to Regime B by replacing Π with its empirical counterpart, without incurring any bias. Concretely, by defining

$$\widehat{\Pi} = V/m, \quad \text{i.e. } \widehat{\Pi}_{ij} = \frac{1}{m}\sum_{r=1}^{m} I(W_{ir} = j), \quad i = 1, \dots, k, j = 1, \dots, d,$$

we arrive at the following counterpart of the XEB estimator, which was already presented in equation (14):

$$\widehat{c}^{\,\mathrm{coll}} = (d/n)\widehat{\Pi}Y - \mathbb{1}_k, \quad \text{i.e. } \widehat{c}_i^{\,\mathrm{coll}} = \frac{d}{nm}\sum_{\ell=1}^{n}\sum_{r=1}^{m} I(Z_\ell = W_{ir}) - 1, \quad i = 1, \dots, k. \tag{B11}$$

As before, we refer to this estimator as the *collision* estimator. We can also form regularized variants of this estimator via thresholding:

$$\widehat{c}^{\,\mathrm{HT}} := \mathcal{H}_\lambda(\widehat{c}^{\,\mathrm{coll}}), \quad \widehat{c}^{\,\mathrm{ST}} = \mathcal{S}_\lambda(\widehat{c}^{\,\mathrm{coll}}), \tag{B12}$$

for some $\lambda > 0$. The following result shows that the hard-thresholding collision estimator achieves the optimal sample complexity stated in Theorem 1.

Proposition 3 (Informal)**.** *There exists a universal constant $C > 0$ such that for all $1 \le n, k \le d$, $m \ge 1$, there exists $\lambda \ge 0$ such that*

$$\sup_{c \in \Delta_k} \mathbb{E}\|\widehat{c}^{\,\mathrm{HT}} - c\|_2 \le C \cdot \min\left\{ \left(\frac{dk}{nm_d}\right)^{\frac{1}{2}}, \left(\frac{d\log k}{nm_d}\right)^{\frac{1}{4}} \right\},$$

with $m_d = \min\{m, d\}$.

A rigorous statement and proof of this result appears in Appendix E. As before, the collision estimator has the advantage of being computationally efficient, and achieves the minimax optimal rate of convergence, but has the downside of relying strongly on condition **(PT)**. We next develop an estimator which somewhat relaxes this assumption.

b. The Errors-in-Variables Estimator

Recall from the previous section that the XEB estimator can be understood as an approximation of the ordinary least squares estimator with an orthonormal design matrix Π (up to centering). One way of generalizing the least squares estimator to Regime B is given by the following estimator:

$$\operatorname*{argmin}_{x \in \mathbb{R}^k} \left\{ x^\top A_V x - 2\frac{Y^\top V^\top x}{m} \right\} = A_V^{-1}\frac{VY}{nm}, \tag{B13}$$

where A_V is an unbiased estimator of $\Pi\Pi^\top$ based on the side information V. Again, one can also regularize this estimator by restricting its feasible set to the simplex, at the expense of losing the closed-form solution:

$$\widehat{c}^{\mathrm{EiV}} = \operatorname*{argmin}_{x\in\Delta_k} \Big\{ x^\top A_V x - 2\frac{Y^\top V^\top x}{m} \Big\}. \tag{B14}$$

Variants of these estimators have previously appeared in the work of [92], where they were motivated by the fact that their objective function is an unbiased estimator of the usual least squares objective $\|Y - n\Pi^\top\gamma\|_2^2$, up to addition of a constant that does not depend on γ. Analogues of such estimators for Poissonian models have appeared in the works of [93, 94], though are based on a log-link parametrization and do not easily extend to our setting.

A variety of matrices A_V can be used in the definition of $\widehat{c}^{\mathrm{EiV}}$. When condition **(PT)** happens to hold, one such estimator is given by

$$A_V := \mathbb{E}[\Pi\Pi^\top \,|\, V], \quad \text{i.e. } (A_V)_{i\ell} = \mu_i^\top \mu_\ell + \frac{(d+m)\|V_{i\cdot}+1\|_1 - \|V_{i\cdot}+1\|_2^2}{(d+m)^2(d+m+1)} I(i=\ell), \tag{B15}$$

where $\mu_i = (V_{i\cdot}+1)/(d+m)$. Below, we will see that with this choice of A_V, the estimator (B14) achieves reasonable performance even on simulated random circuit data where mild deviations from the Porter-Thomas assumption can occur.

c. *The Variational Maximum Likelihood Estimator*

Our third estimator in Regime B is the maximum likelihood estimator (MLE), which was defined in equation (15). As discussed therein, the MLE is nonconvex in this problem, and we proposed to approximate it by the fixed-point equation (16). The goal of this subsection is to derive that approximation. In statistical terms, this fixed-point equation arises as the limit of a mean-field variational expectation-maximization (EM) algorithm [95–97], and is inspired by [70]. We develop it from first principles for completeness.

It will be convenient to rewrite the data generating distribution according to the following hierarchy: Given a matrix Π satisfying the Porter-Thomas assumption **(PT)**, for $i = 1,\dots,k$, $\ell = 1,\dots,n$ and $r = 1,\dots,m$, we draw,

$$U_\ell^c \sim \textstyle\sum_{i=1}^k c_i\delta_i, \quad Z_\ell \,|\, (U_\ell^c, \Pi) \sim \sum_{j=1}^d (\pi_{U_\ell^c j})\delta_{z_j}, \quad W_{ir} \,|\, \Pi \sim \sum_{j=1}^d \pi_{ij}\delta_{z_j}.$$

That is, $U_\ell^c \in \{1,\dots,k\}$ denotes a latent variable which indicates the category from which bitstring Z_ℓ was drawn. In this notation, the likelihood function of c under Regime B can be expressed as:

$$\mathcal{L}(x) = \mathbb{E}_{U^x,\Pi}\left[\prod_{\ell=1}^n \pi_{U_\ell^x Z_\ell} \cdot \prod_{r=1}^m \prod_{i=1}^k \pi_{iW_{ir}}\right], \quad x\in\Delta_k,$$

which depends on x only through the distribution of the vector $U^x = (U_\ell^x)_{\ell=1}^n$. The above expression can be interpreted as the partition function of a classical physical system with states Π, U^x, and Hamiltonian

$$\mathcal{H}(U^x,\Pi|Z,W) = \sum_{\ell=1}^n \log\big(\pi_{U_\ell^x Z_\ell}\big) + \sum_{r=1}^m\sum_{i=1}^k \log(\pi_{iW_{ir}}).$$

By the Gibbs variational principle, one can express the partition function via

$$\log\mathcal{L}(x) = \sup_J \Big\{ \mathbb{E}_{(\overline{U},\overline{\Pi})\sim J}\big[\mathcal{H}(\overline{U},\overline{\Pi}|Z,W)\big] - \mathrm{KL}\big(J \,\|\, P_{U^x,\Pi|Z,W}\big)\Big\}, \tag{B16}$$

where $P_{U^x,\Pi|Z,W}$ denotes the joint probability distribution of the latent variables U^x and Π given the observables Z and W, and the supremum is taken over all probability distributions J on $[k]^n \times \Delta_d^k$. This representation suggests approximating the intractable distribution $P_{U^x,\Pi|Z,W}$ by a tractable family of joint distributions. We will adopt a *mean-field* approximation, taking this family to consist of all *independent* joint distributions J_ϕ on $[k]^n \times \Delta_d^k$, whose first marginal is any discrete distribution $\phi \in \Delta_k^n$, and whose second marginal is given by the posterior law $\mathcal{D}_W$ of Π given W,

$$\Pi \,|\, W \sim \mathcal{D}_W := \bigotimes_{i=1}^{k} \text{Dirichlet}\big(1 + V_{i1}, \ldots, 1 + V_{id}\big), \quad \text{with } V_{ij} = \sum_{r=1}^{m} I(W_{ir} = j).$$

That is, we will restrict the supremum in equation (B16) to the set of joint distributions of the form

$$(\overline{U}, \overline{\Pi}) \sim J_\phi = \left(\bigotimes_{\ell=1}^{n} \sum_{i=1}^{k} \phi_{i\ell} \delta_i \right) \otimes \mathcal{D}_W, \quad \phi \in \Delta_k^n.$$

This leads to the following lower bound on the partition function:

$$\log \mathcal{L}(x) \geq \sup_{\phi \in \Delta_k^n} \mathcal{F}(x, \phi), \quad x \in \Delta_k,$$

where $\mathcal{F}(x, \phi)$ denotes the mean-field free entropy,

$$\mathcal{F}(x, \phi) = \mathbb{E}_{(\overline{U}, \overline{\Pi}) \sim J_\phi} \big[\mathcal{H}(\overline{U}, \overline{\Pi} | Z, W) \big] - \mathrm{KL}\big(J_\phi \,\|\, P_{U^x,\Pi|Z,W} \big),$$

The variational EM algorithm consists of performing coordinate ascent on the mean-field free entropy: Given initial values $x^{(0)} \in \Delta_k$ and $\phi^{(0)} \in \Delta_k^n$, we perform the following iterations for all $t = 0, 1, \ldots$

(a) $\phi^{(t+1)} = \text{argmax}_{\phi \in \Delta_k^n} \mathcal{F}(x^{(t)}, \phi)$.

(b) $x^{(t+1)} = \text{argmax}_{x \in \Delta_k} \mathcal{F}(x, \phi^{(t+1)})$.

In order to derive the maxima in the above steps, notice first that the free entropy can be rewritten as

$$\mathcal{F}(x, \phi) = \mathbb{E}_{(\overline{U}, \overline{\Pi}) \sim J_\phi} \big[\log p_{U^x,\Pi,Z,W}(\overline{U}, \overline{\Pi}, Z, W) \big] + H(J_\phi) - \log p_{Z,W}(Z, W),$$

where H denotes the differential entropy of J_ϕ, and the joint law of the random variables is given, over their support, by

$$p_{U^x,\Pi,Z,W}(\overline{U}, \overline{\Pi}, Z, W) = \prod_{\ell=1}^{n} x_{\overline{U}_\ell} \overline{\pi}_{\overline{U}_\ell Z_\ell} \cdot \prod_{i=1}^{k} \prod_{r=1}^{m} \overline{\pi}_{i W_{ir}}.$$

Letting "$\propto$" denote equality up to additive constants not depending on x, ϕ, we deduce

$$\begin{aligned}
\mathcal{F}(x, \phi) &\propto \sum_{i=1}^{k} \sum_{\ell=1}^{n} \phi_{i\ell} \mathbb{E}_{\mathcal{D}_V} [\log(x_i \overline{\pi}_{i Z_\ell})] - \sum_{i=1}^{k} \sum_{\ell=1}^{n} \phi_{i\ell} \log \phi_{i\ell} \\
&= \sum_{i=1}^{k} \sum_{\ell=1}^{n} \phi_{i\ell} \log \frac{x_i}{\phi_{i\ell}} + \sum_{i=1}^{k} \sum_{\ell=1}^{n} \phi_{i\ell} \big(\psi(1 + W_{iZ_\ell}) - \psi(d + m) \big) \\
&\propto \sum_{i=1}^{k} \sum_{\ell=1}^{n} \phi_{i\ell} \log \left(\frac{x_i \exp\{\psi(1 + W_{iZ_\ell})\}}{\phi_{i\ell}} \right),
\end{aligned}$$

where ψ denotes the di-gamma function, and the second line is obtained in closed form using the fact that the marginal distribution of $\bar{\pi}_{ij}$ is $\mathrm{Beta}(1+W_{iZ_\ell}, d-1+\sum_{s\neq Z_\ell} W_{is})$. By maximizing the above display with respect to each variable x and ϕ, we deduce that the iterations (a) and (b) reduce to:

$$\text{(a)}\ \ \phi_{i\ell}^{(t+1)} = x_i^{(t)} \exp\{\psi(1+W_{iZ_\ell})\} / \textstyle\sum_{r=1}^k x_r^{(t)} \exp\{\psi(1+W_{rZ_\ell})\},$$

$$\text{(b)}\ \ x_i^{(t+1)} = \tfrac{1}{n} \textstyle\sum_{\ell=1}^n \phi_{i\ell}^{(t+1)},$$

for $t = 0, 1, \ldots$ These iterations simplify to

$$x_i^{(t+1)} = \frac{x_i^{(t)}}{n} \sum_{\ell=1}^n \frac{\exp\{\psi(1+W_{iZ_\ell})\}}{\sum_{r=1}^k x_r^{(t)} \exp\{\psi(1+W_{rZ_\ell})\}} = \frac{x_i^{(t)}}{n} \sum_{j=1}^d \frac{Y_j S_{ij}}{\sum_{r=1}^k x_r^{(t)} S_{ij}}, \quad i = 1, \ldots, k, \tag{B17}$$

where $S_{ij} = \exp\{\psi(1+V_{ij})\}$. We refer to these iterates as the *variational EM estimator*. These iterates converge precisely to a solution of the mean-field fixed-point equation

$$n = \sum_{j=1}^d \frac{Y_j S_{ij}}{\sum_{r=1}^k S_{rj} x_r}, \quad i = 1, \ldots, k,$$

which completes our derivation of equation (16).

Algorithm (B17) has the well-known property of increasing the likelihood at each iteration, in the sense that $\mathcal{L}(x^{(t+1)}) \geq \mathcal{L}(x^{(t)})$ for all $t \geq 0$. Nevertheless, this algorithm is not guaranteed to converge to the maximum likelihood estimator, i.e. the global optimum of $\mathcal{L}$. In a closely-related statistical model known as latent Dirichlet allocation, which we will discuss further below, it has recently been shown that the mean-field approximation is negligible for computing Bayesian estimators when $n = o(k)$ [98], however this analysis corresponds most closely to our model with $m = \infty$. In contrast, we believe that the mean-field approximation becomes very poor when m is of lower order than n and d, as will become clear in our numerical comparisons below. As we have already indicated, one avenue for possible improvement of this algorithm is to optimize the Thouless-Anderson-Palmer free entropy [71–73] instead of the mean-field free entropy. We intend to pursue this avenue in future work.

d. Numerical Comparison in Regime B

We again analyze five different estimators in Regime B, which are counterparts to those in Appendix B 1 c above: The variational EM estimator (B17), the error-in-variable least-square estimators with and without the simplex constraint (B14), and the collision-based estimators with and without hard-thresholding (B11)–(B12). We apply these estimators to the same data as in Appendix B 1 c. This time, we fix the number of errors $k = 46$, and vary the side information sample size m.

The ℓ_2 errors of these estimators as a function of the sample size are summarized in Fig. S3. Let us again make a few remarks. When m is sufficiently large (in particular, larger than d), all estimators exhibit qualitatively similar performance as their counterparts in Regime A. On the other hand, smaller values of m have two main impacts on the the ℓ_2 error: 1) They set a lower bound on the attainable ℓ_2 error, 2) although the ℓ_2 error still decreases as $n^{-1/2}$, the prefactor becomes larger (with a factor of d/m). These various observations are consistent with our theoretical predictions in Proposition 3 and Theorem 1. Furthermore, we observe several phenomena which parallel those of Regime A: 1) The various estimators have comparable risks in most regimes, highlighting the fact that simple least squares-based

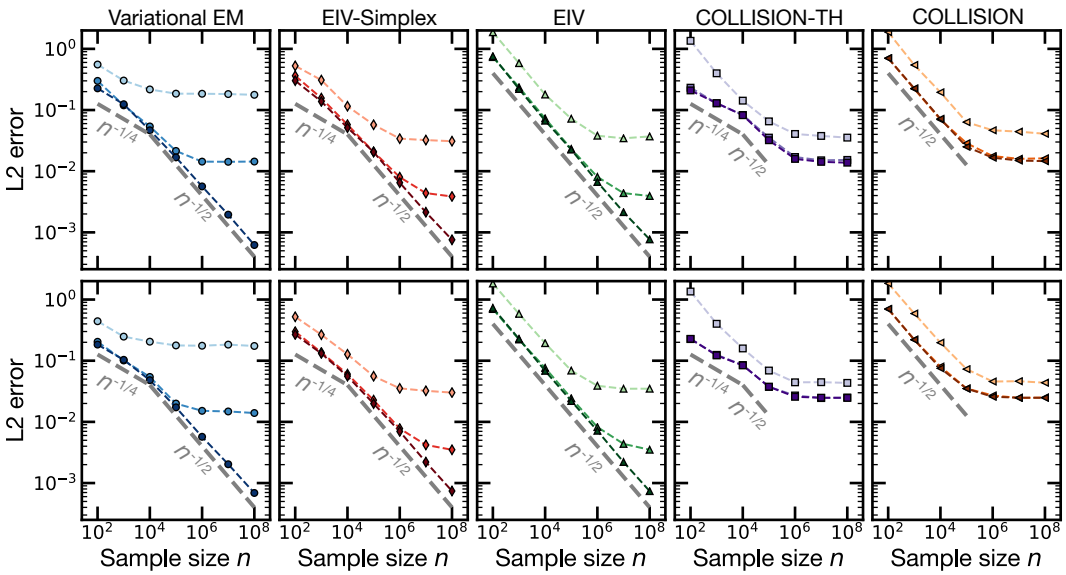

Fig. S2. Sample complexity of various estimators in Regime B. For each panel, from the lightest to the darkest, the side information sample size m is chosen from $\{10^4, 10^6, 10^8\}$. For the top (bottom) row, the matrix Π is taken from the Dirichlet distribution (simulating random unitary circuits). Each data point is obtained by averaging over at least 10 repetitions of simulation.

estimators do not lose much despite the heteroscedastic nature of the problem, and 2) The collision estimators experience a plateau in performance when n exceeds d. The variational EM estimator also experiences a pronounced plateau when m is small, which is likely due to the poor approximation properties of the mean-field free entropy when m is small.

3. Regime C

Unlike Regimes A and B, which are connected to multinomial regression with measurement error, Regime C is closer in spirit to latent variable or blind source separation models, such as nonnegative matrix factorization [99], independent component analysis [100], latent class analysis [101], and particularly topic models, which we discuss in further detail below. Moment-based estimators are known to be information-theoretically optimal for many of these problems (e.g. [102] and references therein). Inspired by this line of work, our focus will be to derive a moment estimator for Regime C. We rely on assumption **(PT)** throughout this subsection.

a. The Moment Estimator

Recall from equation (B2) that the histogram entries Y_j are approximately independent, conditionally on the latent variable Π, and approximately follow a $\mathrm{Poi}(n\sum_{i=1}^k c_i\pi_{ij})$ distribution. Furthermore, as we recall in Lemma 26 below, the flat Dirichlet vector $\Pi_{i\cdot} = (\pi_{i1}, \ldots, \pi_{id})$ is equal in distribution to

$$\Pi_{i\cdot} = \frac{(X_{i1}, \ldots, X_{id})}{\sum_{j=1}^d X_{ij}}, \quad i = 1, \ldots, k,$$

where $X_{ij} \sim \mathcal{E}_d$ are i.i.d. exponential random variables. When the dimension d is large, the denominator in the above display concentrates rapidly around 1, and in this case one can further approximate the histogram Y by a random vector $\overline{Y}$ which follows the distribution:

$$\overline{Y}_j | X \sim \mathrm{Poi}\Big(n \sum_{i=1}^{k} c_i X_{ij}\Big), \quad j = 1, \dots, d.$$

Due to the independence of the rows of $X = (X_{ij})$, the random variables $\overline{Y}_j$ are marginally i.i.d.:

$$\overline{Y}_j \overset{\text{i.i.d.}}{\sim} \mathbf{Q}_c = \int_{\mathbb{R}_+^k} \mathrm{Poi}(n\langle \varpi, c\rangle) d\mathcal{E}_d^{\otimes k}(\varpi), \quad j = 1, \dots, d. \tag{B18}$$

We will show in the following section that model (B18) is statistically indistinguishable from the original model for Y, when the dimension d is sufficiently large. We will also show that model (B18) is identifiable up to sorting the entries of c, in the sense that for any $c, \bar{c} \in \Delta_k$, $\mathbf{Q} = \mathbf{Q}_{\bar{c}}$ implies that $W(c, \bar{c}) = 0$. Taking these facts for granted momentarily, we will derive a moment-based estimator of c motivated by model (B18).

Since we only seek to estimate c up to permutation of its entries, it will suffice to derive estimators for the first k moments of c, namely

$$m(c) = \big(m_1(c), \dots, m_k(c)\big)^\top, \quad \text{with } m_p(c) = \sum_{i=1}^{k} c_i^p, \quad p = 1, \dots, k.$$

These moments uniquely identify c up to permutation of its entries (cf. [74]), since the polynomial

$$f_c(z) = \prod_{i=1}^{k} (z - c_i), \quad z \in \mathbb{C},$$

is uniquely determined, on the one hand, by its set of roots $\{c_1, \dots, c_k\}$, and on the other hand, by its moment vector $m(c)$. This fact follows from *Newton's identities* (recalled in Appendix K 2 a), which imply

$$f_c(z) = z^k + \sum_{j=1}^{k} (-1)^j e_j(c) z^{k-j}, \quad z \in \mathbb{C},$$

$$\text{with } e_0(c) = 1, \quad e_\ell(c) = \frac{k}{\ell} \sum_{j=1}^{\ell} (-1)^{j-1} e_{\ell-j}(c) m_j(c), \quad \ell = 1, \dots, k.$$

Given estimators of the moments of c, say $\widehat{m}_1, \dots, \widehat{m}_k$, which we will define below, one can now construct an estimator of c by forming an estimator of f_c, and returning its k roots. Concretely, define

$$\widehat{e}_0 = 1, \quad \widehat{e}_\ell = \frac{k}{\ell} \sum_{j=1}^{\ell} (-1)^{j-1} \widehat{e}_{\ell-j} \widehat{m}_j, \quad \ell = 1, \dots, k. \tag{B19}$$

and define the fitted k-degree polynomial

$$\widehat{f}(z) = z^k + \sum_{j=1}^{k} (-1)^j \widehat{e}_j z^{k-j}, \quad z \in \mathbb{C}.$$

We then define a pilot estimator $\widetilde{c} = (\widetilde{c}_1, \dots, \widetilde{c}_k)$ as the vector of k (possibly complex) roots of $\widehat{f}$, ordered arbitrarily. To obtain an estimator with real coordinates, we define our final estimator by

$$\widehat{c} = \big(\mathrm{Re}(\widetilde{c}_1), \dots, \mathrm{Re}(\widetilde{c}_k)\big). \tag{B20}$$

We refer the above as the **moment estimator** for Regime C. In order to complete its definition, we need to define the estimators $\widehat{m}_1, \dots, \widehat{m}_k$, which we turn to next.

Our starting point is inspired by the past works of [1, 11, 23, 32], which noted that when $k = 2, 3$, the *cumulants* of the histogram are related to the moments of c. Let $\kappa_p(X)$ denote the p-th cumulant of any random variable X. Given independent exponential random variables $\varpi_1, \dots, \varpi_k \sim \mathcal{E}_d$, define the random variable $\theta = \sum_{i=1}^k c_i \varpi_i$, and write $\xi_p = \kappa_p(\theta)$ for any $p \geq 2$. Recalling that cumulants are additive across sums of independent random variables, we have

$$\xi_p = \sum_{i=1}^k c_i^p \kappa_p(\varpi_i) = \frac{(p-1)!}{d^p} m_p(c). \tag{B21}$$

Thus, to estimate the moment vector $m(c) := (m_1(c), \dots, m_k(c))^\top$, it suffices to estimate the cumulant vector $\xi = (\xi_1, \dots, \xi_k)^\top$. To do so, we recall that the cumulants $\xi_p = \kappa_p(\theta)$ are related to the moments $\eta_p = \mathbb{E}[\theta^p]$ via

$$\xi_p = \sum_{\ell=1}^p (-1)^{\ell-1} (\ell-1)! B_{p,\ell}\big(\eta_1, \dots, \eta_{p-\ell+1}\big), \tag{B22}$$

where $B_{p,\ell}$ are the Bell polynomials (cf. Appendix K 2 c), defined by

$$B_{p,\ell}(x_1, \dots, x_{p-\ell+1}) = p! \sum_{(h_1,\dots,h_{p-\ell+1}) \in \mathcal{H}_{p,\ell}} \prod_{i=1}^{p-\ell+1} \frac{x_i^{h_i}}{(i!)^{h_i} h_i!}, \tag{B23}$$

where $\mathcal{H}_{p,\ell}$ consists of all tuples $(h_1, \dots, h_{p-\ell+1})$ of nonnegative integers such that:

$$\sum_{i=1}^{p-\ell+1} h_i = \ell, \quad \sum_{i=1}^{p-\ell+1} i h_i = p.$$

These expressions suggest that the cumulants ξ_p can be estimated by first estimating the moments η_ℓ, which in turn can be done using the classical unbiased estimators for a Poisson model, given by

$$T_{j,\ell} = \begin{cases} 1/d, & \ell = 1, \\ \frac{Y_j!}{n^\ell (Y_j - \ell)!}, & \ell = 2, 3, \dots \end{cases}$$

Specifically, one has $\eta_\ell = \mathbb{E}[T_{j,\ell}]$ for all $\ell = 1, 2, \dots$. We can use these quantities to build estimators for each term in the summations (B22)–(B23). To this end, given $\mathbf{h} \in \mathcal{H}_{p,\ell}$, let $\mathcal{I}_{\ell,\mathbf{h}}$ be the set of all tuples $(S_1, \dots, S_{p-\ell+1})$ consisting of pairwise disjoint sets $S_1, \dots, S_{p-\ell+1} \subseteq \{1, \dots, d\}$ such that $|S_i| = h_i$ for all i. Notice that some of the S_i could be empty, and that $\big|\bigcup_i S_i\big| = p$ by definition of $\mathcal{H}_{p,\ell}$. Furthermore,

$$|\mathcal{I}_{\ell,\mathbf{h}}| = \frac{d!}{(d-p)! \prod_{i=1}^k h_i!},$$

where the second factor denotes a multinomial coefficient. Now, define the generalized U-Statistic:

$$W_{\mathbf{h}} = \frac{(d-p)!\prod_{i=1}^{k} h_i!}{d!} \sum_{(S_1,\ldots,S_{p-\ell+1})\in\mathcal{I}_{\ell,\mathbf{h}}} \prod_{i=1}^{p-\ell+1} \prod_{j\in S_i} T_{j,i}.$$

If the Poisson model (B18) held true, then the above would be an unbiased estimator of $\prod_{i=1}^{p-\ell+1} \eta_i^{h_i}$. A natural estimator of $B_{p,\ell}(\eta_1,\ldots,\eta_{p-\ell+1})$ is thus the following:

$$p! \sum_{\mathbf{h}\in\mathcal{H}_{p,\ell}} \frac{W_{\mathbf{h}}}{\prod_{i=1}^{p-\ell+1}(i!)^{h_i} h_i!}.$$

Combining these ideas, a natural estimator of ξ_p is given by:

$$\begin{aligned}\widehat{\xi}_p &= p! \sum_{\ell=1}^{p} (-1)^{\ell-1}(\ell-1)! \sum_{\mathbf{h}\in\mathcal{H}_{p,\ell}} \frac{W_{\mathbf{h}}}{\prod_{i=1}^{p-\ell+1}(i!)^{h_i} h_i!} \\ &= \frac{1}{\binom{d}{p}} \sum_{\ell=1}^{p} (-1)^{\ell-1}(\ell-1)! \sum_{\mathbf{h}\in\mathcal{H}_{p,\ell}} \sum_{(S_1,\ldots,S_{p-\ell+1})\in\mathcal{I}_{\ell,\mathbf{h}}} \prod_{i=1}^{p-\ell+1} \prod_{j\in S_i} \frac{T_{j,i}}{i!}.\end{aligned} \tag{B24}$$

By combining this expression with equation (B21), we arrive at the following estimator of $m_p(c)$ (which is unbiased under the Poisson model (B18)):

$$\widehat{m}_p = \frac{d^p}{(p-1)!\binom{d}{p}} \sum_{\ell=1}^{p} (-1)^{\ell-1}(\ell-1)! \sum_{\mathbf{h}\in\mathcal{H}_{p,\ell}} \sum_{(S_1,\ldots,S_{p-\ell+1})\in\mathcal{I}_{\ell,\mathbf{h}}} \prod_{i=1}^{p-\ell+1} \prod_{j\in S_i} \frac{T_{j,i}}{i!}, \quad p = 1,\ldots,k. \tag{B25}$$

Together with equation (B20), this completes our definition of the moment estimator $\widehat{c}^{\,\mathrm{mom}}$. Despite its unwieldy definition, this estimator can be readily implemented in closed form, with the exception of a root-finding step.

In what follows, we present an upper bound on the risk of the moment estimator under the simplified model (B18). In Appendix E 3, we will then show that this upper bound readily extends to our original multinomial sampling model.

Proposition 4. *Given $\gamma > 0$ arbitrarily small, let $n, d, k \geq 1$ satisfy $d^{1-\frac{1}{k}} \leq n \leq d^{\frac{1}{1+\gamma}}$. Assume that condition **(PT)** holds. Then, under model (B18), there exists a constant $C = C(k,\gamma) > 0$ such that*

$$\sup_{c\in\Delta_k} \mathbb{E}_c\Big[W(\widehat{c}^{\,\mathrm{mom}}, c)\Big] \leq C\sqrt{\frac{d^{1-\frac{1}{k}}}{n}}.$$

The proof appears in Appendix E 2. This result shows that the moment estimator achieves the minimax optimal rate of convergence stated in Theorem 2, under the sorted loss function W. Proposition 4 merely studies the estimation rate in the narrow regime $d^{1-1/k} \leq n \leq d^{1-\epsilon}$, for ϵ arbitrarily small. When n falls below $d^{1-1/k}$, Theorem 2 implies that consistent estimation is not possible uniformly over Δ_k. The regime $n > d$ is less relevant for RCS experiments, and in either case, finer estimators would need to be adopted in this regime since our approximate model (B18) becomes unrealistic when $n > d$ (as we discuss further in Appendix C).

Our next results will show that the rate of convergence in Proposition 4 can be significantly improved if mild constraints are placed on the error vector $c \in \Delta_k$. Indeed, we will see that the convergence behavior of the moment estimator is highly heterogeneous across the

parameter space Δ_k, and improves whenever the errors c_i have *different magnitudes*. On a technical level, these improvements can be anticipated from the fact that the root-finding step in the moment estimator is better conditioned when the underlying roots are well-separated (cf. Lemma 29 below for a quantitative statement of this fact).

To elaborate, given $1 \leq k_0 < k$, let Δ_{k,k_0} denote the set of all elements $c^\star \in \Delta_k$ which admit exactly k_0 distinct entries, that is, for which the set $\{c_i^\star : 1 \leq i \leq k\}$ has cardinality equal to k_0. Furthermore, let $\Delta_{k,k_0}(\delta)$ denote the set of elements $c^\star \in \Delta_{k,k_0}$ such that for all $1 \leq i < i' \leq k$, either $c_i^\star = c_{i'}^\star$, or $|c_i^\star - c_{i'}^\star| \geq \delta$. The following result characterizes the minimax rate of estimating vectors c which are in the vicinity of an element in $\Delta_{k,k_0}(\delta)$; roughly-speaking, this means that the vector c has entries which tightly cluster around k_0 different values.

Proposition 5. *Assume the same conditions as Proposition 4. Then, under model (B18), there exists a constant $\epsilon > 0$ depending on k such that for any $\delta > 0$, there exists $C = C(k, \gamma, \delta) > 0$ such that for all $1 \leq k_0 \leq k$,*

$$\sup_{c^\star \in \Delta_{k,k_0}(\delta)} \sup_{\substack{c \in \Delta_k \\ W(c,c^\star) \leq \epsilon}} \mathbb{E}_c\Big[W(\widehat{c}^{\,\mathrm{mom}}, c)\Big] \leq C\left(\frac{d^{k-1}}{n^k}\right)^{\frac{1}{2(k-k_0+1)}}.$$

Notice that Proposition 5 recovers the convergence rate of Proposition 4 when $k_0 = 1$, which corresponds to no separation assumptions. On the other hand, it is strictly faster when $k_0 < k$, and improves as k_0 increases. In the most extreme case where all errors c_i are well-separated, corresponding to $k_0 = k$, we find that the error vector c can be estimated at the rate $\sqrt{\frac{d^{k-1}}{n^k}}$, assuming again that d exceeds n. This dependence on the separation of the elements of c, as measured by the parameter k_0, is qualitatively related to the minimax rate of estimating finite mixture models under partial separation assumptions [74, 103–108].

We next state a final refinement of the minimax estimation rate of c. Building upon Refs. [74, 109], we will now show that, not only can the entire vector c be estimated at faster rates when the coordinates of c are partially separated, but some of the individual coordinates of c enjoy even faster rates than those shown in Proposition 5. We will prove this by showing that Proposition 5 continues to hold when W is replaced by a stronger loss function, which sharply captures the heterogeneity in parameter estimation across the vector c. To elaborate, let $c^\star \in \Delta_{k,k_0}$ be given, and assume that its entries are listed in decreasing order. Recall that the k entries of $c^\star$ are assumed to take on k_0 distinct values, which we denote by $v_1 > \cdots > v_{k_0}$. Let $1 \leq a_j \leq k$ denote the smallest index $i \in \{1, \ldots, k\}$ such that $c_i^\star = v_j$, and let r_j denote the number of entries in c which are equal to v_j: $r_j = a_{j+1} - a_j$. Furthermore, for all $i = 1, \ldots, k$, let $j_i \in \{1, \ldots, k_0\}$ denote the unique index such that $a_{j_i} \leq i < a_{j_i+1}$.

Given elements $c, c' \in \Delta_k$, whose entries we again assume are in decreasing order, we write

$$\mathcal{D}_{c^\star}(c, c') = \sum_{j=1}^{k_0} \left\| c_{a_j:(a_{j+1}-1)} - c'_{a_j:(a_{j+1}-1)} \right\|_2^{r_j}, \tag{B26}$$

where $c_{a_j:(a_{j+1}-1)} \in \mathbb{R}^{a_{j+1}-a_j}$ is the vector with coordinates $c_{a_j}, \ldots, c_{a_{j+1}-1}$. By abuse of notation, when c, c' are not ordered, we still write $\mathcal{D}_{c^\star}(c, c')$ as a shorthand for $\mathcal{D}_{c^\star}(\mathrm{ord}(c), \mathrm{ord}(c'))$, where $\mathrm{ord}(c)$ is the vector consisting of the same coordinates as c, in decreasing order. Before interpreting this divergence further, let us state our final result.

Proposition 6. *Assume the same conditions as Proposition 5. Then, under model (B18), for any $1 \leq k_0 \leq k$ and $\delta > 0$, there exist constants $C, \epsilon > 0$ depending on k, γ, δ such that*

$$\sup_{c^\star \in \Delta_{k,k_0}(\delta)} \sup_{\substack{c \in \Delta_k \\ W(c,c^\star) \leq \epsilon}} \mathbb{E}_c\Big[\mathcal{D}_{c^\star}(\widehat{c}^{\,\mathrm{mom}}, c)\Big] \leq C\sqrt{\frac{d^{k-1}}{n^k}}.$$

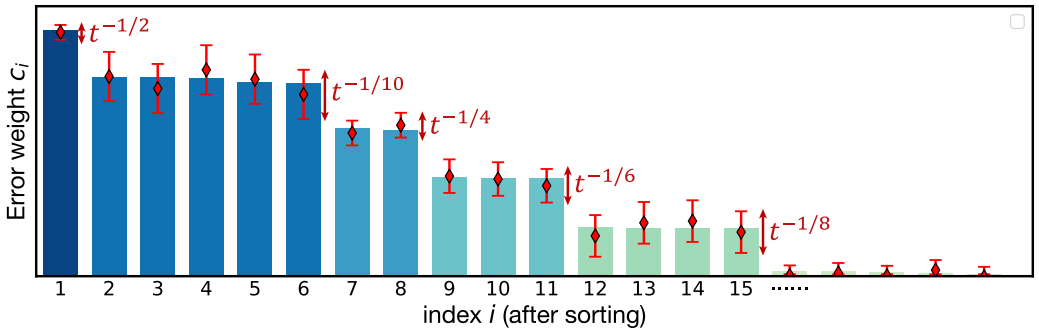

Fig. S3. Illustration of the local minimax upper bound stated in Proposition 6, with $t = n^k/d^{k-1}$. Different entries of the vector c can be estimated at different rates, depending on their local separation structure, thus significantly improving the rate $t^{-1/2k}$ of Proposition 4, which is only sharp when all entries of c_i are near $1/k$.

An implication of Propositions 5–6 is the following. Given $c^\star \in \Delta_{k,k_0}$ and a parameter c in a small neighborhood of $c^\star$, every coordinate of the moment estimator satisfies:

$$\mathbb{E}|\widehat{c}_i^{\,\mathrm{mom}} - c_i| \lesssim \left(\frac{d^{k-1}}{n^k}\right)^{\frac{1}{2r_{j_i}}}, \quad i = 1, \ldots, k. \tag{B27}$$

For example, when $k_0 = 1$, so that no separation assumptions are imposed, r_{j_i} must always be equal to k, and we recover the rate of convergence stated in Proposition 4. When $k_0 = k$, so that all coordinates of c are well-separated, we must always have $r_{j_i} = 1$ and we recover the rate of convergence stated in Proposition 4. When $1 < k_0 < 1$, the values of r_{j_i} are always bounded from above by $k - k_0 + 1$, thus Proposition 6 is never worse than Proposition 5, but generally provides different upper bounds for estimating different elements c_i, depending on their local separation structure.

This refined convergence rate is particularly well-suited to our problem, since we typically expect the fidelity parameter to be appreciably larger than the remaining parameters. If we assume that $c_1 = \|c\|_\infty$ denotes the fidelity, and if we make the reasonable assumption that $c_2, \ldots, c_k < c_1 - \delta$ for a fixed k and fixed $\delta > 0$, then the above result implies $r_1 = 1$, $j_1 = 1$, and the fidelity can then be estimated at the following fast rate:

$$\mathbb{E}|\widehat{c}_1^{\,\mathrm{mom}} - c_1| \lesssim \sqrt{\frac{d^{k-1}}{n^k}}.$$

This is the basis for Proposition 1 of the main text. It is worth emphasizing that these various local convergence rates are achieved *adaptively*: The moment estimator does not rely on any prior information about the possible separation among the atoms of c, and achieves these multiscale convergence rates automatically.

Related Work. Our model in Regime C is closely-related to to a class of hierarchical models for text analysis known as *topic models* [110, 111]. The typical setup of a topic model is to observe a collection of M random variables of the form:

$$Y^{(1)} \sim \mathrm{Mult}(n; \Pi^\top c^{(1)})$$
$$Y^{(2)} \sim \mathrm{Mult}(n; \Pi^\top c^{(2)})$$
$$\vdots$$

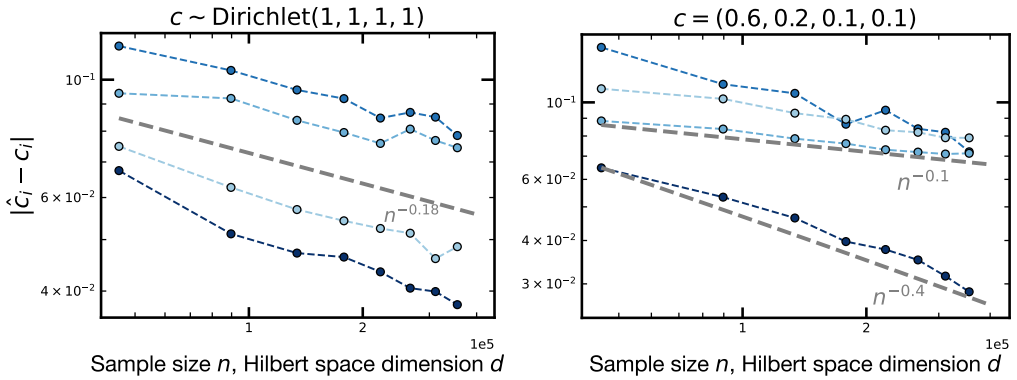

Fig. S4. Sample complexity of the moment estimator in Regime C, with $k = 4$. We arrange the elements of $\widehat{c}$ and c in decreasing order, and plot the average errors $\mathbb{E}|\widehat{c}_i - c_i|$ for $i = 1, \ldots, 4$, with $i = 1$ denoted by the darkest color, and $i = 4$ by the lightest. Across 30 sample sizes $n = d$, we generate 500 replications from each model, under multinomial sampling.

$$Y^{(M)} \sim \text{Mult}(n; \Pi^\top c^{(M)}),$$

where $\Pi \in \mathbb{R}^{k \times d}$ is an unknown stochastic matrix, which is common across observations, and $c^{(i)} \in \Delta_k$ are unknown mixing weights. The goal is to estimate Π, or the matrix $C \in \mathbb{R}^{k \times M}$ comprised of columns $c^{(i)}$, for $i = 1, \ldots, M$. As stated, this problem is not statistically identifiable without further modeling assumptions. A sufficient condition for identifiability is the so-called *anchor word* assumption [99, 112] on the matrix Π, which forms the basis for a wide array of frequentist methods for topic modeling [69, 112–118]. A second approach consists of treating the matrices Π and C as latent variables, endowed with some prior distribution, which makes the model identifiable under very general conditions [119]. The most widely-used method in this second category is *latent Dirichlet allocation* (LDA; [70]), which consists of placing Dirichlet prior distributions on the columns of $\Pi^\top$ and C, and performing inference for these objects via their posterior distribution, typically approximated using variational inference.

When $M = 1$, our model is closely-related to this second line of literature, since we assume that the rows of the matrix Π are drawn from the flat Dirichlet distribution. Despite the wide practical adoption of the LDA model, we are only aware of a few references that analyze the sample complexity of parameter estimation in this model, most of which focus on estimating Π rather than C [113, 120, 121]. The very recent work of [119] studies posterior contraction rates for Bayesian estimators of C, but does so under pointwise and fixed-dimensional asymptotics which are incomparable to our setting.

b. *Numerical Comparison in Regime C*

Although the moment estimator is currently our only practical proposal for addressing Regime C, we close this Appendix by briefly reporting a simulation study to illustrate its empirical sample complexity, in Figure S4. We consider two models: In the first, we draw c from a flat Dirichlet law on Δ_4 at each replication, thus placing c close to the uniform distribution, which is the least-favourable parameter from the standpoint of sample complexity. In the second model, we instead fix $c = (0.6, 0.2, 0.1, 0.1)$, in which case the fidelity 0.6 is appreciably larger than the remaining entries. In the first case, we nearly observe the worst-case rate $n^{-1/8} = 0.125$ predicted by Proposition 4, across all entries of $\widehat{c}$. In the latter case, we instead see significant variation between the convergence rate of the

various parameters. In particular, the estimated fidelity $\widehat{c}_1$ nearly achieves the $n^{-1/2}$ rate predicted by Proposition 6.

Appendix C: Equivalent Statistical Models

In this Appendix, we justify the model approximations which we made in the exposition of the previous section: We show that, when the dimension d is sufficiently large, our model is statistically equivalent to a simpler model in which the sample sizes n and m are Poissonized, and the entries of Π are taken to be independent exponential random variables without normalization. We will prove that these various simplifications do not appreciably alter the sample complexity of estimating c. We will then adopt the reduced model for much of the remainder of this manuscript, without loss of generality.

1. Statistical Models

Recall that the unsorted and sorted minimax estimation risks are defined by

$$\begin{aligned} \mathcal{M}(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c \|\widehat{c}(Z,W) - c\|, \\ \mathcal{M}_<(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c W\big(\widehat{c}(Z,W), c\big), \end{aligned} \tag{C1}$$

where the the infimum is taken over all Borel-measurable functions $\widehat{c}\colon \mathbb{R}^n \times \mathbb{R}^{k\times m} \to \mathbb{R}^k$, and the expectation $\mathbb{E}_c$ is taken over the *marginal* distribution of the random variables $Z = (Z_\ell)$ and $W = (W_{\ell i})$, that is, over the probability law

$$(Z,W) \sim \mathbb{E}_\Pi \left[\left(\sum_{i=1}^k c_i \Pi_{i\cdot} \right)^{\otimes n} \otimes \left(\bigotimes_{i=1}^k \Pi_{i\cdot} \right)^{\otimes m} \right],$$

where the expectation is to be interpreted as marginalization over the law of $\Pi \in \mathbb{R}^{k\times d}$, assuming that its rows are independently distributed according to the flat Dirichlet law $\mathcal{D}_d$. By abuse of notation, we identify the discrete density $\Pi_{i\cdot}$ with the probability distribution that it induces.

A set of sufficient statistics for the observations Z, W is given by the following histograms indexed by $j = 1, \ldots, d$, which we decorate with superscripts in this section only:

$$\begin{aligned} \widetilde{Y}_j &= \sum_{\ell=1}^n I(Z_\ell = j), \qquad (\widetilde{Y}_1, \ldots, \widetilde{Y}_d) \,\big|\, \Pi \sim \mathrm{Mult}(n; \Pi^\top c), \\ \widetilde{V}_{ij} &= \sum_{\ell=1}^m I(W_{\ell i} = j), \quad (\widetilde{V}_{i1}, \ldots, \widetilde{V}_{id}) \,\big|\, \Pi \sim \mathrm{Mult}(m; \Pi_{i\cdot}), \quad i = 1, \ldots, k. \end{aligned}$$

The joint distribution of the random variables $\widetilde{Y} = (\widetilde{Y}_j : 1 \le j \le d)$ and $\widetilde{V} = (\widetilde{V}_{ij} : 1 \le i \le k, 1 \le j \le d)$ is given by

$$\widetilde{\mathbf{Q}}_c = \mathbb{E}_\Pi \left[\mathrm{Mult}(n; \Pi^\top c) \otimes \bigotimes_{i=1}^k \mathrm{Mult}(m; \Pi_{i\cdot}) \right]. \tag{C2}$$

We will refer to the observation model (C2) as the **multinomial model**. It follows by sufficiency of the histograms $(\widetilde{Y}, \widetilde{V})$ that the minimax risks in equation (C1) can equivalently

be written as

$$\begin{aligned}\mathcal{M}(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}(\widetilde{Y},\widetilde{V}) - c\|,\\ \mathcal{M}_<(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c W\big(\widehat{c}(\widetilde{Y},\widetilde{V}), c\big),\end{aligned} \tag{C3}$$

where the infimum is over all Borel-measurable maps $\widehat{c}$ from $\mathbb{R}^d \times \mathbb{R}^{k\times d}$ into $\mathbb{R}^k$.

Let us now introduce two alternative sampling models which have comparable minimax risks to the multinomial model, which we already alluded to in Section B. The first of these models will be referred to as the **normalized Poisson model**, under which the practitioner observes random vectors $(\overline{Y},\overline{V})$ drawn from the following joint distribution:

$$\overline{\mathbf{Q}}_c = \mathbb{E}_\Pi\left[\bigotimes_{j=1}^d \left(\mathrm{Poi}(n\Pi_{\cdot j}^\top c)\otimes \bigotimes_{i=1}^k \mathrm{Poi}(m\pi_{ij})\right)\right]. \tag{C4}$$

Model (C4) can be viewed as a variant of model (C2) in which the sample size n is replaced by a Poisson random variable $N\sim\mathrm{Poi}(n)$, drawn independently of all other random variables. The unordered and ordered minimax risks under the normalized Poisson model are given by:

$$\begin{aligned}\mathcal{R}'(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}(\overline{Y},\overline{V}) - c\|,\\ \mathcal{R}'_<(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c W\big(\widehat{c}(\overline{Y},\overline{V}), c\big),\end{aligned} \tag{C5}$$

where the expectation is taken over $(\overline{Y},\overline{V})\sim\overline{\mathbf{Q}}_c$.

Our final model is the **unnormalized Poisson model**, in which the practitioner observes random variables (Y,V) drawn from the product measure $\mathbf{Q}_c^{\otimes d}$, where

$$\mathbf{Q}_c = \mathbb{E}_\varpi\left[\mathrm{Poi}(n\langle\varpi,c\rangle)\otimes\bigotimes_{i=1}^k \mathrm{Poi}(m\varpi_i)\right]. \tag{C6}$$

Here, the expectation is taken over a random variable $\varpi\sim\mathcal{E}_d^k$ consisting of independent $\mathrm{Exp}(d)$ entries. The unnormalized Poisson model can be viewed as a proxy of the normalized Poisson model, in which the flat Dirichlet law of the rows of Π are approximated by the law of ϖ. The minimax risks under this model are defined by

$$\begin{aligned}\mathcal{R}(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}(Y,V) - c\|,\\ \mathcal{R}_<(n,d,k,m) &= \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c W\big(\widehat{c}(Y,V), c\big),\end{aligned} \tag{C7}$$

where the expectations are now taken over the law of a pair (Y,V) drawn from $\mathbf{Q}_c^{\otimes d}$.

2. Equivalence of Minimax Risks

Let us begin by showing that the minimax risks for the multinomial and unnormalized Poisson models are comparable.

Lemma 1. *There exists a universal constant $C>0$ such that for all $n,m,d,k\geq 1$,*

$$\begin{aligned}&\mathcal{M}(2n,d,k,2m)\cdot\big(1 - Ce^{-n/C} - C\mathcal{I}ke^{-m/C}\big)\\ &\qquad\leq \mathcal{R}'(n,d,k,m) \leq \mathcal{M}(n/2,d,k,m/2) + Ce^{-n/C} + C\mathcal{I}ke^{-m/C},\end{aligned}$$

where $\mathcal{I} = I(m=0)$. An analogous assertions holds for the sorted minimax risks.

The proof appears in Appendix G 1 a, and is based on a well-known Poissonization technique [122, 123]. As a result of Lemma 1, we find that the minimax risks $\mathcal{M}$ and $\mathcal{R}'$ are comparable, up to additive error terms which decay exponentially with n and m. The following result further relates the normalized and unnormalized Poisson models, but this time at the level of their distributions rather than their induced risks.

Lemma 2. *There exists a universal constant $C > 0$ such that for all $n, d, m, k \geq 1$, and $c \in \Delta_k$,*

$$\mathrm{TV}(\overline{\mathbf{Q}}_c, \mathbf{Q}_c^{\otimes d}) \leq C\left(\sqrt{\frac{n}{d}} + \sqrt{\frac{mk}{d}}\right).$$

The proof appears in Appendix G 1 b. By definition of the total variation distance, one can find a coupling between any random pairs $(\overline{Y}, \overline{V}) \sim \overline{\mathbf{Q}}_c$ and $(Y, V) \sim \mathbf{Q}_c^{\otimes d}$ such that the equality $(Y, V) = (\overline{Y}, \overline{V})$ holds with probability at least $1 - \mathrm{TV}(\overline{\mathbf{Q}}_c, \mathbf{Q}_c^{\otimes d})$. Due to the boundedness of the parameter space Δ_k, it then follows from Lemma 2 that

$$\mathcal{R}(n, d, k, m) = \mathcal{R}'(n, d, k, m) + O\left(\sqrt{\frac{n}{d}} + \sqrt{\frac{mk}{d}}\right). \tag{C8}$$

Together with Lemma 1, the above bound will allow us to reduce the problem of bounding the minimax risk $\mathcal{M}$ to that of bounding the risk $\mathcal{R}$, at least whenever the scaling of these risks is of lower order than $\sqrt{n/d} + \sqrt{mk/d}$. Furthermore, we will use Lemma 2 directly in our development of lower bounds on the various minimax risks.

Remark 1 (Sharpness of Lemma 2)**.** It is worth noting that the upper bound of Lemma 2 can perhaps be improved quadratically, but not further. To elaborate, let us first note that, by Lemma 26, we may write

$$\mathbf{Q}_c^{\otimes d} = \mathbb{E}_{\Pi, G}\left[\bigotimes_{j=1}^{d} \mathrm{Poi}\big(n \textstyle\sum_{i=1}^{k} c_i G_i \pi_{ij}\big) \otimes \bigotimes_{i=1}^{k} \mathrm{Poi}(m\pi_{ij})\right],$$

where $G = (G_1, \ldots, G_k)^\top$ is a vector of independent $\mathrm{Gamma}(d, d)$-distributed random variables. Now, let us consider a proxy of this model and of the model $\overline{\mathbf{Q}}_c$ in which $k = 1$, and $m = 0$. Let Y and $\overline{Y}$ be drawn from these corresponding models:

$$Y \sim \mathbb{E}_{\Pi, G}\left[\bigotimes_{j=1}^{d} \mathrm{Poi}(nG\pi_{1j})\right], \quad \overline{Y} \sim \mathbb{E}_{\Pi}\left[\bigotimes_{j=1}^{d} \mathrm{Poi}(n\pi_{1j})\right].$$

Define the random variables

$$S = \sum_{j=1}^{d} Y_j \sim \mathrm{NB}(d, (1 + n/d)^{-1}), \quad \text{and,} \quad \overline{S} = \sum_{j=1}^{d} \overline{Y}_j \sim \mathrm{Poi}(n)$$

where we obtained the law of S by noting that $\sum_j Y_j \,|\, G \sim \mathrm{Poi}(nG)$, and any Poisson mixture with Gamma mixing measure has negative binomial distribution (cf. Lemma 26). One has

$$\mathrm{KL}\big(\mathrm{Law}(\overline{Y}) \,\big\|\, \mathrm{Law}(Y)\big) = \mathrm{KL}\big(\mathrm{Law}(\overline{S}) \,\big\|\, \mathrm{Law}(S)\big) = \mathrm{KL}\big(\mathrm{NB}(d, (1 + n/d)^{-1})\big) \,\big\|\, \mathrm{Poi}(n)\big).$$

Now, using the weak-lower semicontinuity of the KL divergence, and taking $d \to \infty$ such that $\tau = n/d$, one has the Gaussian approximation

$$\mathrm{KL}\big(\mathrm{NB}(d, (1 + n/d)^{-1})\big) \,\big\|\, \mathrm{Poi}(n)\big) \geq \mathrm{KL}\big(N(d\tau, d\tau(1 + \tau)), N(d\tau, d\tau)\big) + o(1)$$

$$= \frac{1}{2}\log\frac{1}{1+\tau} + \frac{1+\tau}{2} - \frac{1}{2} + o(1)$$
$$= \frac{1}{2}\big(\tau - \log(1+\tau)\big) = O(\tau^2).$$

This suggests that $\mathrm{KL}(\mathrm{Law}(\overline{Y}) \,\big\|\, \mathrm{Law}(Y))$, and hence $\mathrm{TV}^2(\overline{\mathbf{Q}}_c \| \mathbf{Q}_c^{\otimes d})$, cannot scale faster than $\tau^2 \asymp (n/d)^2$, thus suggesting that, at least for $k = 1$ and $m = 0$, Lemma 2 can only be improved quadratically, with no change in the trade-off between n and d.

Remark 2 (Mixtures of Products)**.** From a technical lens, our reduction from model $\widetilde{\mathbf{Q}}_c$ to model $\mathbf{Q}_c^{\otimes d}$ will be fruitful since the former is a mixture of product distributions, whereas the latter is a product of mixture distributions, which is significantly simpler to handle. We refer to the recent manuscript [119] for a more systematic comparison of mixtures-of-products and products-of-mixtures in discrete latent variable models.

3. Identifiability

Having reduced our problem to that of controlling the minimax risk under the unnormalized Poisson model, we will now establish the identifiability of that model, even in the absense of side information.

Lemma 3. *For all $c, \bar{c} \in \Delta_k$, the following assertions hold.*

(i) Assume $m = 0$. Then, $\mathbf{Q}_c = \mathbf{Q}_{\bar{c}}$ implies $W(c, \bar{c}) = 0$.

(ii) Assume $m \geq 1$. Then, $\mathbf{Q}_c = \mathbf{Q}_{\bar{c}}$ implies $c = \bar{c}$.

Proof. Mixtures of products of Poisson measures are identifiable in terms of their mixing measures [124, 125]. To prove the claim, it will therefore suffice to establish the identifiability of the collection $\{\mu_c : c \in \Delta_k\}$ of mixing measures with respect to the parameter c, where

$$\mu_c := \mathrm{Law}(n\langle \varpi, c\rangle) \otimes \bigotimes_{i=1}^{k} \mathrm{Law}(m\varpi_i), \quad c \in \Delta_k,$$

and $\varpi \sim \mathcal{E}_d^{\otimes k}$. The law of μ_c is uniquely characterized by its multivariate characteristic function, which is given by

$$\varphi_c(t, s) = \mathbb{E}_{\varpi}\left[\exp\left(\mathbf{i}tn\langle \varpi, c\rangle + \mathbf{i}m\langle \varpi, s\rangle\right)\right], \quad t \in \mathbb{R}, s \in \mathbb{R}^k,$$

with $\mathbf{i} = \sqrt{-1}$. One has

$$\varphi_c(t, s) = \prod_{i=1}^{k} \mathbb{E}\left[\exp\left(\mathbf{i}(ntc_i + ms_i)\varpi_i\right)\right] = \prod_{i=1}^{k} \frac{d}{d - \mathbf{i}(ntc_i + ms_i)}.$$

The right-hand side of the above display defines the reciprocal of a polynomial in t and s. By setting $s = 0$, this polynomial depends only on the univariate parameter t, and is uniquely characterized by its unordered collection of roots, namely $\{d/(\mathbf{i}nc_i) : 1 \leq i \leq k\}$. It follows that φ_c, and hence μ_c, is uniquely characterized by the unordered collection of entries of c, and claim (i) readily follows from this observation. Furthermore, when $m \geq 1$, the equality $\varphi_c(t, s) = \varphi_{\bar{c}}(t, s)$ can only hold uniformly over all $(t, s) \in \mathbb{R} \times \mathbb{R}^k$ if $c = \bar{c}$, which proves part (ii). □

With these various reductions in place, we are now in a position to prove the main results of this manuscript, beginning with sample complexity lower bounds.

Appendix D: Proofs of Lower Bounds

The goal of this Appendix is to prove the lower bounds on the sample complexity stated in Theorems 1–2.

1. Preliminaries

The bulk of our lower bound arguments will be contained in the following two Lemmas, which characterize the divergence between elements of our model in terms of their parameter separation. In what follows, for any given $k \geq 3$, $1 \leq s \leq k-2$, and $0 \leq \beta \leq 1/s$, we denote by $\Sigma_{k,s}(\beta)$ the set of elements $c \in \Delta_k$ for which exactly s of the first $k-2$ entries of c are equal to β, and the last two entries of c are given by $c_{k-1} = c_k = (1-\beta s)/2$. Our first result deals with the regime where m is polynomially smaller than d, under the unnormalized Poisson model.

Lemma 4. *Let $n, m, d, k \geq 1$ satisfy the conditions $2 \leq k \leq d$, $(n+m)^{1+\gamma} \leq d$, and $k^{1+\gamma} \leq d/m$, for an arbitrarily small constant $\gamma > 0$. Let $c, \bar{c} \in \Delta_k$ be two vectors such that*

$$m_j(c) = m_j(\bar{c}), \quad \text{for all } j = 0, \dots, k-1.$$

Furthermore, let $q_1 := |m_k(\bar{c}) - m_k(c)|$ and $q_2 = \|\bar{c} - c\|_2$. Assume $q_2 \geq 1/d$. Then, there exist constants $C_1 = C_1(\gamma, k) > 0$ and $C_2 = C_2(\gamma) > 0$ such that the following assertions hold.

1. *We have,*

$$\mathrm{TV}(\mathbf{Q}_{\bar{c}}^{\otimes d}, \mathbf{Q}_c^{\otimes d}) \leq C_1 n^{\frac{k}{2}} d^{-\frac{k-1}{2}} q_1 + C_2 q_2 \sqrt{\frac{nm}{d}}.$$

2. *Assume $W(c, \bar{c}) = 0$. Then,*

$$\chi^2(\mathbf{Q}_{\bar{c}}^{\otimes d}, \mathbf{Q}_c^{\otimes d}) \leq C_2 q_2^2 \frac{nm}{d}.$$

The proof of Lemma 4 appears in Appendix D 4. Our next result provides an upper bound which is effective in the regime $m \geq d$. In this case, we directly analyze the multinomial model.

Lemma 5. *There exists a universal constant $C > 0$ such that if $3 \leq k \leq d$, then for all $1 \leq s \leq k-2$, $0 \leq \beta \leq s$, $n, m \geq 1$, and $c, \bar{c} \in \Sigma_{k,s}(\beta)$,*

$$\mathrm{KL}(\widetilde{\mathbf{Q}}_c \| \widetilde{\mathbf{Q}}_{\bar{c}}) \leq Cn \|\bar{c} - c\|_2^2.$$

Furthermore, for all $k \geq 2$, there exists a constant $C_k > 0$ such that for all $m, n, d \geq 1$ and $c, \bar{c} \in \Delta_k$ satisfying $\|c\|_\infty \vee \|\bar{c}\|_\infty \leq 3/4$,

$$\mathrm{KL}(\widetilde{\mathbf{Q}}_c \| \widetilde{\mathbf{Q}}_{\bar{c}}) \leq C_k n \|\bar{c} - c\|_2^2.$$

The proof of Lemma 5 appears in Appendix D 5. A remarkable feature of Lemmas 4–5 is the fact that they exhibit a quadratic scaling with respect to $q_2 = \|\bar{c} - c\|_2^2$, similarly to divergences between Gaussian location models. This reflects the fact that, despite the Poissonian nature of our problem, the heteroscedasticity of our observations is mild, due to assumption **(PT)**.

Before proving these two results, let us show how they lead to our various minimax lower bounds.

2. Minimax Lower Bound for the Sorted Loss Function

Our aim in this section is to prove the following minimax lower bound.

Proposition 7. *Let $d, k, m, n \geq 1$ satisfy condition (S). Then, there exists a constant $C_{k,\gamma} > 0$ such that*

$$\mathcal{M}_<(n, d, k, m) \geq C_{k,\gamma} \left(\sqrt{\frac{d}{n(m + d^{1/k})}} + \frac{1}{\sqrt{n}} \right).$$

Proof of Proposition 7. By Le Cam's Lemma (cf. [126]), it will suffice to show that there exist parameters $c, \bar{c} \in \Delta_k$ such that

$$\mathrm{TV}(\widetilde{\mathbf{Q}}_c^{\otimes d}, \widetilde{\mathbf{Q}}_{\bar{c}}^{\otimes d}) \leq 1/2, \quad \text{and} \quad W(c, \bar{c}) \gtrsim \sqrt{\frac{d}{n(m + d^{1/k})}} + \frac{1}{\sqrt{n}}. \tag{D1}$$

Recall that, under condition (S), we either have $m^{1+\gamma} \leq d$ or $m > d$. Let us begin by proving the claim under the former condition. We will make use of the following Lemma to obtain least-favorable parameters $c, \bar{c}$. In what follows, recall that $m_p(u) = \frac{1}{k} \sum_{i=1}^k u_i^p$ for any $u \in \Delta_k$.

Lemma 6. *For any $k \geq 1$, there exists a constant $C_k > 0$ and vectors $u, v \in \mathbb{R}^k$ such that $\|u\|_\infty \vee \|v\|_\infty \leq 1$, and*

(i) $m_1(u) = 0$.

(ii) $m_p(u) = m_p(v)$ *for all* $p = 1, \ldots, k-1$.

(iii) $W(u, v) \wedge |m_k(u) - m_k(v)| \geq C_k$.

The proof of Lemma 6 appears in Appendix G 2 a. Now, given $\epsilon > 0$ and $u, v \in \mathbb{R}^k$ as in Lemma 6, define

$$c = \left(\frac{1}{k} + \epsilon u_1, \ldots, \frac{1}{k} + \epsilon u_k \right), \quad \bar{c} = \left(\frac{1}{k} + \epsilon v_1, \ldots, \frac{1}{k} + \epsilon v_k \right).$$

For all sufficiently small ϵ, the conditions of Lemma 6 ensure that $c, \bar{c}$ lie in the simplex Δ_k. Furthermore, we have $\|c - \bar{c}\|_2 \lesssim \epsilon$ and for all $p = 1, \ldots, k$,

$$\begin{aligned}
m_p(\bar{c}) - m_p(c) &= \sum_{i=1}^k \left[\left(\frac{1}{k} + \epsilon u_i \right)^p - \left(\frac{1}{k} + \epsilon v_i \right)^p \right] \\
&= \sum_{i=1}^k \sum_{j=1}^p \binom{p}{j} k^{-(p-j)} (u_i^j - v_i^j) \epsilon^j \\
&= \sum_{j=1}^p \binom{p}{j} k^{-(p-j)} \big(m_j(u) - m_j(v) \big) \epsilon^j \\
&= \epsilon^p \big(m_p(u) - m_p(v) \big),
\end{aligned}$$

where we used property (ii) of Lemma 6. Together with property (iii), we deduce that

$$|m_p(u) - m_p(v)| \asymp \epsilon^p \cdot I(p = k), \quad \text{for all } p = 1, \ldots, k.$$

We may therefore apply Lemma 4 with $q_1 \asymp \epsilon^k$ and $q_2 \asymp \epsilon$ to obtain

$$\mathrm{TV}(\mathbf{Q}_{\bar{c}}^{\otimes d}, \mathbf{Q}_c^{\otimes d}) \leq C_1 n^{\frac{k}{2}} d^{-\frac{k-1}{2}} \epsilon^k + C_2 \epsilon \sqrt{\frac{nm}{d}},$$

and thus, by Lemma 2, we deduce the following bound for the normalized Poisson model:

$$\mathrm{TV}(\overline{\mathbf{Q}}_{\bar{c}}^{\otimes d}, \overline{\mathbf{Q}}_c^{\otimes d}) \lesssim n^{\frac{k}{2}} d^{-\frac{k-1}{2}} \epsilon^k + \epsilon \sqrt{\frac{nm}{d}} + \sqrt{\frac{n}{d}} + \sqrt{\frac{mk}{d}}.$$

Under condition **(S)**, the above quantity is bounded by $1/2$ if we choose ϵ to be a sufficiently small multiple of $\sqrt{\frac{d}{n(d^{1/k}+m)}}$. On the other hand, Lemma 6 implies that $W(\bar{c}, c) = \epsilon W(u, v) \gtrsim \epsilon$. We thus deduce the following lower bound for the minimax risk in the normalized Poisson model, from Le Cam's Lemma:

$$\mathcal{R}'_<(n, d, k, m) \gtrsim \sqrt{\frac{d}{n(d^{1/k}+m)}}.$$

Finally, let us deduce a lower bound for the multinomial model. Notice that for any given (n, d, k), the map $m \mapsto \mathcal{M}_<(n, d, k, m)$ is monotonically decreasing, thus it suffices to prove the lower bound in the regime $m \geq d^{1/k}$. By combining the above display with Lemma 1, we have

$$\mathcal{M}_<(n, d, k, m) \gtrsim \mathcal{R}'_<(2n, d, k, 2m) - e^{-n/C} - e^{-m/C} \gtrsim \sqrt{\frac{d}{n(d^{1/k}+m)}},$$

where the final inequality uses the fact that the terms $e^{-n/C}$ and $e^{-m/C}$ are both of low order when $m \geq d^{1/k}$. This proves the claim in the regime $m^{1+\gamma} \leq d$. Finally, the claim for $m \geq d$ follows immediately from Le Cam's Lemma together with Lemma 5. □

3. Minimax Lower Bound for the Unsorted Loss Function

Our aim is now to derive lower bound for the unsorted minimax risk.

Proposition 8. *Assume that condition **(S)** holds with $k \geq 2$, and assume that $nm_d \geq d\log(k)$ with $m_d = \min\{m, d\}$. Then, there exists a constant $C_\gamma > 0$ such that*

$$\mathcal{M}(n, d, k, m) \geq C_\gamma \cdot \min\left\{ \left(\frac{d\log(k)}{nm_d}\right)^{\frac{1}{4}}, \left(\frac{dk}{nm_d}\right)^{\frac{1}{2}} \right\}.$$

Proof. Let us begin by proving the claim in the special case $k = 2$. Since the unsorted minimax risk is bounded from below by the sorted minimax risk, which in turn is always bounded from below by $1/\sqrt{n}$ (by Proposition 7), it suffices to consider the regime $m^{1+\gamma} < d$, and to prove the lower bound

$$\mathcal{M}(n, d, 2, m) \gtrsim \sqrt{\frac{d}{nm}}.$$

By Lemma 1, it further suffices to lower bound $\mathcal{R}'(n, d, 2, m)$, since the conditions $m^{1+\gamma} < d$ and $nm > d\log(k)$ imply that $n \wedge m \gtrsim d^\epsilon$ for some $\epsilon > 0$. To prove this lower bound, we

can reason similarly as in the proof of Proposition 7. Given $0 \le \epsilon, \delta \le 1/4$, define the parameters

$$c = \left(\frac{1}{2} + \delta, \frac{1}{2} - \delta\right), \quad \bar{c} = \left(\frac{1}{2} + \delta + \epsilon, \frac{1}{2} - \delta - \epsilon\right).$$

Notice that $\|c - \bar{c}\|_2 \asymp \epsilon$ and $|m_2(c) - m_2(\bar{c})| \asymp \epsilon|\epsilon - \delta|$. Thus, by applying Lemmas 2 and 4 under condition **(S)**, as well as Le Cam's Lemma, it suffices to show that there exists a choice of ϵ, δ such that

$$\frac{n}{\sqrt{d}}\epsilon|\epsilon - \delta| + \epsilon\sqrt{\frac{nm}{d}} = \epsilon\sqrt{\frac{n}{d}}\left(\sqrt{n}|\epsilon - \delta| + \sqrt{m}\right) \le 1/C,$$

for a sufficiently large constant $C > 0$. The above display is satisfied by choosing $\epsilon = c_0\sqrt{d/nm}$ and $\delta = \epsilon + (1/4) \wedge \sqrt{m/n}$, for a sufficiently small constant $c_0 > 0$. This proves the claim for $k = 2$.

Notice that the map $k \mapsto \mathcal{M}(n, d, k, m)$ is monotonically increasing. Thus, in view of the preceding lower bound for $k = 2$, it suffices to assume that $k \ge 30$ in what follows. For this regime, we will invoke Fano's Lemma (cf. [126]), a special case of which we recall next.

Lemma 7 (Fano's Lemma)**.** *Let $M \ge 2$ and let $c^{(1)}, \dots, c^{(M)} \in \Delta_k$ be a collection of parameters satisfying*

$$\epsilon := \min_{j \ne j'} \|c^{(j)} - c^{(j')}\|_2 > 0.$$

Let J be a random variable uniformly-distributed over $\{1, \dots, M\}$, and let $\widetilde{U}$ be a random variable such that $\widetilde{U} \,|\, J \sim \widetilde{\mathbf{Q}}_{c^{(J)}}$. Then, there exists a universal constant $C_1 > 0$ such that if $I(\widetilde{U}; J) \le C_1 \cdot \log M$, then

$$\mathcal{M}(n, d, k, m) \ge \epsilon/C_1.$$

In view of applying Fano's Lemma, let us begin by exhibiting an ϵ-packing of Δ_k. Let $1 \le s \le k/10 \le k - 2$ be an integer to be defined below. By the sparse Varshamov-Gilbert Lemma (cf. Theorem 27.6 of [126]), there exist an integer $M \ge 1$, a constant $C > 0$, and bitstrings $\omega^{(1)}, \dots, \omega^{(M)} \in \{0, 1\}^{k-2}$ satisfying the following three properties (where d_H denotes the Hamming distance):

(i) $d_H(\omega^{(j)}, \omega^{(j')}) \ge s/2$, for all $j \ne j'$.

(ii) $\log M \asymp s \log(k/s)$.

(iii) $\|\omega^{(j)}\|_0 = s$ for all $j = 1, \dots, M$.

Given a constant $\beta \in [0, 1/s]$ to be defined below, define for $j = 1, \dots, M$ the vectors

$$c_i^{(j)} = \beta\omega_i^{(j)}, \ i = 1, \dots, k-2, \quad \text{and} \quad c_{k-1}^{(j)} = c_k^{(j)} = (1 - \beta s)/2,$$

which lie in the set $\Sigma_{k,s}(\beta)$ in view of condition (iii). With this choice, notice that

$$\epsilon := \min_{j \ne j'} \|c^{(j)} - c^{(j')}\|_2 = \beta \cdot \min_{j \ne j'} \|\omega^{(j)} - \omega^{(j')}\|_2 \asymp \beta\sqrt{s},$$

by condition (i). Now, let $\widetilde{U}$ and J be defined as in Lemma 7. We will bound their mutual information separately in the case $m^{1+\gamma} \le d$ and $m \ge d$, beginning with the former. Define

a random variable U via $U \,|\, J \sim \mathbf{Q}_{c^{(J)}}^{\otimes d}$. We will bound the mutual information $I(J;\widetilde{U})$ by passing through $I(J;U)$:

$$\big|I(\widetilde{U};J) - I(U;J)\big| \le \big|H(J|U) - H(J|\widetilde{U})\big| = \big|\mathbb{E}[H(P_{J|U}) - H(P_{J|\widetilde{U}})]\big| \le (\log M)\cdot \mathrm{TV}(U,\widetilde{U}),$$

where $P_{J|U}$ is the conditional law of J given U, whose entropy is bounded above by $\log M$. Now, we simply have

$$\mathrm{TV}(U,\widetilde{U}) \le \frac{1}{M}\sum_{j=1}^{M} \mathrm{TV}(\mathbf{Q}_{c^{(j)}}^{\otimes d}, \widetilde{\mathbf{Q}}_{c^{(j)}}^{\otimes d}) \lesssim \sqrt{n/d} + \sqrt{mk/d} =: \delta,$$

by Lemma 2. Notice that δ vanishes under condition **(S)**. We now have

$$\begin{aligned} I(\widetilde{U};J) &\lesssim (\log M)\delta + I(U;J) \lesssim (\log M)\delta + \frac{1}{M^2}\sum_{j,j'=1}^{M} \mathrm{KL}(\mathbf{Q}_{c^{(j)}}^{\otimes d} \| \mathbf{Q}_{c^{(j')}}^{\otimes d}) \\ &\lesssim (\log M)\delta + \frac{nm}{dM^2}\sum_{j,j'=1}^{M} \|c^{(j)} - c^{(j')}\|_2^2, \end{aligned}$$

where the final inequality follows from Lemma 4(ii). On the other hand, when $m > d$, we may apply Lemma 5 to directly bound the mutual information by:

$$I(\widetilde{U};J) \le \frac{1}{M^2}\sum_{j,j'=1}^{M} \mathrm{KL}\big(\widetilde{\mathbf{Q}}_{c^{(j)}} \| \widetilde{\mathbf{Q}}_{c^{(j')}}\big) \lesssim \frac{n}{M^2}\sum_{j,j'}^{M} \|c^{(j)} - c^{(j')}\|_2^2.$$

It follows that for all m satisfying condition **(S)**, we have

$$I(\widetilde{U};J) \lesssim (\log M)\delta + \frac{nm_d}{dM^2}\sum_{j,j'}^{M} \|c^{(j)} - c^{(j')}\|_2^2 \lesssim (\log M)\delta + \frac{nm_d\beta^2 s}{d}.$$

With this bound in place, let us apply Fano's Lemma, for which we will use different choices of β, s depending on the magnitude of k. If $k < \sqrt{nm_d/d}$, we may choose a small enough constant $C > 0$ such that if $s = \lfloor k/10 \rfloor$ and $\beta = C\sqrt{d/nm_d}$, then $I(\widetilde{U};J) \le C_1 \log M$, where C_1 is the constant appearing in the statement of Lemma 7, and we used property (ii) above. Thus, when $k < \sqrt{nm_d/d}$, we obtain the minimax lower bound

$$\mathcal{M}(n,d,k,m) \gtrsim \beta\sqrt{s} = \sqrt{\frac{kd}{nm_d}}. \tag{D2}$$

Due to condition **(S)**, it remains to handle the case $k \ge (\sqrt{nm_d/d})^{1+\gamma}$. Pick $\beta = 1/s$, with s the smallest integer satisfying

$$s \ge c_0\sqrt{\frac{nm_d}{d\log(dk/(nm_d))}},$$

for a sufficiently small constant $c_0 > 0$ to be defined below. Then,

$$I(\widetilde{U};J) \lesssim \frac{nm_d}{ds} \asymp s\log(dk/(nm_d)) \asymp s\log(k/s).$$

Therefore, by property (ii) above, we have $I(\widetilde{U};J) \le C_1 \log M/2$ provided c_0 is chosen sufficiently small. Thus, by Fano's Lemma, one has

$$\mathcal{M}(n,d,k,m) \gtrsim \beta\sqrt{s} \asymp \sqrt{\frac{d\log(k)}{nm_d}},$$

where we used the fact that $\log(kd/(nm_d)) \asymp \log(k)$ under the stated assumption on k.

Altogether, we have thus shown that, under condition **(S)** and $nm > d \cdot \log k$, we have

$$\mathcal{M}(n,d,k,m) \gtrsim \min\left\{\left(\frac{d\log(k)}{nm_d}\right)^{\frac{1}{4}}, \left(\frac{dk}{nm_d}\right)^{\frac{1}{2}}\right\}.$$

This proves the claim. □

4. Proof of Lemma 4

Throughout the proof, let $\varpi = (\varpi_1, \ldots, \varpi_k)$ denote a random vector with entries

$$\varpi_i \overset{\text{i.i.d.}}{\sim} \mathcal{E}_d, \quad i = 1, \ldots, k.$$

Recall that we write

$$\mathbf{Q}_c = \mathbb{E}_\varpi\left[\mathrm{Poi}\left(n\langle c, \varpi\rangle\right) \otimes \bigotimes_{i=1}^{k} \mathrm{Poi}(m\varpi_i)\right],$$

and we denote the density of $\mathbf{Q}_c$, defined over $I := \mathbb{N}_0 \times \mathbb{N}_0^k$, as

$$\mathbf{q}_c(x,y) = \mathbb{E}_\varpi\left[f(x; n\langle \varpi, c\rangle)\prod_{i=1}^{k} f(y_i; m\varpi_i)\right], \quad (x,y) \in I,$$

where $f(\cdot;\lambda)$ is the $\mathrm{Poi}(\lambda)$ density. Fix once and for all the truncation parameter $t = (n+m)^{-\gamma_0} d^{\gamma_0 - 1}$ with $\gamma_0 = \gamma/(1+\gamma)$, and write

$$\mathbf{Q}_c^t = \mathbb{E}_\varpi\left[\mathrm{Poi}\left(n\langle c, \varpi^t\rangle\right) \otimes \bigotimes_{i=1}^{k} \mathrm{Poi}(m\varpi_i^t)\right], \quad \text{where } \varpi^t = \left(\varpi_1 \wedge t, \ldots, \varpi_k \wedge t\right),$$

with corresponding density denoted $\mathbf{q}_c^t(x,y)$. The following Lemma reduces our problem to that of bounding the χ^2-divergence between the truncated measures.

Lemma 8. *Let $1 \le k \le d$. Under assumption **(S)**, there exist constants $C_1, C_2, a > 0$ depending only on γ such that the following assertions hold for all $c, \bar{c} \in \Delta_k$.*

1. *We have,*

$$\mathrm{TV}\left(\mathbf{Q}_c^{\otimes d}, \mathbf{Q}_{\bar{c}}^{\otimes d}\right) \le C_1\left(\sqrt{d \cdot \chi^2(\mathbf{Q}_c^t, \mathbf{Q}_{\bar{c}}^t)} + e^{-C_2 d^a}\right).$$

2. *If $W(c,\bar{c}) = 0$, then*

$$\mathrm{KL}\left(\mathbf{Q}_c^{\otimes d}, \mathbf{Q}_{\bar{c}}^{\otimes d}\right) \le C_1\left(d \cdot \chi^2(\mathbf{Q}_c^t, \mathbf{Q}_{\bar{c}}^t) + e^{-C_2 d^a}\right).$$

The proof appears in Section G 2 c. Now, for the remainder of the proof, let $c, \bar{c} \in \Delta_k$ be any elements such that

$$m_j(c) = m_j(\bar{c}), \quad j = 1, \ldots, k-1.$$

Notice that $q_1 = 0$ when $W(c, \bar{c}) = 0$, thus, in view of Lemma 8, both assertions of the claimed Lemma 4 will follow if we are able to prove the following upper bound on the χ^2-divergence between *truncated* distributions:

$$\chi^2(\mathbf{Q}_c^t, \mathbf{Q}_{\bar{c}}^t) \lesssim q_1^2 \left(\frac{n}{d}\right)^k + \frac{nmq_2^2}{d^2}. \tag{D3}$$

The remainder of the proof is devoted to deriving equation (D3). Our approach consists of expanding the χ^2-divergence in terms of moment differences of the mixing measures of $\mathbf{Q}_c$ and $\mathbf{Q}_{\bar{c}}$. Expansions of this type have been used for Gaussian mixture models since the early work of [127, 128]; for Poisson mixture models, related ideas have been used for instance by [129, 130]. Although these methods often scale poorly for high-dimensional mixtures [131, 132], our earlier truncation step ensures that the relevant mixing measures have support near zero, and thus have moments which decay exponentially in k.

Define the following quantities:

$$\begin{aligned} \widetilde{U}_c &= n\langle c, \varpi^t \rangle, \quad \widetilde{V}_i = m\varpi_i^t, \quad \lambda = \mathbb{E}[\varpi_1^t], \\ U_c &= \widetilde{U}_c - n\lambda, \quad V_i = \widetilde{V}_i - m\lambda, \quad i = 1, \ldots, k. \end{aligned}$$

Notice that the random variables U_c and V_i all have mean zero. We will make use of the following elementary bounds on λ.

Lemma 9. *It holds that*

$$\frac{1}{d}\left(1 - tde^{-td}\right) \leq \lambda \leq \frac{1}{d}.$$

The proof appears in Section G 2 d. In particular, by definition of t and the fact that $(n+m)^{1+\gamma} \leq d$, we have $\lambda \asymp 1/d$. Now, notice that

$$\chi^2(\mathbf{Q}_{\bar{c}}^t, \mathbf{Q}_c^t) = \sum_{x, y_1, \ldots, y_k = 0}^{\infty} \frac{\left\{\mathbb{E}_\varpi\left[\left(f(x; \widetilde{U}_{\bar{c}}) - f(x; \widetilde{U}_c)\right) \prod_{i=1}^k f(y_i; \widetilde{V}_i)\right]\right\}^2}{\mathbb{E}_\varpi\left[f(x; \widetilde{U}_c) \prod_{i=1}^k f(y_i; \widetilde{V}_i)\right]}.$$

We lower bound the denominator using the following.

Lemma 10. *Assume condition **(S)**. Then, there exists a constant $C = C(\gamma) > 0$ such that for all $(x, y) \in \mathbb{N}_0$,*

$$\mathbb{E}_\varpi\left[f(x; \widetilde{U}_c) \prod_{i=1}^k f(y_i; \widetilde{V}_i)\right] \geq \frac{1}{C} f(x; n\lambda) \prod_{i=1}^k f(y_i; m\lambda).$$

The proof appears in Appendix G 2 e. We may thus write,

$$\begin{aligned} &\chi^2(\mathbf{Q}_{\bar{c}}^t, \mathbf{Q}_c^t) \\ &\leq C \sum_{x, y_1, \ldots, y_k = 0}^{\infty} \frac{\left\{\mathbb{E}_\varpi\left[\left(f(x; n\lambda + U_{\bar{c}}) - f(x; n\lambda + U_c)\right) \prod_{i=1}^k f(y_i; m\lambda + V_i)\right]\right\}^2}{f(x; n\lambda) \prod_{i=1}^k f(y_i; m\lambda)}. \end{aligned} \tag{D4}$$

We will proceed by expanding the numerator of the above display in the basis of Charlier polynomials, whose definition and basic properties are recalled in Appendix K 2 b. Let $\{\varphi_\ell(\cdot; \lambda)\}_{\ell=0}^\infty$ denote the univariate family of Charlier polynomials with parameter $\lambda > 0$. Given $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_k) \in \mathbb{R}_+^{k+1}$, we define the following tensor-product family of Charlier polynomials

$$\varphi_{\alpha,\beta}(x, y; \boldsymbol{\lambda}) = \varphi_\alpha(x; \lambda_0) \cdot \prod_{i=1}^k \varphi_{\beta_i}(y_i; \lambda_i), \quad x \in \mathbb{R}, y \in \mathbb{R}^k,$$

for any multi-indices $(\alpha, \beta) \in I$, where $I := \mathbb{N}_0 \times \mathbb{N}_0^k$. In what follows, we show that they form an orthogonal basis with respect to the $L^2(g_{\boldsymbol{\lambda}})$ inner product, where

$$g_{\boldsymbol{\lambda}}(x, y) = f(x; \lambda_0) \cdot \prod_{i=1}^{k} f(y_i; \lambda_i), \quad x \in \mathbb{N}_0, y \in \mathbb{N}_0^k.$$

Lemma 11. *The polynomial family* $\{\varphi_{\alpha,\beta}(\cdot, \cdot; \boldsymbol{\lambda})\}_{(\alpha,\beta)\in I}^{\infty}$ *with parameter* $\boldsymbol{\lambda} > 0$ *is an orthogonal basis of* $L^2(g_{\boldsymbol{\lambda}})$*, such that*

$$\mathbb{E}_{(X,Y)\sim g_{\boldsymbol{\lambda}}}\Big[\varphi_{\alpha,\beta}(X, Y; \boldsymbol{\lambda})\varphi_{\alpha',\beta'}(X, Y; \boldsymbol{\lambda})\Big] = \alpha!\lambda_0^{\alpha} \cdot \prod_{i=1}^{k} \beta_i!\lambda_i^{\beta_i} \cdot \mathbb{I}\big((\alpha, \beta) = (\alpha', \beta')\big),$$

for any $(\alpha, \beta), (\alpha', \beta') \in I$*. Furthermore, one has the relation*

$$g_{\boldsymbol{\lambda}+\mathbf{u}}(x, y) = g_{\boldsymbol{\lambda}}(x, y) \sum_{(\alpha,\beta)\in I} \varphi_{\alpha,\beta}(x, y; \boldsymbol{\lambda}) \frac{u_0^{\alpha}}{\alpha!\lambda_0^{\alpha}} \prod_{i=1}^{k} \frac{u_i^{\beta_i}}{\beta_i!\lambda_i^{\beta_i}}, \quad (x, y) \in I,$$

for any $\mathbf{u} = (u_0, \ldots, u_k) \in \mathbb{R}_+^{k+1}$ *such that* $\boldsymbol{\lambda} + \mathbf{u}$ *has positive entries.*

The proof appears in Appendix G 2 f. Now, fixing $\boldsymbol{\lambda} = (n\lambda, m\lambda, \ldots, m\lambda)$, we deduce from Lemma 11 that

$$\begin{aligned}\mathbb{E}_{\varpi}\left[\big(f(x; n\lambda + U_{\bar{c}}) - f(x; n\lambda + U_c)\big) \prod_{i=1}^{k} f(y_i; m\lambda + V_i)\right] \\ = g_{\boldsymbol{\lambda}}(x, y) \sum_{(\alpha,\beta)\in I} \varphi_{\alpha,\beta}(x, y; \boldsymbol{\lambda}) \frac{\Delta_{\alpha,\beta}}{\alpha!\beta!(n\lambda)^{\alpha}(m\lambda)^{|\beta|}},\end{aligned}$$

where for any multi-indices $(\alpha, \beta) \in I$, we write $|\beta| = \sum_i \beta_i$, $\beta! = \beta_1! \cdots \beta_k!$, and

$$\Delta_{\alpha,\beta} = \mathbb{E}_{\varpi}\Big[(U_{\bar{c}}^{\alpha} - U_c^{\alpha})V_1^{\beta_1} \cdots V_k^{\beta_k}\Big].$$

Thus, returning to equation (D4) , and using the orthogonality relation from Lemma 11, we arrive at

$$\begin{aligned}\chi^2(\mathbf{Q}_{\bar{c}}, \mathbf{Q}_c) &\le C \sum_{(x,y)\in I} \left(\sum_{(\alpha,\beta)\in I} \varphi_{\alpha,\beta}(x, y; \boldsymbol{\lambda}) \frac{\Delta_{\alpha,\beta}}{\alpha!\beta!(n\lambda)^{\alpha}(m\lambda)^{|\beta|}}\right)^2 g_{\boldsymbol{\lambda}}(x, y) \\ &= C \sum_{(\alpha,\beta)\in I} \frac{\Delta_{\alpha,\beta}^2}{\alpha!\beta!(n\lambda)^{\alpha}(m\lambda)^{|\beta|}} \\ &=: C(S_1 + S_2),\end{aligned}$$

where

$$S_1 = \sum_{\alpha=1}^{\infty} \frac{\Delta_{\alpha}^2}{\alpha!(n\lambda)^{\alpha}}, \quad S_2 = \sum_{\alpha=1}^{\infty} \sum_{\substack{\beta\in\mathbb{N}_0^k \\ |\beta|\ge 1}} \frac{\Delta_{\alpha,\beta}^2}{\alpha!\beta!(n\lambda)^{\alpha}(m\lambda)^{|\beta|}}, \tag{D5}$$

and where we abbreviate

$$\Delta_{\alpha} := \Delta_{\alpha,0} = \mathbb{E}_{\varpi}\big[U_{\bar{c}}^{\alpha} - U_c^{\alpha}\big].$$

We bound the terms S_1 and S_2 separately, beginning with the former. Recall that $\kappa_\alpha(X)$ denotes the α-th cumulant of a random variable X. By translational invariance of cumulants (except when $\alpha = 1$), it holds that:

$$\kappa_\alpha(U_c)/n^\alpha = \kappa_\alpha(\langle c, \varpi^t\rangle) - \lambda \cdot I(\alpha = 1),$$

for any $\alpha = 1, 2, \ldots$. Furthermore, since the elements of the vector ϖ are independent, we have

$$\kappa_\alpha(\langle c, \varpi^t\rangle) = \sum_{i=1}^k c_i^\alpha \kappa_\alpha(\varpi_i^t) = \kappa_\alpha(\varpi_1^t) \cdot m_\alpha(c),$$

where we recall that $m_\alpha(c)$ is the α-th moment of the uniform distribution over $\{c_1, \ldots, c_k\}$. Since we assumed that $m_\alpha(c) = m_\alpha(\bar{c})$ for all $1 \le \alpha \le k-1$, it must follow from the preceding two displays that $\kappa_\alpha(U_c) = \kappa_\alpha(U_{\bar{c}})$ for all such α, and we deduce that

$$S_1 \lesssim \sum_{\alpha=k}^\infty \frac{\Delta_\alpha^2}{\alpha!(n\lambda)^\alpha}.$$

Our aim is now to bound the remaining terms Δ_α, for $\alpha \ge k$. Notice that

$$\begin{aligned}
\Delta_\alpha/n^\alpha &= \mathbb{E}[(\langle \bar{c}, \varpi^t\rangle - \lambda)^\alpha] - \mathbb{E}[(\langle c, \varpi^t\rangle - \lambda)^\alpha] \\
&= \sum_{j=k}^\alpha \binom{\alpha}{j} \lambda^{\alpha-j} \mathbb{E}\left[\langle \bar{c}, \varpi^t\rangle^j - \langle c, \varpi^t\rangle^j\right] \\
&= \sum_{j=k}^\alpha \binom{\alpha}{j} \lambda^{\alpha-j} (\bar{\eta}_j - \eta_j), \qquad \text{(D6)}
\end{aligned}$$

where we define the quantities

$$\begin{aligned}
\xi_j &= \kappa_j(\langle \varpi^t, c\rangle), \quad \eta_j = \mathbb{E}[\langle \varpi^t, c\rangle^j], \\
\bar{\xi}_j &= \kappa_j(\langle \varpi^t, \bar{c}\rangle), \quad \bar{\eta}_j = \mathbb{E}[\langle \varpi^t, \bar{c}\rangle^j], \quad j = 1, 2, \ldots.
\end{aligned}$$

By equation (K11) of Appendix K 2 c, the moments η_j can be expressed in terms of the cumulants ξ_j via the following expansion in the Bell polynomial system:

$$\eta_j = \sum_{\ell=1}^j B_{j,\ell}(\xi_1, \ldots, \xi_{j-\ell+1}) = \sum_{\ell=1}^j j! \sum_{(h_1,\ldots,h_{j-\ell+1}) \in \mathcal{H}_{j,\ell}} \prod_{i=1}^{j-\ell+1} \frac{\xi_i^{h_i}}{(i!)^{h_i} h_i!},$$

thus, for any $j = k, \ldots, \alpha$, we have

$$\bar{\eta}_j - \eta_j = \sum_{\ell=1}^j j! \sum_{(h_1,\ldots,h_{j-\ell+1}) \in \mathcal{H}_{j,\ell}} \left(\prod_{i=1}^{j-\ell+1} \frac{\xi_i^{h_i}}{(i!)^{h_i} h_i!} - \prod_{i=1}^{j-\ell+1} \frac{\bar{\xi}_i^{h_i}}{(i!)^{h_i} h_i!} \right).$$

Let $K = K(j, \alpha) > 0$ denote a constant depending on j, α, whose value may change from line to line. We have

$$|\bar{\eta}_j - \eta_j| \le K \left| \sum_{\ell=1}^j \sum_{(h_1,\ldots,h_{j-\ell+1}) \in \mathcal{H}_{j,\ell}} \left(\prod_{i=1}^{j-\ell+1} \xi_i^{h_i} - \prod_{i=1}^{j-\ell+1} \bar{\xi}_i^{h_i} \right) \right|.$$

Now, recall that $m_i(c) = m_i(\bar{c})$, and hence $\xi_i = \bar{\xi}_i$, for all $i = 1, \dots, k-1$. Thus

$$|\bar{\eta}_j - \eta_j| \le K \left| \sum_{\ell=1}^{j} \sum_{(h_1,\dots,h_{j-\ell+1}) \in \mathcal{H}_{j,\ell}} \left(\prod_{i \le k-1} \xi_i^{h_i} \right) \left(\prod_{i \ge k} \xi_i^{h_i} - \prod_{i \ge k} \bar{\xi}_i^{h_i} \right) \right|,$$

where all products are to be understood as ranging over all integers $1 \le i \le j - \ell + 1$ satisfying the stated conditions, with the convention that empty products equal 1. Now recalling that $\xi_i = \sum_{r=1}^{k} c_r^i \kappa_i(\varpi_1 \wedge t)$ for all i, due to the independence of the entries of ϖ^t, we have (cf. Appendix K 2 c):

$$\begin{aligned}
\kappa_i(\varpi_1 \wedge t) &\le \sum_{\ell=1}^{i} (\ell-1)! B_{i,\ell}(\mathbb{E}[\varpi_1 \wedge t], \dots, \mathbb{E}[(\varpi_1 \wedge t)^{i-\ell+1}]) \\
&\le \sum_{\ell=1}^{i} (\ell-1)! B_{i,\ell}\left(\frac{1}{d}, \dots, \frac{(i-\ell+1)!}{d^{i-\ell+1}}\right) \\
&\le K \sum_{\ell=1}^{i} B_{i,\ell}\left(\frac{1}{d}, \dots, \frac{1}{d^{i-\ell+1}}\right) \\
&\le K d^{-i}.
\end{aligned}$$

Thus, $\xi_i \le K d^{-i}$, and similarly, $\bar{\xi}_i \le K d^{-i}$. Using the definition of $\mathcal{H}_{j,\ell}$, we have $\sum_{i=1}^{j-\ell+1} i h_i = j$, thus

$$\begin{aligned}
|\bar{\eta}_j - \eta_j| &\le K d^{-j} \sum_{\ell=1}^{j} \sum_{(h_1,\dots,h_{j-\ell+1}) \in \mathcal{H}_{j,\ell}} \left| \prod_{i \ge k} (d^i \xi_i)^{h_i} - \prod_{i \ge k} (d^i \bar{\xi}_i)^{h_i} \right| \\
&\le K d^{-j} \sum_{\ell=1}^{j} \sum_{(h_1,\dots,h_{j-\ell+1}) \in \mathcal{H}_{j,\ell}} \sum_{i \ge k} \left| (d^i \xi_i)^{h_i} - (d^i \bar{\xi}_i)^{h_i} \right| \\
&\le K d^{-j} \sum_{\ell=1}^{j} \sum_{(h_1,\dots,h_{j-\ell+1}) \in \mathcal{H}_{j,\ell}} \sum_{i \ge k} d^i \left| \xi_i - \bar{\xi}_i \right| \\
&\le K d^{-j} \sum_{i=k}^{j} d^i \left| \xi_i - \bar{\xi}_i \right| \\
&= K d^{-j} \sum_{i=k}^{j} d^i \left| \kappa_i(\varpi_1 \wedge t) \sum_{r=1}^{k} (c_r^i - \bar{c}_r^i) \right| \le K d^{-j} q_1.
\end{aligned}$$

Returning to equation (D6), we deduce that for any $\alpha \ge k$, there exists a constant $K_\alpha > 0$ such that

$$|\Delta_\alpha| \le K_\alpha n^\alpha \sum_{j=k}^{\alpha} \binom{\alpha}{j} \lambda^{\alpha-j} d^{-j} q_1 \le K_\alpha q_1 \left(\frac{n}{d}\right)^\alpha.$$

On the other hand, we also have the naive bound

$$|\Delta_\alpha| \le n^\alpha \sum_{j=k}^{\alpha} \binom{\alpha}{j} \lambda^{\alpha-j} t^j \le n^\alpha (t+\lambda)^\alpha \le (2nt)^\alpha.$$

Thus, returning to equation (D5), we have shown that for all $p \geq k$, there exists $K_p > 0$ such that

$$\begin{aligned} S_1 &\leq K_p q_1^2 \sum_{\alpha=k}^{p} \frac{(n/d)^{2\alpha}}{\alpha! \lambda^\alpha} + \sum_{\alpha=p+1}^{\infty} \frac{(2nt)^{2\alpha}}{\alpha! \lambda^\alpha} \\ &\leq K_p q_1^2 \sum_{\alpha=k}^{p} \frac{(n/d)^{\alpha}}{\alpha!} + \sum_{\alpha=p+1}^{\infty} \frac{(2n/d)^{\alpha/2}}{\alpha!} \\ &\lesssim K_p q_1^2 (2n/d)^k + (2n/d)^{p/2}. \end{aligned}$$

Under our conditions on d and n, we can choose p sufficiently large, as a function only of γ, such that $(2n/d)^{p/2} \leq 1/d^5$, to obtain

$$S_1 \leq C q_1^2 (n/d)^k + d^{-5}, \tag{D7}$$

for a constant $C = C(k, \gamma) > 0$. We now turn to bounding the quantity

$$S_2 = \sum_{\alpha=1}^{\infty} \sum_{\substack{\beta \in \mathbb{N}_0^k \\ |\beta| \geq 1}} \frac{\Delta_{\alpha,\beta}^2}{\alpha! \beta! (n\lambda)^\alpha (m\lambda)^\beta}.$$

For any $\alpha, |\beta| \geq 1$, we have

$$|\Delta_{\alpha,\beta}| \leq n^\alpha m^{|\beta|} \mathbb{E} \left| \Big((\langle \varpi^t, \bar{c} \rangle - \lambda)^\alpha - (\langle \varpi^t, c \rangle - \lambda)^\alpha \Big) \prod_{i=1}^{k} (\varpi_i \wedge t - \lambda)^{\beta_i} \right|.$$

By the mean value theorem, one has

$$\big| (\langle \varpi^t, \bar{c} \rangle - \lambda)^\alpha - (\langle \varpi^t, c \rangle - \lambda)^\alpha \big| \leq \alpha \big(|\langle \varpi^t, c \rangle - \lambda| + |\langle \varpi^t, \bar{c} \rangle - \lambda| \big)^{\alpha - 1} |\langle \bar{c} - c, \varpi^t \rangle|,$$

thus, $|\Delta_{\alpha,\beta}| \leq n^\alpha m^{|\beta|} \alpha T_1^{1/2} (T_2 T_3)^{1/4}$, where

$$\begin{aligned} T_1 &= \mathbb{E} |\langle \bar{c} - c, \varpi^t \rangle|^2 \\ T_2 &= \mathbb{E} \big(|\langle \varpi^t, c \rangle - \lambda| + |\langle \varpi^t, \bar{c} \rangle - \lambda| \big)^{4(\alpha-1)} \\ T_3 &= \mathbb{E} \prod_{i=1}^{k} |\varpi_i \wedge t - \lambda|^{4\beta_i}. \end{aligned}$$

We have,

$$T_1 = \operatorname{Var}[\langle \bar{c} - c, \varpi^t \rangle] = \sum_{i=1}^{k} (\bar{c}_i - c_i)^2 \operatorname{Var}[\varpi^t] \lesssim q_2^2 / d^2.$$

To bound T_2, apply Jensen's inequality to obtain

$$\mathbb{E} |\langle \varpi^t, \bar{c} \rangle - \lambda|^{4(\alpha-1)} = \mathbb{E} \left| \sum_{i=1}^{k} \bar{c}_i (\varpi_i^t - \lambda) \right|^{4(\alpha-1)} \leq \sum_{i=1}^{k} \bar{c}_i \mathbb{E} |\varpi_i^t - \lambda|^{4(\alpha-1)} \leq \frac{(4\alpha)!}{d^{4(\alpha-1)}} \wedge t^{4(\alpha-1)}.$$

It follows that

$$T_2 \lesssim 2^{4(\alpha-1)} \left(\frac{(4\alpha)!}{d^{4(\alpha-1)}} \wedge t^{4(\alpha-1)} \right).$$

Next, we have,

$$T_3 = \prod_{i=1}^{k} \mathbb{E}|\varpi_i \wedge t - \lambda|^{4\beta_i} \lesssim \prod_{i=1}^{k} 2^{4\beta_i} \left(\frac{(4\beta_i)!}{d^{4\beta_i}} \wedge t^{4\beta_i} \right) \leq 2^{4|\beta|} \left(\frac{(4\beta)!}{d^{4|\beta|}} \wedge t^{4|\beta|} \right)$$

We thus obtain

$$\begin{aligned}
\frac{|\Delta_{\alpha,\beta}|}{n^\alpha m^{|\beta|}} &\leq \alpha T_1^{1/2} (T_2 T_3)^{1/4} \\
&\leq \alpha \cdot \frac{q_2}{d} \cdot 2^{\alpha-1} \left(\frac{(4\alpha)!}{d^{(\alpha-1)}} \wedge t^{(\alpha-1)} \right) \cdot 2^{|\beta|} \left(\frac{(4\beta)!}{d^{|\beta|}} \wedge t^{|\beta|} \right) \\
&\leq 2^{\alpha+|\beta|} q_2 \left(\frac{(4\alpha)!}{d^\alpha} \wedge \frac{t^{(\alpha-1)}}{d} \right) \left(\frac{(4\beta)!}{d^{|\beta|}} \wedge t^{|\beta|} \right) \\
&\leq 2^{\alpha+|\beta|} q_2 \left(\frac{(4\alpha)!}{d^\alpha} \wedge t^\alpha \right) \left(\frac{(4\beta)!}{d^{|\beta|}} \wedge t^{|\beta|} \right) \\
&\leq 2^{\alpha+|\beta|} q_2 \left(\frac{(4\alpha)!(4\beta)!}{d^{\alpha+|\beta|}} \wedge t^{\alpha+|\beta|} \right).
\end{aligned}$$

It follows that for any fixed $\ell \geq 1$, there exists a constant $C_\ell > 0$ (which potentially grows factorially in ℓ) such that:

$$\begin{aligned}
S_2 &= \sum_{\substack{\alpha \in \mathbb{N}_0, \beta \in \mathbb{N}_0^k \\ |\beta| \geq 1, \alpha+|\beta| < \ell}} \frac{\Delta_{\alpha,\beta}^2}{\alpha!\beta!(n\lambda)^\alpha (m\lambda)^{|\beta|}} + \sum_{\substack{\alpha \in \mathbb{N}_0, \beta \in \mathbb{N}_0^k \\ |\beta| \geq 1, \alpha+|\beta| \geq \ell}} \frac{\Delta_{\alpha,\beta}^2}{\alpha!\beta!(n\lambda)^\alpha (m\lambda)^{|\beta|}} \\
&\lesssim C_\ell \sum_{\substack{\alpha \in \mathbb{N}_0, \beta \in \mathbb{N}_0^k \\ |\beta| \geq 1, \alpha+|\beta| < \ell}} \frac{(q_2 n^\alpha m^{|\beta|}/d^{\alpha+|\beta|})^2}{(n\lambda)^\alpha (m\lambda)^{|\beta|}} + \sum_{\substack{\alpha \in \mathbb{N}_0, \beta \in \mathbb{N}_0^k \\ |\beta| \geq 1, \alpha+|\beta| \geq \ell}} \frac{(q_2 n^\alpha m^{|\beta|} (2t)^{\alpha+|\beta|})^2}{\alpha!\beta!(n\lambda)^\alpha (m\lambda)^{|\beta|}} \\
&\lesssim q_2^2 \left\{ C_\ell \frac{nm}{d^2} + \sum_{\substack{\alpha \in \mathbb{N}_0, \beta \in \mathbb{N}_0^k \\ |\beta| \geq 1, \alpha+|\beta| \geq \ell}} \frac{(4t^2 d)^{(\alpha+|\beta|)} n^\alpha m^{|\beta|}}{\alpha!\beta!} \right\} \\
&\lesssim q_2^2 \left\{ C_\ell \frac{nm}{d^2} + \sum_{\alpha \geq \ell} \frac{(4t^2 dn)^\alpha}{\alpha!} \cdot \left(\sum_{b \geq \ell} \frac{(4t^2 dm)^b}{b!} \right)^k \right\} \\
&\leq q_2^2 \cdot C_\ell \left\{ \frac{nm}{d^2} + (4t^2 dn)^\ell e^{4t^2 dn} \cdot \left((4t^2 dm)^\ell e^{4t^2 dm} \right)^k \right\}.
\end{aligned}$$

Notice that $t^2 d(n+m) = \sqrt{(n+m)/d} = o(1)$, and we have $e^{4t^2 dm} \vee e^{4t^2 dn} \leq C < \infty$, thus we obtain

$$S_2 \lesssim q_2^2 \cdot C_\ell \left\{ \frac{nm}{d^2} + \left(\frac{4(n+m)}{d} \right)^{\ell/2} \cdot \left(C \left(\frac{4(n+m)}{d} \right)^{\ell/2} \right)^k \right\}.$$

By choosing $\ell = 5$, the second term in the above display is of lower order than the first for large enough d (irrespective of the magnitude of k), and we obtain $S_2 \lesssim q_2^2 nm/d$. Combining this bound with equation (D7) and the fact that $q_2 \geq 1/d$, we have thus shown that

$$\chi^2(\mathbf{Q}_{\bar{c}}^t, \mathbf{Q}_c^t) \leq C_1 q_1^2 \left(\frac{n}{d} \right)^k + C_2 \frac{q_2^2 nm}{d^2},$$

where $C_1 = C_1(\gamma, k)$ and $C_2 = C_2(\gamma)$. This proves equation (D3), and the claim follows. $\square$

5. Proof of Lemma 5

Let $c, \bar{c} \in \Delta_k$. Let $\Pi \in \mathbb{R}^{k\times d}$ be a random matrix whose rows are independently drawn from the flat Dirichlet law $\mathcal{D}_d$. Let $Y^c \in \mathbb{R}^d$ and $V \in \mathbb{R}^{k\times d}$ be drawn conditionally independently according to

$$Y^c|\Pi \sim \mathrm{Mult}(n; \Pi^\top c), \quad V_{i\cdot}|\Pi \sim \bigotimes_{i=1}^{k} \mathrm{Mult}(m; \Pi_{i\cdot}), \quad i = 1, \dots, k,$$

so that $(Y^c, V) \sim \widetilde{\mathbf{Q}}_c$. Given random variables X, X' which are absolutely continuous with respect to a common dominating measure, we denote by $p_{X,X'}$ their joint density, and by p_X the marginal density of X. We also denote the Markov kernel of X conditionally on X' by $p_{X|X'}$, whenever it exists.

By the chain rule for the KL divergence, one has

$$\mathrm{KL}(\widetilde{\mathbf{Q}}_c \| \widetilde{\mathbf{Q}}_{\bar{c}}) = \mathrm{KL}(f_{Y^c,V} \| f_{Y^{\bar{c}},V}) = \mathbb{E}_V\Big[\mathrm{KL}\big(f_{Y^c|V} \| f_{Y^{\bar{c}}|V}\big)\Big].$$

Notice that the conditional law of Y^c given V is

$$\begin{aligned} f_{Y^c|V}(y|v) &= \frac{1}{f_V(v)} \int_{\mathbb{R}_+^k} f_{Y^c,V|\Pi}(y, v|\pi) d\mathcal{D}_d^{\otimes k}(\pi) \\ &= \frac{1}{f_V(v)} \int_{\mathbb{R}_+^k} f_{Y^c|\Pi}(y|\pi) \cdot f_{V|\Pi}(v|\pi) d\mathcal{D}_d^{\otimes k}(\pi) \\ &= \int_{\mathbb{R}_+^k} f_{Y^c|\Pi}(y|\pi) \cdot f_{\Pi|V}(\pi|v) d\mathcal{D}_d^{\otimes k}(\pi), \end{aligned} \tag{D8}$$

where we used Bayes' rule and the fact that the law of the rows of Π are uniformly-distributed over the d-simplex. By conjugacy of the multinomial and Dirichlet distributions, notice that the conditional law of Π given V is given by

$$\Pi_{i,\cdot}|V \sim \mathrm{Dirichlet}(1 + V_{i1}, \dots, 1 + V_{id}), \quad i = 1, \dots, k.$$

Now, equation (D8) shows that the conditional law of Y^c given V is simply given by the law $\mathbb{E}_\Pi[\mathrm{Poi}(n\Pi^\top c) \,|\, V]$, which we use as a shorthand to denote the posterior distribution of $\mathrm{Poi}(n\Lambda^\top c)$ when Λ is drawn conditionally on V from the Dirichlet law in the above display. We thus have

$$\begin{aligned} \mathrm{KL}(\widetilde{\mathbf{Q}}_c \| \widetilde{\mathbf{Q}}_{\bar{c}}) &= \mathbb{E}_V\Big\{\mathrm{KL}\Big(\mathbb{E}_\Pi[\mathrm{Poi}(n\Pi^\top c) \,|\, V] \,\|\, \mathbb{E}_\Pi[\mathrm{Poi}(n\Pi^\top \bar{c}) \,\|\, V]\Big)\Big\} \\ &\le \mathbb{E}_V\Big\{\mathbb{E}_\Pi\big[\mathrm{KL}\big(\mathrm{Poi}(n\Pi^\top c) \,\|\, \mathrm{Poi}(n\Pi^\top \bar{c})\big) \,|\, V\big]\Big\} \\ &= \mathbb{E}_\Pi\Big\{\mathrm{KL}\big(\mathrm{Poi}(n\Pi^\top c) \,\|\, \mathrm{Poi}(n\Pi^\top \bar{c})\big)\Big\}, \end{aligned}$$

where we used the convexity of the KL divergence in the second line. Notice that

$$\begin{aligned} \mathbb{E}_\Pi\Big\{\mathrm{KL}\big(f_{U_V^c|\Pi} \| f_{U_V^{\bar{c}}|\Pi}\big)\Big\} &\lesssim n \sum_{j=1}^d \mathbb{E}_\Pi\left[\frac{(\Pi^\top(c - \bar{c}))_j^2}{(\Pi^\top \bar{c})_j}\right] \\ &\le n \sum_{j=1}^d \Big(\mathbb{E}_\Pi\left[(\Pi^\top(c - \bar{c}))_j^6\right]\Big)^{\frac{1}{3}} \Big(\mathbb{E}\Big[(\Pi^\top \bar{c})_j^{-3/2}\Big]\Big)^{\frac{2}{3}}. \end{aligned}$$

From here, we prove claims (i) and (ii) separately. To prove claim (i), assume $c, \bar{c} \in \Sigma_{k,s}(\beta)$. Then, there exists two index sets $S_1, S_2 \subseteq \{1, \dots, k-2\}$ of cardinality s such that $c_i = \beta \cdot I(i \in S_1)$ and $\bar{c}_i = \beta \cdot I(i \in S_2)$ for all $i = 1, \dots, k-2$, and $c_{k-1} = \bar{c}_{k-1} = c_k = \bar{c}_k = (1-\beta s)/2$. It follows that

$$\|c - \bar{c}\|_2^2 = \beta^2 \cdot \big(|S_1 \setminus S_2| + |S_2 \setminus S_1|\big) = 2\beta^2 |S_1 \setminus S_2|.$$

Now, notice that $(\Pi^\top \bar{c})_j = \beta \sum_{i \in S_2} \pi_{ij} + \frac{1-s\beta}{2}(\pi_{kj} + \pi_{(k-1)j})$, where we recall that $\pi_{ij} \overset{iid}{\sim} \mathrm{Beta}(1, d-1)$ for all i, j, independently across i. We will make use of the following.

Lemma 12. *Let $L \geq 2$ and let $X_1, \dots, X_L \sim \mathrm{Beta}(1, d-1)$ be independent random variables. Then, there exists a universal constant $C > 0$ such that*

$$\mathbb{E}\Big[\big(\textstyle\sum_{i=1}^L X_i\big)^{-3/2}\Big] \leq C(d/L)^{3/2}.$$

The proof appears in Appendix G 2 g. It follows that

$$\begin{aligned}
\mathbb{E}\left[(\Pi^\top \bar{c})_j^{-3/2}\right] &\leq \min\left\{\mathbb{E}\left[(\beta \textstyle\sum_{i \in S_2} \pi_{ij})_j^{-3/2}\right], \mathbb{E}\left[(1-s\beta)(\pi_{kj} + \pi_{(k-1)j})/2\right]\right\} \\
&\leq C d^{3/2} \cdot \min\left\{(\beta s)^{-3/2}, (1-s\beta)/2^{5/2}\right\} \lesssim d^{3/2}.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&\mathbb{E}_\Pi\left[(\Pi^\top(c - \bar{c}))_j^6\right] \\
&= \beta^6 \cdot \mathbb{E}\left[\left(\sum_{i \in S_1 \setminus S_2} \pi_{ij} - \sum_{i \in S_2 \setminus S_1} \pi_{ij}\right)^6\right] \\
&= \beta^6 \cdot \mathbb{E}\left[\left(\sum_{i \in S_1 \setminus S_2} (\pi_{ij} - \frac{1}{d}) - \sum_{i \in S_2 \setminus S_1} (\pi_{ij} - \frac{1}{d})\right)^6\right] \\
&\lesssim \beta^6 \cdot \mathbb{E}\left[\left(\sum_{i \in S_1 \setminus S_2} (\pi_{ij} - \frac{1}{d})\right)^6\right] + \beta^6 \cdot \mathbb{E}\left[\left(\sum_{i \in S_2 \setminus S_1} (\pi_{ij} - \frac{1}{d})\right)^6\right].
\end{aligned}$$

Recall that $\mathbb{E}[\pi_{ij}] = 1/d$, and that the π_{ij} are i.i.d. across i, for any fixed j. Now, apply Rosenthal's inequalities [133, 134] to obtain that for any $j = 1, \dots, d$,

$$\mathbb{E}\left[\left(\sum_{i \in S_1 \setminus S_2} (\pi_{ij} - \frac{1}{d})\right)^6\right] \lesssim |S_1 \setminus S_2|^3 \cdot \mathrm{Var}^3[\pi_{11}] + |S_1 \setminus S_2| \cdot \mathbb{E}|\pi_{11}|^6 \lesssim \frac{|S_1 \setminus S_2|^3}{d^6}.$$

After repeating a symmetric argument, we thus obtain

$$\mathbb{E}_\Pi\left[(\Pi^\top(c - \bar{c}))_j^6\right] \lesssim \beta^6 \frac{|S_1 \setminus S_2|^3 + |S_2 \setminus S_1|^3}{d^6} \asymp \|c - \bar{c}\|_2^6 / d^6.$$

Altogether, we have thus shown

$$\begin{aligned}
\mathbb{E}_\Pi\Big\{\mathrm{KL}\big(f_{U_V^c|\Pi} \| f_{U_V^c|\Pi}\big)\Big\} &\leq n \sum_{j=1}^d \Big(\mathbb{E}_\Pi\left[(\Pi^\top(c - \bar{c}))_j^6\right]\Big)^{\frac{1}{3}} \Big(\mathbb{E}\Big[(\Pi^\top \bar{c})_j^{-3/2}\Big]\Big)^{\frac{2}{3}} \\
&\lesssim nd \cdot (\|c - \bar{c}\|_2^6 / d^6)^{-\frac{1}{3}} \cdot d \lesssim n\|c - \bar{c}\|_2^2.
\end{aligned}$$

and the first claim follows. To prove the second claim, notice that when $c, \bar{c} \in \Delta_k$ satisfy $\|c\|_\infty \vee \|\bar{c}\|_\infty \leq 3/4$, at least two entries of $\bar{c}$ are bounded from below by $3/(4k)$. Assuming without loss of generality that these correspond to the first two entries of $\bar{c}$, one has

$$\mathbb{E}\Big[(\Pi^\top \bar{c})_j^{-3/2}\Big] \gtrsim \mathbb{E}[(\pi_{11} + \pi_{21})^{-3/2}] \gtrsim d^{3/2},$$

where the symbol '$\lesssim$' now hides constants depending on k, and where we applied Lemma 12. Furthermore, one has

$$\mathbb{E}\big[(\Pi^\top (c - \bar{c})_j^6\big] \lesssim \mathbb{E}\big[\|\Pi_{\cdot j}\|_2^6\big] \cdot \|c - \bar{c}\|_2^6 \lesssim d^{-6}\|c - \bar{c}\|_2^2.$$

From here, the claim can be deduced as before. □

Appendix E: Proofs of Upper Bounds

The goal of this section is to prove Propositions 3 and 4.

1. Proof of Proposition 3

We prove Proposition 3 in two steps. We begin by showing that the unregularized collision estimator

$$\widehat{c}_i^{\,\mathrm{coll}} = \frac{d+1}{nm} \sum_{\ell=1}^{n} \sum_{r=1}^{m} I(Z_\ell = W_{ir}) - 1, \quad i = 1, \dots, k,$$

already achieves the optimal convergence rate when $k < \sqrt{nm_d/d}$.

Lemma 13. *Under the multinomial model, there exists a universal constant $C > 0$ such that for all $1 \leq n \leq d$ and all $m, k \geq 1$,*

$$\sup_{c \in \Delta_k} \mathbb{E}\|\widehat{c} - c\|_2 \leq C\sqrt{\frac{dk}{nm_d}},$$

with $m_d = \min\{m, d\}$.

Second, we will prove that for appropriate $\lambda > 0$, the hard-thresholded collision estimator achieves the optimal convergence rate when $k > \sqrt{nm_d/d}$. We will analyze this estimator under the normalized Poisson model, which will allow us to deduce the following upper bound.

Lemma 14. *Let $n, k, d, m \geq 1$ and $m_d = \min\{m, d\}$. Assume that $k \leq d$ and $nm_d > d^{1+\gamma}$ for some $\gamma > 0$. Then, there exists a constant $C = C(\gamma) > 0$ such that*

$$\mathcal{M}(n, d, k, m) \leq C\left(\frac{d \log k}{nm_d}\right)^{\frac{1}{4}},$$

with $m_d = \min\{m, d\}$.

Let us now prove these Lemmas in turn.

a. Proof of Lemma 13

Throughout the proof, we repeatedly make use of elementary moment bounds for Dirichlet random variables, as summarized in Lemma 26. Abbreviate $\widehat{c}^{\text{coll}}$ by $\widehat{c}$. It suffices to show that for each i, $\mathbb{E}|\widehat{c}_i - c_i| \lesssim \sqrt{d/nm_d}$. Without loss of generality, fix $i = 1$. Recall that the collections of random variables $\{Z_\ell\}_\ell$ and $\{W_{1r}\}_r$ are each exchangeable, and mutually independent conditionally on Π, thus,

$$\mathbb{E}[\widehat{c}_1] = (d+1) \cdot \mathbb{P}(Z_1 = W_{11}) - 1.$$

Furthermore,

$$\begin{aligned}
\mathbb{P}(Z_1 = W_{11}) &= \sum_{j=1}^{d} \mathbb{E}\big[\mathbb{P}(Z_1 = z_j, W_{11} = z_j|\Pi)\big] \\
&= \sum_{j=1}^{d} \mathbb{E}\big[\mathbb{P}(Z_1 = z_j|\Pi) \cdot \mathbb{P}(W_{11} = z_j|\Pi)\big] \\
&= \sum_{j=1}^{d} \mathbb{E}\left[\sum_{i=1}^{k} c_i \pi_{ij} \pi_{1j}\right] \\
&= \sum_{j=1}^{d} \left[c_1 \frac{2}{d(d+1)} + \sum_{i=2}^{k} c_i \frac{1}{d(d+1)}\right] \\
&= \frac{1}{d(d+1)} \sum_{j=1}^{d} [1 + c_1] = \frac{1+c_1}{d+1}.
\end{aligned}$$

Thus, $\widehat{c}_1$ is unbiased. To bound its variance, notice that

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{\ell=1}^{n}\sum_{r=1}^{m} I(Z_\ell = W_{1r})\right)^2\right] \\
&= \sum_{\ell,\ell'=1}^{n}\sum_{r,r'=1}^{m} \mathbb{E}\big[I(Z_\ell = W_{1r}) I(Z_{\ell'} = W_{1r'})\big] \\
&= \sum_{\ell=1}^{n}\sum_{r=1}^{m} \mathbb{E}\big[I(Z_\ell = Z_{1r})\big] + \sum_{\ell=1}^{n}\sum_{r\neq r'} \mathbb{E}\big[I(Z_\ell = W_{1r} = W_{1r'})\big] \\
&\quad + \sum_{\ell\neq\ell'}\sum_{r=1}^{m} \mathbb{E}\big[I(Z_\ell = Z_{\ell'} = Z_{1r})\big] + \sum_{\ell\neq\ell'}\sum_{r\neq r'} \mathbb{E}\big[I(Z_\ell = W_{1r}) I(Z_{\ell'} = W_{1r'})\big] \\
&=: (I) + (II) + (III) + (IV).
\end{aligned}$$

We compute these terms in turn. First, our earlier bias calculations imply that

$$(I) = nm \frac{1+c_1}{d+1} \lesssim \frac{nm}{d}.$$

To compute term (II), notice that

$$\mathbb{P}(W_1 = Z_{11} = Z_{12}) = \sum_{j=1}^{d} \mathbb{E}\big[\mathbb{P}(W_1 = z_j|\Pi)\mathbb{P}(Z_{11} = z_j|\Pi)\mathbb{P}(Z_{12} = z_j|\Pi)\big]$$

$$= \sum_{j=1}^{d} \sum_{i=1}^{k} c_i \mathbb{E}\big[\pi_{ij}\pi_{1j}^2\big] = \sum_{j=1}^{d} \left\{ c_1 \mathbb{E}\big[\pi_{1j}^3\big] + \sum_{i=2}^{k} c_i \mathbb{E}\big[\pi_{ij}\pi_{1j}^2\big] \right\} \lesssim 1/d^2,$$

thus,

$$(II) \lesssim \frac{nm(m-1)}{d^2} \lesssim \frac{nm^2}{d^2}.$$

A similar argument shows that $(III) \lesssim mn^2/d^2$. To bound (IV), notice that

$$\begin{aligned}
&\mathbb{E}\big[I(W_1 = Z_{11})I(W_2 = Z_{12})\big] \\
&= \mathbb{E}\Big[\mathbb{P}(W_1 = Z_{11}|\Pi) \cdot \mathbb{P}(W_2 = Z_{12}|\Pi)\Big] \\
&= \sum_{j,j'=1}^{d} \sum_{i,i'} c_i c_{i'} \mathbb{E}[\pi_{ij}\pi_{1j}\pi_{i'j'}\pi_{1j'}] = (a) + (b),
\end{aligned}$$

where

$$(a) = \sum_{j=1}^{d} \sum_{i,i'} c_i c_{i'} \mathbb{E}[\pi_{ij}\pi_{1j}\pi_{i'j}\pi_{1j}] \leq C'_k d^{-3},$$

for a sufficiently large constant $C' > 0$, and

$$\begin{aligned}
(b) &= \sum_{j \neq j'} \sum_{i,i'} c_i c_{i'} \mathbb{E}[\pi_{ij}\pi_{1j}\pi_{i'j'}\pi_{1j'}] \\
&= \sum_{j \neq j'} \left(\sum_{i=1}^{k} c_i \mathbb{E}[\pi_{ij}\pi_{1j}] \right)^2 + \sum_{j \neq j'} \sum_{i,i'} c_i c_{i'} \mathrm{Cov}(\pi_{ij}\pi_{1j}, \pi_{i'j'}\pi_{1j'}) \\
&\leq \sum_{j \neq j'} \left(\sum_{i=1}^{k} c_i \mathbb{E}[\pi_{ij}\pi_{1j}] \right)^2 + C'' d^{-3} \\
&= d(d-1) \left(\frac{1+c_1}{d(d+1)} \right)^2 + C'' d^{-3} \leq \left(\frac{1+c_1}{d+1} \right)^2 + C'' d^{-3},
\end{aligned}$$

for another universal constant $C'' > 0$. We thus have

$$(IV) = n(n-1)m(m-1)\big[(a)+(b)\big] \leq (nm)^2 \left[\frac{C'''}{d^3} + \left(\frac{1+c_1}{d+1} \right)^2 \right],$$

for $C'''_k = C'_k + C''_k$. Altogether, we deduce that

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{\ell=1}^{n} \sum_{j=1}^{m} I(W_\ell = Z_{1j}) \right] \\
&= (I) + (II) + (III) + (IV) - \Big[\mathbb{E}\Big(\textstyle\sum_{\ell=1}^{n} \sum_{j=1}^{m} I(W_\ell = Z_{1j}) \Big) \Big]^2 \\
&= (I) + (II) + (III) + (IV) - (nm)^2 \left(\frac{1+c_1}{d+1} \right)^2 \\
&\lesssim \frac{nm}{d} + \frac{nm^2}{d^2} + \frac{mn^2}{d^2} + \frac{(nm)^2}{d^3}.
\end{aligned}$$

It follows that

$$\mathrm{Var}[\widehat{c}_1] \lesssim \frac{d^2}{(nm)^2}\,\mathrm{Var}\left[\left(\sum_{\ell=1}^{n}\sum_{j=1}^{m} I(W_\ell = Z_{ij})\right)^2\right] \lesssim \frac{d}{nm} + \frac{1}{n} + \frac{1}{m} + \frac{1}{d}.$$

Since $n \le d$, the claim follows. □

b. *Proof of Lemma 14*

By condition **(S)** and Lemma 1, it will suffice to derive an upper bound on the normalized Poisson minimax risk $\mathcal{R}'(n, d, k, m)$. Recall that, under this model, one observes histograms (Y, V) with entries that are conditionally independent given Π, with Π admitting independent rows distributed according to the flat Dirichlet law. We will analyze the associated hard-thresholded collision estimator, defined by

$$\widetilde{c}_i = \widehat{c}_i \cdot I(\widehat{c}_i \ge \lambda), \quad \text{where } \widehat{c}_i = \frac{d}{nm}(VY)_i - 1 = \frac{d}{nm}\sum_{j=1}^{d} V_{ij}Y_j - 1, \quad i = 1, \dots, k,$$

where $\lambda = \sqrt{ad\log(k)/(nm_d)}$, for a constant $a > 0$ depending only on γ, to be specified below. Throughout the proof, $C > 0$ denotes a universal constant whose value may change from one expression to the next. Our starting point is the following basic inequality for hard-thresholding estimators [65]:

$$(\widetilde{c}_i - c_i)^2 \le 4(c_i \wedge \lambda)^2 + |\widehat{c}_i - c_i|^2 I(|\widehat{c}_i - c_i| > \lambda/2), \quad i = 1, \dots, k.$$

Notice that

$$\sum_{i=1}^{k} c_i^2 \wedge \lambda^2 = \sum_{i:c_i \le \lambda} c_i^2 + \sum_{i:c_i > \lambda} \lambda^2 \le \lambda\|c\|_1 + |\{i : c_i > \lambda\}|\lambda^2 \lesssim \lambda \asymp \sqrt{\frac{da\log(k)}{nm_d}},$$

thus, to prove the claim, it will suffice to show that $\mathbb{E}[R] \lesssim \sqrt{d\log(k)/nm_d}$, where

$$R = \max_{1\le i\le k} R_i, \quad R_i = k \cdot \mathbb{E}\Big[|\widehat{c}_i - c_i|^2 \cdot I(|\widehat{c}_i - c_i| > \lambda/2)\,\Big|\,\Pi\Big].$$

Let us now derive a concentration bound for $\widehat{c}_i$ conditionally on Π. We will repeatedly make use of the following concentration bounds for sub-Weibull random variables, which are due to Kuchibhotla and Chakrabortty [135].

Lemma 15 (Concentration of sub-Weibull Random Variables)**.** *Let $X_1, \dots, X_d$ be independent mean-zero random variables in $\mathbb{R}$, such that for some $\alpha \in (0, 1]$ and $\zeta > 0$,*

$$\max_{1\le j\le d} \|X_j\|_{\psi_\alpha} \le \zeta, \quad \textit{and define } \sigma^2 = \sum_{j=1}^{d} \mathbb{E}[X_i^2].$$

Then, there exists a constant $C_\alpha > 0$ such that the following assertions hold.

1. *(Theorem 3.1, [135]) For all $t > 0$, it holds with probability at least $1 - 2e^{-t}$ that*

$$\left|\sum_{j=1}^{d} X_j\right| \le C_\alpha\zeta\big(\sqrt{dt} + t^{1/\alpha}\big).$$

2. (Theorem 3.4, [135]) For all $t > 0$, it holds with probability at least $1 - 3e^{-t}$ that

$$\left|\sum_{j=1}^{d} X_j\right| \le C_\alpha \left(\sqrt{\sigma^2 t} + \zeta (t \log d)^{1/\alpha}\right).$$

We will also make use of the following bound from [135], which characterizes the tail behavior of products of sub-Weibull random variables.

Lemma 16 ([135], Proposition D.2)**.** *If $X_1, \dots, X_d$ are (possibly dependent) random variables satisfying $\|X_j\|_{\psi_{\alpha_j}} < \infty$ for some $\alpha_j > 0$, then*

$$\left\|\prod_{j=1}^{d} X_j\right\|_{\psi_\beta} \le \prod_{j=1}^{d} \|X_j\|_{\psi_{\alpha_j}}, \quad \text{where } \frac{1}{\beta} = \sum_{j=1}^{d} \frac{1}{\alpha_j}.$$

Now, using Proposition 6.5 of [136], it can be shown that the ψ_1-Orlicz norm of a Poisson random variable $X \sim \mathrm{Poi}(\lambda)$ satisfies $\|X - \lambda\|_{\psi_1} \lesssim 1 \vee \sqrt{\lambda}$. Thus, denoting by $\|\cdot\|_{\psi_{\alpha,\Pi}}$ the Orlicz norm taken with respect to the conditional law $\mathbb{P}(\cdot|\Pi)$, one has

$$\big\|Y_j - \mathbb{E}[Y_j|\Pi]\big\|_{\psi_1,\Pi} \lesssim 1 \vee \mu_{Y_j}, \quad \big\|V_{ij} - \mathbb{E}[V_{ij}|\Pi]\big\|_{\psi_1,\Pi} \lesssim 1 \vee \mu_{V_{ij}},$$

with $\mu_{Y_j} = n\Pi_{\cdot j}^\top c$ and $\mu_{V_{ij}} = m\pi_{ij}$, hence,

$$\begin{aligned}
&\big\|V_{ij}Y_j - \mathbb{E}[V_{ij}Y_j|\Pi]\big\|_{\psi_{1/2,\Pi}} \\
&\quad\lesssim \big\|(V_{ij} - \mathbb{E}[V_{ij}|\Pi])(Y_j - \mathbb{E}[Y_j|\Pi])\big\|_{\psi_{1/2,\Pi}} + \mu_{Y_j}\|V_{ij}\|_{\psi_{1/2,\Pi}} + \mu_{V_{ij}}\|Y_j\|_{\psi_{1/2,\Pi}} + \mu_{Y_j}\mu_{V_{ij}} \\
&\quad\lesssim (1 \vee \mu_{Y_j})(1 \vee \mu_{V_{ij}}),
\end{aligned}$$

where we used Lemma 16 on the final line. We write

$$\zeta_{\Pi,i} = \max_{1 \le j \le d} (1 \vee \mu_{Y_j})(1 \vee \mu_{V_{ij}}), \quad i = 1, \dots, k.$$

We may then apply the concentration inequality of Lemma 15(ii) for $\psi_{1/2}$-random variables to obtain

$$\mathbb{P}\left(\left|\sum_{j=1}^{d} \big(V_{ij}Y_j - \mathbb{E}(V_{ij}Y_j|\Pi)\big)\right| > x \,\middle|\, \Pi\right) \le 2\exp\left\{-\frac{1}{C}\left(\frac{x^2}{\sigma^2_{\Pi,i}} \wedge \frac{\sqrt{x/\zeta_{\Pi,i}}}{\log d}\right)\right\}, \tag{E1}$$

for all $x > 0$, where

$$\sigma^2_{\Pi,i} = \sum_{j=1}^{d} \mathrm{Var}[V_{ij}Y_j|\Pi].$$

It follows that

$$\mathbb{P}\big(|\widehat{c}_i - \mathbb{E}[\widehat{c}_i \,|\, \Pi]| > x \,|\, \Pi\big) \le 2\exp\left\{-\frac{1}{C}\left(\frac{(nm)^2 x^2}{d^2 \sigma^2_{\Pi,i}} \wedge \frac{1}{\log d}\sqrt{\frac{nmx}{d\zeta_{\Pi,i}}}\right)\right\}. \tag{E2}$$

In particular, denoting by $\beta_{\Pi,i} = |\mathbb{E}[\widehat{c}_i|\Pi] - c_i|$ the conditional bias of $\widehat{c}_i$, we deduce that

$$\mathbb{P}\big(|\widehat{c}_i - c_i| > x \,|\, \Pi\big) \le 2\exp\left\{-\frac{1}{C}\left(\frac{(nm)^2 (x - \beta_{\Pi,i})_+^2}{d^2 \sigma^2_{\Pi,i}} \wedge \frac{1}{\log d}\sqrt{\frac{nm(x - \beta_{\Pi,i})_+}{d\zeta_{\Pi,i}}}\right)\right\}. \tag{E3}$$

where the transition between the sub-Gaussian and sub-Weibull tail occurs when the point x exceeds $\tau_{\Pi,i} := (d/nm)(\sigma^2_{\Pi,i}/\sqrt{\zeta_{\Pi,i}})^{2/3} + \beta_{\Pi,i}$. With these preliminaries in place, we obtain

$$\begin{aligned} R_i &= \int_{\lambda^2/4}^{\infty} \mathbb{P}\big(|\widehat{c}_i - c_i|^2 > u \,|\, \Pi\big) du \\ &\lesssim k \cdot \int_{\lambda^2/4}^{\infty} \exp\left\{ -\frac{1}{C_0}\left(\frac{(nm)^2(\sqrt{x} - \beta_{\Pi,i})_+^2}{d^2\sigma^2_{\Pi,i}} \wedge \frac{1}{\log d}\sqrt{\frac{nm(\sqrt{x} - \beta_{\Pi,i})_+}{d\zeta_{\Pi,i}}} \right) \right\} dx, \end{aligned}$$

for a sufficiently large constant $C_0 > 0$. Now, define for all $i = 1, \ldots, k$ the quantities

$$\begin{aligned} \lambda_{\Pi,i} &= (\lambda/2 - \beta_{\Pi,i})_+ \\ K_{\Pi,i} &= k \exp\left\{ -\frac{(nm)^2\lambda^2_{\Pi,i}}{C_0 d^2 \sigma^2_{\Pi,i}} \right\} + k \exp\left\{ -\frac{1}{C_0 \log d}\sqrt{\frac{nm\lambda_{\Pi,i}}{d\zeta_{\Pi,i}}} \right\}. \end{aligned}$$

In particular, notice that $K_{\Pi,i}$ solves

$$k \le K_{\Pi,i} \cdot \exp\left\{ -\frac{1}{C_0}\left(\frac{(nm)^2(\sqrt{x} - \beta_{\Pi,i})_+^2}{d^2\sigma^2_{\Pi,i}} \wedge \frac{1}{\log d}\sqrt{\frac{nm(\sqrt{x} - \beta_{\Pi,i})_+}{d\zeta_{\Pi,i}}} \right) \right\} \quad \text{for all } x > \lambda^2/4,$$

thus we obtain

$$\begin{aligned} R_i &\le K_{\Pi,i} \int_{\lambda^2/4}^{\tau^2_{\Pi,i}} \exp\left\{ -\frac{(nm)^2(\sqrt{x} - \beta_{\Pi,i})_+^2}{C_0 d^2 \sigma^2_{\Pi,i}} \right\} dx + K_{\Pi,i} \int_{\tau^2_{\Pi,i}}^{\infty} \exp\left\{ -\frac{1}{C_0}\sqrt{\frac{nm(\sqrt{x} - \beta_{\Pi,i})_+}{d(\log d)^2\zeta_{\Pi}}} \right\} dx \\ &\le K_{\Pi,i}\left\{ (\beta^2_{\Pi,i} - \frac{\lambda^2}{4})_+ + \frac{d^2\sigma^2_{\Pi,i}}{(nm)^2} + \beta_{\Pi,i}\frac{d\sigma_{\Pi,i}}{nm} \right\} + K_{\Pi,i}\left\{ \frac{d^2\,(\log d)^4\,\zeta^2_{\Pi,i}}{(nm)^2} + \beta_{\Pi,i}\,\frac{d\,(\log d)^2\,\zeta_{\Pi,i}}{nm} \right\} \end{aligned}$$

where we used the following elementary integral identities, which hold for all $a, b, f > 0$,

$$\begin{aligned} \int_a^{\infty} e^{-b(\sqrt{x} - f)_+^2} dx &\lesssim (f^2 - a)_+ + \frac{1}{b} + \frac{f}{\sqrt{b}}, \\ \int_a^{\infty} e^{-b(\sqrt{x} - f)_+^{1/2}} dx &\lesssim (f^2 - a)_+ + \frac{1}{b^4} + \frac{f}{b^2}. \end{aligned}$$

Thus, if we define the event

$$A = \bigcap_{i=1}^{k} \left\{ \sigma^2_{\Pi,i} \le C_1\frac{nm}{d} + C_2\frac{nm^2}{d^2} + C_3\frac{n^2m}{d^2} \right\} \cap \big\{ \beta_{\Pi,i} \le \frac{\lambda}{4} \big\} \cap \big\{ \zeta_{\Pi,i} \le 1 + \frac{m}{d} \big\} \cap \big\{ K_{\Pi,i} \le 1 \big\}, \tag{E4}$$

for large enough constants $C_1, C_2, C_3 > 0$ to be defined below, then, under condition **(S)**, we obtain

$$\begin{aligned} \mathbb{E}[R] &\lesssim \mathbb{E}[R \cdot I(A^{\mathsf{c}})] \\ &\quad + \mathbb{E}\left\{ \max_{1 \le i \le k} K_{\Pi,i} \left[(\beta^2_{\Pi,i} - \frac{\lambda^2}{4})_+ + \frac{d^2\sigma^2_{\Pi,i}}{(nm)^2} + \beta_{\Pi,i}\frac{d\sigma_{\Pi,i}}{nm} + \left(\frac{d(\log d)^2\zeta_{\Pi,i}}{nm} \right)^2 \right] \right\} \\ &\lesssim d \cdot \mathbb{P}(A^{\mathsf{c}}) + \lambda, \end{aligned} \tag{E5}$$

where we used the fact that $\widehat{c}_i \le d$, thus $R \lesssim d$. To complete the claim, it will thus suffice to show that the event A occurs with sufficiently high probability. To this end, we will

provide high-probability bounds on the quantities $\sigma^2_{\Pi,i}$, $\beta_{\Pi,i}$, $\zeta_{\Pi,i}$ and $K_{\Pi,i}$, in turn.

Bounding term $\sigma^2_{\Pi,i}$. Recall that the entries of the matrix Π take the form $\pi_{ij} = \varpi_{ij}/S_i$, where $S_i = \sum_{\ell=1}^d \varpi_{i\ell}$, and $\varpi_{ij} \sim \mathcal{E}_d$ are i.i.d. exponential random variables. Furthermore, write $X = (\varpi_{ij} : 1 \le i \le k, 1 \le j \le d) \in \mathbb{R}^{k \times d}$. We have

$$\begin{aligned}
\sigma^2_{\Pi,i} &= \sum_{j=1}^d \mathrm{Var}[V_{ij}Y_j|\Pi] \\
&= \sum_{j=1}^d \Big(\mathrm{Var}[V_{ij}|\Pi]\,\mathrm{Var}[Y_j|\Pi] + \mathrm{Var}[Y_j|\Pi]\big(\mathbb{E}[V_{ij}|\Pi]\big)^2 + \mathrm{Var}[V_{ij}|\Pi]\big(\mathbb{E}[Y_j|\Pi]\big)^2 \Big) \\
&= \sum_{j=1}^d \Big(m\pi_{ij} n(\Pi^\top c)_j + n(\Pi^\top c)_j (m\pi_{ij})^2 + m\pi_{ij}(n(\Pi^\top c)_j)^2 \Big) \\
&= \sum_{j=1}^d \sum_{s=1}^k c_s \Big(mn\pi_{ij}\pi_{sj} + nm^2 \pi_{sj}\pi_{ij}^2 + mn^2 \sum_{s'=1}^k c_{s'}\pi_{s'j}\pi_{ij}\pi_{sj} \Big) \\
&= mn \sum_{j=1}^d \sum_{s=1}^k c_s \frac{\varpi_{ij}\varpi_{sj}}{S_i S_s} + nm^2 \sum_{j=1}^d \sum_{s=1}^k c_s \frac{\pi_{ij}^2 \pi_{sj}}{S_i^2 S_s} + mn^2 \sum_{j=1}^d \sum_{s,s'=1}^k c_s c_{s'} \frac{\pi_{ij}\pi_{sj}\pi_{s'j}}{S_i S_s S_{s'}} \\
&=: T_{1i} + T_{2i} + T_{3i}.
\end{aligned}$$

We bound the terms T_{1i}, T_{2i}, and T_{3i} in turn. By a sub-exponential tail bound, notice that one has

$$\mathbb{P}\big(\max_{1 \le i \le k} |S_i - 1| > x \big) \lesssim k \cdot \exp\big(-Cd(x \wedge x^2)\big), \quad \text{for all } x > 0. \tag{E6}$$

Under condition (S), we deduce that the event $\mathcal{A}_1 = \bigcap_{i=1}^k \{|S_i - 1| \le 1/2\}$ satisfies $\mathbb{P}(\mathcal{A}_1) \ge 1 - Ce^{-d/C}$. Over the event $\mathcal{A}_1$, we deduce that for all $i = 1, \dots, k$,

$$\frac{T_{1i}}{nm} \asymp \sum_{j=1}^d \sum_{s=1}^k c_s \varpi_{ij}\varpi_{sj} = \sum_{j=1}^d F_{ij}, \quad \text{where } F_{ij} = \sum_{s=1}^k c_s \varpi_{ij}\varpi_{sj}. \tag{E7}$$

Recalling the Orlicz norms $\|\cdot\|_{\psi_\alpha}$ defined in equation (A1), notice that for all $j = 1, \dots, d$,

$$\|F_{ij}\|_{\psi_{1/2}} \le \sum_{s=1}^k c_s \|\varpi_{ij}\varpi_{sj}\|_{\psi_{1/2}} \le \sum_{s=1}^k c_s \|\varpi_{ij}\|_{\psi_1} \|\varpi_{rj}\|_{\psi_1} \lesssim \frac{1}{d^2},$$

where the penultimate inequality follows from Lemma 16, and the final inequality is a simple consequence of the fact that the random variables $d\varpi_{ij}$ are (sub-)exponential with fixed modulus. We deduce that, up to rescaling, $F_{ij} - \mathbb{E}[F_{ij}]$ are (1/2)-sub-Weibull random variables, which are independent across $j = 1, \dots, d$. Applying Lemma 15(i), we deduce that for all $x > 0$,

$$\mathbb{P}\left(\max_{1 \le i \le k} \left| \sum_{j=1}^d \big(F_{ij} - \mathbb{E}[F_{ij}]\big) \right| \ge (C/d^2)\big(\sqrt{dx} + x^2\big) \right) \lesssim ke^{-x}. \tag{E8}$$

Furthermore, we readily have $\mathbb{E}[F_{ij}] \lesssim d^{-2}$, thus the above display implies that, for a large enough constant $C > 0$, the event $\mathcal{A}_2 := \{\max_{1 \le i \le k} |\sum_{j=1}^d F_{ij}| \le C/d\}$ has probability

content $\mathbb{P}(\mathcal{A}_2^\complement) \leq ke^{-\sqrt{d}} \lesssim e^{-\sqrt{d}/C}$. We deduce that over $\mathcal{A}_1 \cap \mathcal{A}_2$, it holds that

$$\max_{1\leq i\leq k} T_{1i} \leq C_1 \frac{nm}{d}.$$

To bound T_{2i}, we adopt a similar proof. We again have, over the event $\mathcal{A}_1$,

$$\frac{T_{2i}}{nm^2} \asymp \sum_{j=1}^{d}\sum_{s=1}^{k} c_s \varpi_{ij}^2 \varpi_{sj} = \sum_{j=1}^{d} L_{ij}, \quad \text{where } L_{ij} = \sum_{s=1}^{k} c_s \varpi_{ij}^2 \varpi_{sj}. \tag{E9}$$

The random variables $d^3 L_{ij}$ are (1/3)-sub-Weibull, since, reasoning as before, one has

$$\|L_{ij}\|_{\psi_{1/3}} \leq \sum_{s=1}^{k} c_s \|\varpi_{ij}\|_{\psi_1}^2 \|\varpi_{rj}\|_{\psi_1} \lesssim \frac{1}{d^3}.$$

Furthermore, one has $\mathbb{E}[L_{ij}] \lesssim d^{-3}$, thus by again applying the sub-Weibull tail bound from Lemma 15(i), we arrive at

$$\mathbb{P}\bigg(\max_{1\leq i\leq k} \sum_{j=1}^{d} L_{ij} \geq C/d^2 + (C/d^3)\big(\sqrt{dx} + x^3\big)\bigg) \lesssim ke^{-x}, \quad x > 0,$$

which implies that the event $\mathcal{A}_3 := \{\max_{1\leq i\leq k} |\sum_{j=1}^{d} L_{ij}| \leq C/d^2\}$ satisfies $\mathbb{P}(\mathcal{A}_3) \geq 1 - e^{-d^{1/3}/C}$. Over the event $\mathcal{A}_1 \cap \mathcal{A}_3$, we thus obtain

$$\max_{1\leq i\leq k} T_{2i} \leq C_2 \frac{nm^2}{d^2}.$$

for a large enough choice of the constant $C_2 > 0$. An analogous proof can be used to show that, for an event $\mathcal{A}_4$ satisfying $\mathbb{P}(\mathcal{A}_4) \geq 1 - e^{-d^{1/3}/C}$,

$$\max_{1\leq i\leq k} T_{3i} \leq C_3 \frac{n^2 m}{d^2}.$$

Altogether, we have thus shown that

$$\max_{1\leq i\leq k} \sigma_{\Pi,i}^2 \leq C_1 \frac{nm}{d} + C_2 \frac{nm^2}{d^2} + C_3 \frac{n^2 m}{d^2}. \tag{E10}$$

over the event $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4$. This completes our upper bound of $\sigma_{\Pi,i}^2$.

Bounding term $\beta_{\Pi,i}$. Next, we provide a high-probability bound on the conditional bias term $\beta_{\Pi,i}$. We will make use of the following simple Lemma.

Lemma 17. *Define the random variables $B_i = (d+1)\sum_{j=1}^{d}\sum_{r=1}^{k} c_r \varpi_{ij}\varpi_{rj} - 1$. Then, there exists a constant $C_3 > 0$ such that the event*

$$\mathcal{A}_5 = \bigcap_{i=1}^{k} \Big\{ \Big| \big(\mathbb{E}[\widehat{c}_i|\Pi] - c_i\big) - \big(B_i - \mathbb{E}[B_i]\big) \Big| \leq C_3/d \Big\}$$

satisfies $\mathbb{P}(\mathcal{A}_5) \geq 1 - C_3 e^{-\sqrt{d}/C_3}$.

The proof appears in Appendix G 3 a. In view of Lemma 17, we can bound $\beta_{\Pi,i}$ in the same way as $\sigma^2_{\Pi,i}$. Indeed, notice that $B_i + 1 \asymp d\sum_{j=1}^d F_{ij}$, thus, using equation (E8), we have for all $x > 0$,

$$\mathbb{P}\left(\max_{1\le i\le k} \left|B_i - \mathbb{E}[B_i]\right| > (C/d)(\sqrt{dx} + x^2)\right) \lesssim ke^{-x}.$$

Notice that $\lambda \ge n^{-1/2} \ge d^{-\frac{1}{2(1+\gamma)}}$ under condition **(S)**, thus, by choosing $x = d^{\gamma/(2(1+\gamma))-\epsilon}$ for any fixed $\epsilon > 0$, we deduce that the event $\mathcal{A}_6 = \{\max_{1\le i\le k} |B_i - \mathbb{E}[B_i]| > \lambda/8\}$ satisfies $\mathbb{P}(\mathcal{A}_6^{\mathsf{c}}) \lesssim e^{-d^b/C}$ for a fixed constant $b = b(\gamma) > 0$. It thus follows that, over the event $\mathcal{A}_5 \cap \mathcal{A}_6$,

$$\max_{1\le i\le k} \beta_{\Pi,i} \le \lambda/4. \tag{E11}$$

Bounding term $\zeta_{\Pi,i}$. By repeating the same arguments as in the previous steps, there exists an event $\mathcal{A}_7$ of probability content at least $1 - e^{-\sqrt{d}/C}$ over which it holds that

$$\mu_{Y_j} = n\sum_{i=1}^k c_i\pi_{ij} \le 2n/d, \quad \text{and,} \quad \mu_{V_{ij}} = m\pi_{ij} \le 2m/d,$$

uniformly in i, j. Under condition **(S)**, it follows that $\zeta_{\Pi,i} \le C(1 + \sqrt{m/d})$.

Bounding term $K_{\Pi,i}$. The preceding steps readily lead to an upper bound on $K_{\Pi,i}$. Indeed, under condition **(S)**, over the event $\bigcap_{s=1}^7 \mathcal{A}_s$, we have for all $i = 1, \dots, k$ that $\lambda_{\Pi,i} \ge \lambda/2$ and

$$\begin{aligned}
K_{\Pi,i} &= k\exp\left\{-\frac{(nm)^2\lambda^2_{\Pi,i}}{C_0 d^2\sigma^2_{\Pi,i}}\right\} + o(1)\\
&\lesssim k\exp\left\{-\frac{1}{C}\frac{nm^2 a\log(k)}{dm_d(nm/d + nm^2/d^2 + mn^2/d^2)}\right\} + o(1)\\
&\lesssim k\exp\left\{-\frac{a\log(k)}{C}\right\} + o(1)\\
&\lesssim k\cdot k^{-a/C} + o(1) < 1,
\end{aligned} \tag{E12}$$

for a sufficiently large choice of a, depending only on γ.

Concluding the proof. By combining equations (E10), (E11), and (E12), we deduce that the set A defined in equation (E4) satisfies

$$\mathbb{P}(A^{\mathsf{c}}) \lesssim \sum_{i=1}^5 \mathbb{P}(\mathcal{A}_i^{\mathsf{c}}) \lesssim e^{-d^b},$$

for a sufficiently small exponent $b > 0$ depending only on γ. Therefore, returning to equation (E5), we arrive at

$$\mathbb{E}[R] \lesssim de^{-d^b} + \lambda \lesssim \lambda.$$

The claim follows from here. □

2. Proof of Propositions 4–6

We will prove the three claims by first analyzing the risk of the cumulant estimators $\widehat{\xi}_p$, under the unnormalized Poisson model.

Lemma 18. *Assume condition (S) and let $1 \le p \le k$. Then, under the unnormalized Poisson model, there exists a constant $C_{p,\gamma} > 0$ such that*

$$\mathbb{E}|\widehat{\xi}_p - \xi_p| \le \frac{C_{p,\gamma}}{\sqrt{n^p d^{p+1}}}.$$

Proof. We will make use of the following fact.

Lemma 19. *Under the unnormalized Poisson model, for all $i = 1, \dots, k$, there exists a constant $C_i > 0$ such that*

$$\mathbb{E}[T_{1,i}] = \eta_i, \quad \mathrm{Var}[T_{1,i}] \le \frac{C_i}{\min\{n,d\}^i d^i}.$$

The proof appears in Appendix G 2 b. It follows from Lemma 19 that

$$\mathbb{E}[W_{\mathbf{h}}] = \mathbb{E}\left[\prod_{i=1}^{p-\ell+1} \prod_{j \in S_i} T_{j,i}\right] = \prod_{i=1}^{p-\ell+1} \prod_{j \in S_i} \mathbb{E}\left[T_{j,i}\right] = \prod_{i=1}^{p-\ell+1} \prod_{s=1}^{h_i} \eta_i = \prod_{i=1}^{p-\ell+1} \eta_i^{h_i}, \tag{E13}$$

thus it is clear that $\mathbb{E}[\widehat{\xi}_p] = \xi_p$. Let us now compute the variance. Notice that:

$$\mathrm{Var}[\widehat{\xi}_p] \lesssim \sum_{\mathbf{h} \in \mathcal{H}_{p,\ell}} \mathrm{Var}[W_{\mathbf{h}}],$$

where the implicit constants depend on k, p. Now, let us rewrite $W_{\mathbf{h}}$ as the p-th order U-Statistic:

$$W_{\mathbf{h}} = \frac{1}{\binom{d}{p}} \sum_{1 \le j_1 < \cdots < j_p \le d} \zeta_{\mathbf{h}}(j_1, \dots, j_p),$$

where

$$\zeta_{\mathbf{h}}(j_1, \dots, j_p) = \frac{\prod_{i=1}^k h_i!}{p!} \sum_{(A_1, \dots, A_{p-\ell+1})} \prod_{i=1}^{p-\ell+1} \prod_{j \in A_i} T_{j,i},$$

and where the summation is taken over all partitions $A_1, \dots, A_{p-\ell+1}$ of $\{j_1, \dots, j_p\}$ such that $|A_i| = h_i$ for all i. By [137], we have for any $\mathbf{h} \in \mathcal{H}_{p,\ell}$ and any large enough d that

$$\mathrm{Var}[W_{\mathbf{h}}] \lesssim \frac{1}{d} \mathrm{Var}\left[\zeta_{\mathbf{h}}(1, \dots, p)\right] \lesssim \frac{1}{d} \mathrm{Var}\left[\prod_{i=1}^{p-\ell+1} \prod_{j \in A_i} T_{j,i}\right],$$

for any fixed partition $(A_1, \dots, A_{p-\ell+1})$ of $\{1, \dots, j\}$ with $|A_i| = h_i$. The random variables appearing in the above product are independent, thus, together with Lemma 19, we have

$$\mathrm{Var}\left[\prod_{i=1}^{p-\ell+1} \prod_{j \in A_i} T_{j,i}\right] \le \prod_{i=1}^{p-\ell+1} \prod_{j \in A_i} \left(\mathrm{Var}\left[T_{j,i}\right] + \mathbb{E}\left[T_{j,i}\right]^2\right)$$

$$\lesssim \prod_{i=1}^{p-\ell+1} \prod_{j\in A_i} \left(\frac{1}{n^i d^i} + \frac{1}{d^{2i}}\right)$$
$$\lesssim \prod_{i=1}^{p-\ell+1} \left(\frac{1}{n^i d^i}\right)^{h_i} = (nd)^{-\sum_i i h_i} = n^{-p} d^{-p}.$$

The claim follows. □

The following is now an immediate consequence of Lemma 18 and the definition of the estimator $\widehat{m} = (\widehat{m}_1, \dots, \widehat{m}_p)^\top$ defined in equation (B25).

Lemma 20. *Assume condition (S). Then, under the unnormalized Poisson model, there exists a constant $C = C(k, \gamma) > 0$ such that*

$$\mathbb{E}\|\widehat{m} - m(c)\|_2 \le C\sqrt{\frac{d^{k-1}}{n^k}}. \tag{E14}$$

With Lemma 20 in hand, we are now ready to prove Propositions 4–6. For the proofs of Propositions 4–5, notice that $W(\widehat{c}, c) \le W(\widetilde{c}, c)$ (cf. Lemma 48 of [74]), thus it suffices to prove upper bounds for the possibly complex-valued estimator $\widetilde{c}$. By Newton's identity (K3), the moments of $\widetilde{c}$ are given by

$$m_p(\widetilde{c}) = \widehat{m}_p, \quad p = 1, \dots, k.$$

Therefore, we may apply Lemma 28 to obtain

$$\mathbb{E}W(\widetilde{c}, c) \lesssim \mathbb{E}\big\|m(\widetilde{c}) - m(c)\big\|^{\frac{1}{k}} \lesssim \sqrt{\frac{d^{1-\frac{1}{k}}}{n}},$$

where the implicit constants depend only on k, γ. This proves Proposition 4. Proposition 5 follows similarly, by now replacing the $1/k$-modulus of continuity in the above display by $1/(k - k_0 + 1)$, as a result of Lemma 29.

Finally, to prove Proposition 6, invoke Proposition 4 to deduce that the loss functions $\mathcal{D}_{c^\star}(\widehat{c}, c)$ and $\overline{\mathcal{D}}_{c^\star}(\widehat{c}, c)$ (defined in Appendix K 2 a) coincide for all large enough d. By again applying Lemma 48 of [74], we thus obtain

$$\mathcal{D}_{c^\star}(\widehat{c}, c) = \overline{\mathcal{D}}_{c^\star}(\widehat{c}, c) \le \overline{\mathcal{D}}_{c^\star}(\widetilde{c}, c).$$

The claim now follows as before, invoking Lemma 29 to bound $\mathcal{D}_{c^\star}$ in terms of the moment differences. The claim follows. □

3. Moment Estimator in the Multinomial Model

We now show that Propositions 4–6 lead to upper bounds on the sorted minimax risk $\mathcal{M}_<(n, d, k, 0)$ for the original multinomial sampling model, as well as for the local minimax risk:

$$\mathcal{M}_<(n, d, k; c^\star, \epsilon) = \inf_{\widehat{c}} \sup_{\substack{c\in\Delta_k \\ W(c,c^\star)\le\epsilon}} \mathbb{E}_c\Big[\mathcal{D}_{c^\star}(\widehat{c}(Y, V), c)\Big],$$

which is defined for any $c^\star \in \Delta_k$ and $\epsilon \ge 0$, where the expectation is taken over a realization (Y, V) from the multinomial model with parameter c.

Corollary 1. *Assume condition (S), and $n > d^{1-\frac{1}{k}}$. Then, the following assertions hold.*

(i) There exists a constant $C = C(k, \gamma) > 0$ such that

$$\mathcal{M}_<(n, d, k, 0) \le C\sqrt{\frac{d^{1-\frac{1}{k}}}{n}}.$$

(ii) For any $1 \le k_0 \le k$, $\delta > 0$, there exist constants $C, \epsilon > 0$ depending on k, γ, δ such that for any $c^\star \in \Delta_{k,k_0}(\delta)$,

$$\mathcal{M}_<(n, d, k; c^\star, \epsilon) \le C\sqrt{\frac{d^{k-1}}{n^k}}.$$

To prove Corollary 1(i), notice that

$$\mathcal{M}_<(n, d, k, 0) \lesssim \mathcal{R}_<(n, d, k, 0) + e^{-n/C_1} \lesssim \mathcal{R}'_<(n, d, k, 0) + e^{-n/C_1} + \sqrt{\frac{n}{d}},$$

for a large enough constant $C_1 > 0$, due to Lemmas 1–2. By Proposition 4, we deduce

$$\mathcal{M}_<(n, d, k, 0) \lesssim \sqrt{\frac{d^{1-\frac{1}{k}}}{n}} + e^{-n/C_1} + \sqrt{\frac{n}{d}}.$$

Under the condition $n \ge d^{1-\frac{1}{k}}$, the first term on the right-hand side of the above display is dominant, which proves Corollary 1(i). The second claim can be proven analogously, by again using the fact that $\sqrt{d^{k-1}/n^k}$ dominates $\sqrt{n/d}$ in the regime $n \ge d^{1/k}$. □

Appendix F: Proofs of Main Results

Our main results—namely Theorems 1–2 and Proposition 1—now follow from the lower and upper bounds developed in the preceding two appendices. Concretely, Theorem 1 follows from the lower bound in Proposition 8 and the upper bound in Proposition 3, while Theorem 2 follows from the lower bound in Proposition 7 and the upper bounds in Proposition 4 (where we recall that $W \le \|\cdot\|$), and Corollary 1(i). Finally, Proposition 1 is a direct consequence of Corollary 1(ii) with $k_0 = 2$ and $r_1 = 1$.

Appendix G: Proofs Deferred from Appendices C–E

1. Proofs Deferred from Appendix C

a. *Proof of Lemma 1*

Our proof follows a standard Poissonization argument [122]. We prove the claim for the unordered minimax risk, and a similar proof can then be used for the ordered risk. Furthermore, we focus on the case $m > 0$; adaptations to the special case $m = 0$ are straightforward.

Let $(Y^{(n)}, V^{(m)})$ be random variables drawn from the multinomial model $\widetilde{Q}_c$ with sample sizes n and m. Let $N \sim \mathrm{Poi}(n)$ and $M \sim \mathrm{Poi}(m)$ be independent of all other random

variables, and notice that $(Y^{(N)}, V^{(M)})$ is distributed according to $\overline{Q}_c$. In this notation, the unordered minimax risks are given by

$$\mathcal{M}(n,d,k,m) := \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}(Y^{(n)}, V^{(m)}) - c\|$$
$$\mathcal{R}'(n,d,k,m) := \inf_{\widehat{c}} \sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}(Y^{(N)}, V^{(M)}) - c\|.$$

Now, given $\epsilon > 0$, let $\widehat{c}_\epsilon$ be a near-optimal estimator satisfying

$$\sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}_\epsilon(Y^{(n)}, V^{(m)}) - c\| \le \mathcal{M}(n,d,k,m) + \epsilon.$$

Writing $E_{n'm'} = \{N = n', M = m'\}$, we thus have,

$$\begin{aligned}
&\sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}_\epsilon(Y^{(N)}, V^{(M)}) - c\| \\
&\quad= \sup_{c\in\Delta_k} \mathbb{E}_\Pi\left\{\mathbb{E}_c\left[\|\widehat{c}_\epsilon(Y^{(N)}, V^{(M)}) - c\| \,\middle|\, \Pi\right]\right\} \\
&\quad= \sup_{c\in\Delta_k} \sum_{n',m'=0}^{\infty} \mathbb{E}_\Pi\left\{\mathbb{E}_c\left[\|\widehat{c}_\epsilon(Y^{(N)}, V^{(M)}) - c\| \,\middle|\, \Pi\right]\mathbb{P}(E_{n'm'}|\Pi)\right\} \\
&\quad= \sup_{c\in\Delta_k} \sum_{n',m'=0}^{\infty} \mathbb{E}_\Pi\left\{\mathbb{E}_c\left[\|\widehat{c}_\epsilon(Y^{(n')}, V^{(m')}) - c\| \,\middle|\, \Pi\right]\right\}\mathbb{P}(E_{n'm'}) \\
&\quad\le \sup_{c\in\Delta_k} \sum_{n',m'=0}^{\infty} \mathcal{M}(n',d,k,m')\mathbb{P}(E_{n'm'}) + \epsilon.
\end{aligned}$$

Since the risk function $\mathcal{M}(n,d,k,m)$ is monotonically decreasing in n and m, we deduce that

$$\begin{aligned}
&\sup_{c\in\Delta_k} \mathbb{E}_c\|\widehat{c}_\epsilon(Y^{(N)}, V^{(M)}) - c\| \\
&\qquad\le \mathcal{M}(n/2,d,k,m/2) + \mathbb{P}(N < n/2) + \mathbb{P}(M < m/2) + \epsilon \\
&\qquad\le \mathcal{M}(n/2,d,k,m/2) + Ce^{-n/C} + Ce^{-m/C} + \epsilon,
\end{aligned}$$

where the final inequality holds for a sufficiently large universal constant $C > 0$ by standard Chernoff bounds for the Poisson distribution (e.g. equation (C.1) of [122]). Since ϵ was arbitrary, it follows that

$$\mathcal{R}'(n,d,k,m) \le \mathcal{M}(n/2,d,k,m/2) + Ce^{-n/C} + Cke^{-m/C}. \tag{G1}$$

To prove a converse bound, we reason similarly as in Lemma 1 of [130], and use the fact that the worst-case Bayes risk provides a lower bound on the minimax risk:

$$\mathcal{R}'(n,d,k,m) \ge \sup_{\rho} \inf_{\widehat{c}} \mathbb{E}_{c\sim\rho}\big\{\mathbb{E}_c\|\widehat{c}(Y^{(N)}, V^{(M)}) - c\|\big\},$$

where the supremum is taken over all probability distributions on Δ_k. Reasoning similarly as before, we have for any prior ρ,

$$\begin{aligned}
&\inf_{\widehat{c}} \mathbb{E}_{c\sim\rho}\big\{\mathbb{E}_c\|\widehat{c}(Y^{(N)}, V^{(M)}) - c\|\big\} \\
&\quad= \inf_{\widehat{c}} \sum_{n',m'=0}^{\infty} \mathbb{E}_{c\sim\rho}\big\{\mathbb{E}_c\|\widehat{c}(Y^{(n')}, V^{(m')}) - c\|\big\} \cdot \mathbb{P}(N = n', M = m')
\end{aligned}$$

$$\geq \sum_{n'=0}^{2n}\sum_{m'=0}^{2m} \inf_{\widehat{c}} \mathbb{E}_{c\sim\rho}\big\{\mathbb{E}_c\|\widehat{c}(Y^{(n')},V^{(m')})-c\|\big\}\cdot\mathbb{P}(N=n',M=m')$$
$$\geq \mathbb{E}_{c\sim\rho}\big\{\mathbb{E}_c\|\widehat{c}(Y^{(2n)},V^{(2m)})-c\|\big\}\cdot\mathbb{P}(N\geq 2n,M\geq 2m),$$

where we again used the fact that the map $(n',m')\mapsto \inf_{\widehat{c}}\mathbb{E}_{c\sim\rho}\{\mathbb{E}_c\|\widehat{c}(Y^{(n')},V^{(m')})-c\|\}$ is decreasing in both of its coordinates. Now, by again applying a Poisson Chernoff bound, we obtain $\mathbb{P}(N\leq 2n,M\leq 2m)\leq 1-Ce^{-n/C}-Ce^{-m/C}$. Taking the supremum over ρ on both sides of the above display, we thus obtain

$$\mathcal{R}'(n,d,k,m)\geq \mathcal{M}(2n,d,k,2m)\cdot\big(1-Ce^{-n/C}-Cke^{-m/C}\big). \tag{G2}$$

This proves the claim. □

b. *Proof of Lemma 2*

Let $C>0$ be a universal constant, whose value may change from one display to the next.

Let $\Pi\in\mathbb{R}^{k\times d}$ be a random matrix with rows independently drawn from the flat Dirichlet law $\mathcal{D}_d$. Let $G_1,\dots,G_k\sim\text{Gamma}(d,d)$ be independent Gamma-distributed random variables, which are independent of Π, and notice that the matrix $X:=\text{diag}(G)\Pi$ consists of i.i.d. Exp(d)-distributed entries (cf. Lemma 26). Thus, we can write:

$$\mathbf{Q}_c^{\otimes d}=\mathbb{E}_{\Pi,G}[\mathbf{Q}_{c|\Pi,G}],\quad \text{with } \mathbf{Q}_{c|\Pi,G}=\bigotimes_{j=1}^{d}\left(\text{Poi}(n\sum_{i=1}^{k}G_ic_i\pi_{ij})\otimes\bigotimes_{i=1}^{k}\text{Poi}(mG_i\pi_{ij})\right)$$
$$\overline{\mathbf{Q}}_c=\mathbb{E}_{\Pi}[\overline{\mathbf{Q}}_{c|\Pi}],\qquad \text{with } \mathbf{Q}_{c|\Pi}=\bigotimes_{j=1}^{d}\left(\text{Poi}(n\sum_{i=1}^{k}c_i\pi_{ij})\otimes\bigotimes_{i=1}^{k}\text{Poi}(m\pi_{ij})\right).$$

By convexity of the TV distance, one has

$$\text{TV}(\overline{\mathbf{Q}}_c,\mathbf{Q}_c^{\otimes d})\leq\mathbb{E}_{\Pi,G}\Big[\text{TV}(\overline{\mathbf{Q}}_{c|\Pi},\mathbf{Q}_{c|\Pi,G})\Big]$$
$$\leq\mathbb{E}_{\Pi,G}\Big[\text{TV}(\overline{\mathbf{Q}}_{c|\Pi},\mathbf{Q}_{c|\Pi,G})\cdot I(A)\Big]+\mathbb{P}(A^{\mathsf{c}}),$$

where A denotes the event that $\pi_{ij}\leq 1$ and $|G_i-1|\leq d^{-1/4}$ for all $i=1,\dots,k$ and $j=1,\dots,d$. Notice that $\pi_{1i}\sim\text{Beta}(1,d-1)$, and is therefore sub-Gaussian with variance proxy $1/4(d+1)$ by [138]. Furthermore, $G_1\sim\text{Gamma}(d,d)$ can be expressed as $G_1=\frac{1}{d}\sum_{j=1}^{d}X_j$ where $X_j\sim\text{Exp}(1)$ are i.i.d sub-exponential random variables. By a sub-Gaussian and sub-exponential tail bound, one readily obtains

$$\mathbb{P}(A^{\mathsf{c}})\leq kd\cdot\mathbb{P}(\pi_{11}>1)+k\cdot\mathbb{P}(|G_1-1|>d^{-1/4})\lesssim kd\cdot e^{-d/C}+k\cdot e^{-\sqrt{d}/C}\lesssim e^{-\sqrt{d}/C}.$$

It thus remains to bound the mean value of $\text{TV}(\overline{\mathbf{Q}}_{c|\Pi},\mathbf{Q}_{c|\Pi,G})$ over the event A. By Pinsker's inequality, it will suffice to bound the KL divergence $\text{KL}^2(\overline{\mathbf{Q}}_{c|\Pi},\mathbf{Q}_{c|\Pi,G})$, which, over the event A, is bounded from above as follows (cf. Lemma 23):

$$\text{KL}(\overline{\mathbf{Q}}_{c|\Pi},\mathbf{Q}_{c|\Pi,G})$$
$$\lesssim\sum_{j=1}^{d}\left\{\text{KL}\Big(\text{Poi}\big(n\textstyle\sum_i c_iG_i\pi_{ij}\big),\text{Poi}\big(\sum_i c_i\pi_{ij}\big)\Big)+\sum_{i=1}^{k}\text{KL}\Big(\text{Poi}(mG_i\pi_{ij}),\text{Poi}(m\pi_{ij})\Big)\right\}$$
$$\lesssim n\sum_{j=1}^{d}\frac{(\sum_{i=1}^{k}c_i\pi_{ij}(G_i-1))^2}{\sum_{i=1}^{k}c_i\pi_{ij}}+m\sum_{j=1}^{d}\sum_{i=1}^{k}\frac{((G_i-1)\pi_{ij})^2}{\pi_{ij}}.$$

We thus have,

$$\begin{aligned}
&\mathbb{E}_{\Pi,G}\Big[\mathrm{TV}^2(\overline{\mathbf{Q}}_{c|\Pi}, \mathbf{Q}_{c|\Pi,G})\Big] \\
&\quad\lesssim e^{-\sqrt{d}/C} + nd\cdot\mathbb{E}_{\Pi,G}\left[\frac{(\sum_{i=1}^k c_i\pi_{ij}(G_i-1))^2}{\sum_{i=1}^k c_i\pi_{ij}}\right] + mdk\mathbb{E}_{\Pi,G}[(G_1-1)^2\pi_{11}] \\
&\quad\le e^{-\sqrt{d}/C} + n\cdot\mathbb{E}_{\Pi}\left[\frac{\mathrm{Var}_G\left[\sum_{i=1}^k c_i\pi_{ij}G_i \,\middle|\, \Pi\right]}{\sum_{i=1}^k c_i\pi_{ij}}\right] + mdk\,\mathrm{Var}[G_1]\cdot\mathbb{E}[\pi_{11}] \\
&\quad= e^{-\sqrt{d}/C} + \frac{n}{d}\mathbb{E}_{\Pi}\left[\frac{\sum_{i=1}^k c_i^2\pi_{ij}^2}{\sum_{r=1}^k c_r\pi_{rj}}\right] + \frac{mk}{d} \\
&\quad\le e^{-\sqrt{d}/C} + \frac{n}{d}\mathbb{E}_{\Pi}\left[\sum_{i=1}^k c_i\pi_{ij}\right] + \frac{mk}{d^2} = e^{-\sqrt{d}/C} + \frac{n}{d^2} + \frac{mk}{d^2}.
\end{aligned}$$

The claim follows. □

2. Proofs Deferred from Appendix D

a. *Proof of Lemma 6*

We will construct the vectors u, v using a procedure inspired by [74]. Let u be any fixed vector with mean zero, and with entries satisfying

$$-1/2 < u_1 < u_k < 1/2, \quad u_{i+1} > u_i + \frac{1}{4k}, \quad i = 1,\dots,k-1. \tag{G3}$$

Define the polynomial

$$f_u(z) = \prod_{i=1}^k (z - u_i), \quad z \in \mathbb{C}.$$

Now, consider the perturbed polynomial $f_v(z) = f_u(z) + (1/4k)^k$. In view of the separation condition (G3), Lemma 25 ensures that the polynomial f_v has k real roots $v_1,\dots,v_k$ contained in the interval $[-1,1]$. Since f_v is monic, it takes the form

$$f_v(z) = \prod_{i=1}^k (z - v_i).$$

Now, let us apply Vieta's formula (cf. Appendix K 2 a) to obtain

$$f_u(z) = z^k + \sum_{j=1}^k (-1)^j e_j(u_1,\dots,u_k) z^{k-j}, \quad f_v(z) = z^k + \sum_{j=1}^k (-1)^j e_j(v_1,\dots,v_k) z^{k-j},$$

where e_j denote the elementary symmetric polynomials. Since f_u and f_v only differ in their zeroth-order coefficient, we deduce that

$$e_j(u_1,\dots,u_k) = e_j(v_1,\dots,v_k), \quad j = 1,\dots,k-1.$$

By Newton's identities (equation (K3)), it follows from here that $\widetilde{m}_j(u) = \widetilde{m}_j(v)$ for all $j = 1,\dots,k-1$. Furthermore, by again using Newton's identities, we have for all $z \in \mathbb{C}$:

$$(1/4k)^k = f_v(z) - f_u(z)$$

$$
\begin{aligned}
&= (-1)^k \Big[e_k(v_1, \dots, v_k) - e_k(u_1, \dots, u_k) \Big] \\
&= (-1)^k \sum_{j=1}^{k} (-1)^{j-1} \Big[e_{k-j}(v_1, \dots, v_k) m_j(v) - e_{k-j}(u_1, \dots, u_k) m_j(u) \Big] \\
&= m_k(v) - m_k(u) \lesssim W(u, v),
\end{aligned}
$$

where the implicit constant depends on k. The claim readily follows from here. □

b. *Proof of Lemma 19*

Let $\theta = \langle c, \varpi \rangle$, and $Y \sim \mathrm{Poi}(n\theta)$. To prove the claim, it suffices to show that

$$\widehat{U}_\ell = \frac{Y!}{n^\ell (Y - \ell)!}$$

satisfies $\mathbb{E}[\widehat{U}_\ell] = \eta_\ell$, and $\mathrm{Var}[\widehat{U}_\ell] \lesssim \frac{1}{n^\ell d^\ell} + \frac{1}{d^{2\ell}}$. For any $\ell \geq 1$, one has

$$\mathbb{E}[\widehat{U}_\ell \,|\, \varpi] = \theta^\ell, \quad \mathrm{Var}[\widehat{U}_\ell \,|\, \varpi] \lesssim \frac{1}{n^{2\ell}} (n\theta)^\ell \big((n\theta + \ell)^\ell - (n\theta)^\ell \big),$$

by Lemma 24. To deduce the unconditional bound, notice that

$$\mathrm{Var}[\widehat{U}_\ell] = \mathbb{E}\big\{ \mathrm{Var}[T_{1,\ell} | \varpi] \big\} + \mathrm{Var}\left\{ \mathbb{E}[T_{1,\ell} | \varpi] \right\}.$$

The first term satisfies

$$
\begin{aligned}
\mathbb{E}\big\{ \mathrm{Var}[\widehat{U}_\ell | \varpi] \big\} &\lesssim \frac{1}{d^2 n^{2\ell}} \sum_{j=1}^{d} \mathbb{E}\Big[(n\theta)^\ell + (n\theta)^{2\ell - 1} \Big] \\
&\lesssim \frac{1}{n^{2\ell}} \Big((n/d)^\ell + (n/d)^{2\ell - 1} \Big) \lesssim \frac{1}{n^\ell d^\ell} + \frac{1}{n d^{2\ell - 1}},
\end{aligned}
$$

where we used elementary bounds on the moments of exponential random variables (cf. Lemma 26) whereas the second satisfies

$$\mathrm{Var}\left\{ \mathbb{E}[\widehat{U}_\ell | \varpi] \right\} \leq \mathrm{Var}[\theta^\ell] \lesssim \frac{1}{d^{2\ell}}.$$

The claim follows. □

c. *Proof of Lemma 8*

We begin by noting the following simple bound.

Lemma 21. *There exist constants $C, a > 0$ depending on γ such that*

$$\sup_{c \in \Delta_k} \mathrm{KL}(\mathbf{Q}_c^t \| \mathbf{Q}_c) \leq C \cdot e^{-d^a/2}.$$

The proof appears below. By Lemma 21, together with Pinsker's inequality and the tensorization property of the KL divergence, one has for all $c, \bar{c} \in \Delta_k$,

$$\mathrm{TV}(\mathbf{Q}_{\bar{c}}^{\otimes d}, \mathbf{Q}_c^{\otimes d}) \leq \mathrm{TV}\big(\mathbf{Q}_{\bar{c}}^{\otimes d}, (\mathbf{Q}_{\bar{c}}^t)^{\otimes d} \big) + \mathrm{TV}\big((\mathbf{Q}_{\bar{c}}^t)^{\otimes d}, (\mathbf{Q}_c^t)^{\otimes d} \big) + \mathrm{TV}\big((\mathbf{Q}_c^t)^{\otimes d}, \mathbf{Q}_c^{\otimes d} \big)$$

$$\lesssim \sqrt{d \cdot \mathrm{KL}(\mathbf{Q}_{\bar{c}}^t \,\|\, \mathbf{Q}_c^t)} + \sup_{\widetilde{c} \in \Delta_k} \sqrt{d \cdot \mathrm{KL}(\mathbf{Q}_{\widetilde{c}}^t \,\|\, \mathbf{Q}_{\bar{c}})}$$
$$\lesssim \sqrt{d \cdot \chi^2\big(\mathbf{Q}_{\bar{c}}^t, \mathbf{Q}_c^t\big)} + \sqrt{d} \cdot e^{-d^a/2}$$
$$\lesssim \sqrt{d \cdot \chi^2\big(\mathbf{Q}_{\bar{c}}^t, \mathbf{Q}_c^t\big)} + e^{-d^a/4}.$$

This proves the first claim. To prove the second claim, we make use of the following observation:

Lemma 22. *There exist constants $C_1, C_2 > 0$ depending on γ such that for all $c, \bar{c} \in \Delta_k$ satisfying $W(c, \bar{c}) = 0$,*

$$\chi^2(\mathbf{Q}_c \| \mathbf{Q}_{\bar{c}}) \le C_1\Big(H^2(\mathbf{Q}_c \| \mathbf{Q}_{\bar{c}}) + e^{-C_2 d^a}\Big).$$

The proof appears below. We deduce from Lemma 22 that

$$\begin{aligned}
&\mathrm{KL}(\mathbf{Q}_c^{\otimes d} \| \mathbf{Q}_{\bar{c}}^{\otimes d}) \\
&\quad\lesssim d \cdot \chi^2(\mathbf{Q}_c^{\otimes d} \| \mathbf{Q}_{\bar{c}}^{\otimes d}) \\
&\quad\lesssim d\Big(H^2(\mathbf{Q}_c \| \mathbf{Q}_{\bar{c}}) + e^{-C_2 d}\Big) \\
&\quad\lesssim d\Big(H^2(\mathbf{Q}_c \| \mathbf{Q}_c^t) + H^2(\mathbf{Q}_c^t \| \mathbf{Q}_{\bar{c}}^t) + H^2(\mathbf{Q}_{\bar{c}}^t \| \mathbf{Q}_{\bar{c}}) + e^{-C_2 d}\Big) \\
&\quad\lesssim d\Big(\mathrm{KL}(\mathbf{Q}_c \| \mathbf{Q}_c^t) + \chi^2(\mathbf{Q}_c^t \| \mathbf{Q}_{\bar{c}}^t) + \mathrm{KL}(\mathbf{Q}_{\bar{c}}^t \| \mathbf{Q}_{\bar{c}}) + e^{-C_2 d}\Big) \\
&\quad\lesssim d \cdot \chi^2(\mathbf{Q}_c^t \| \mathbf{Q}_{\bar{c}}^t) + e^{-C_2 d^a},
\end{aligned}$$

for a possibly smaller constant $C_2 > 0$, where we used Lemma 21 in the final inequality. The claim follows. □

It thus remains to prove Lemmas 21–22.

Proof of Lemma 21. By convexity and tensorization of the KL divergence, one has

$$\begin{aligned}
\mathrm{KL}\big(\mathbf{Q}_c^t \,\|\, \mathbf{Q}_c\big) \le\; & \mathbb{E}_{\varpi}\Big[\mathrm{KL}\Big(\mathrm{Poi}(n\langle \varpi^t, c\rangle) \,\Big\|\, \mathrm{Poi}(n\langle \varpi, c\rangle)\Big)\Big] \\
& + k \cdot \mathbb{E}_{\varpi}\Big[\mathrm{KL}(\mathrm{Po}(m\varpi_1^t) \,\|\, \mathrm{Poi}(m\varpi_1))\Big],
\end{aligned} \tag{G4}$$

where the means are taken over $\varpi \sim \mathcal{E}_d^{\otimes k}$. To bound the first term, notice that the inequality $|\langle c, \varpi - \varpi^t\rangle| \le \langle c, \varpi\rangle$ always holds, thus we may apply Lemma 23 to obtain

$$\begin{aligned}
\mathbb{E}_{\varpi}\Big[\mathrm{KL}\Big(\mathrm{Poi}(n\langle \varpi, c^t\rangle) \,\|\, \mathrm{Poi}(n\langle \varpi, c\rangle)\Big)\Big] &\le \mathbb{E}_{\varpi}\left[\frac{\langle c, \varpi - \varpi^t\rangle^2}{\langle c, \varpi\rangle}\right] \\
&\le \mathbb{E}_{\varpi}\left[\frac{\Big(\sum_{i=1}^k c_i \varpi_i \cdot I(\varpi_i > t)\Big)^2}{\sum_{i=1}^k c_i \varpi_i}\right] \\
&\le \mathbb{E}_{\varpi}\left[\sum_{i=1}^k c_i \varpi_i \,\Big|\, \max_i \varpi_i > t\right] \cdot \mathbb{P}\Big(\max_i \varpi_i > t\Big) \\
&\le \sum_{i=1}^k c_i \mathbb{E}_{\varpi}\left[\varpi_i \,|\, \varpi_i > t\right] \cdot \mathbb{P}\Big(\max_i \varpi_i > t\Big).
\end{aligned}$$

By the memoryless property of the exponential distribution, one has

$$\mathbb{E}[\varpi_i \,|\, \varpi_i > t] = t + \mathbb{E}[\varpi_i] = t + \frac{1}{d} \leq 2t.$$

We thus have,

$$\mathbb{E}_{\varpi}\Big[\mathrm{KL}\Big(\mathrm{Poi}(n\langle \varpi, c^t\rangle) \,\|\, \mathrm{Poi}(n\langle \varpi, c\rangle)\Big)\Big] \leq 2t \cdot \mathbb{P}\Big(\max_i \varpi_i > t\Big) \lesssim tk \cdot e^{-dt} \lesssim e^{-d^a/2},$$

where the final inequality holds for a sufficiently small constant $a = a(\gamma) > 0$ by definition of t, and using the fact that $k \leq d$. An analogous upper bound can be obtained on the second term in equation (G4), and the claim then follows. □

Proof of Lemma 22. Let $c, \bar{c}$ satisfy $W(c, \bar{c}) = 0$. Under this condition, we will begin by showing that the ratio of the densities $\mathbf{q}_c(x, y)$ and $\mathbf{q}_{\bar{c}}(x, y)$ is bounded from above and below by positive constants over a large range of values (x, y). Indeed, we have for all $(x, y) \in I$

$$\begin{aligned}
\mathbf{q}_c(x, y) &= \frac{1}{x!y!}\mathbb{E}\left[f(x; n\langle \varpi, c\rangle)\prod_{i=1}^k f(y_i; m\varpi_i)\right] \\
&= \frac{1}{x!y!}\sum_{j:|j|=x}\binom{x}{j}\prod_{i=1}^k \mathbb{E}\left[(n\varpi_i c_i)^{j_i} e^{-n\varpi_i c_i}(m\varpi_i)^{y_i} e^{-m\varpi_i}\right],
\end{aligned}$$

where the summation is taken over all $j = (j_1, \ldots, j_k) \in \mathbb{N}_0^k$ such that $\sum_i j_i = x$, and we write $\binom{x}{j} = x!/j_1! \ldots j_k!$. Thus,

$$\begin{aligned}
\mathbf{q}_c(x, y) &= \frac{1}{x!y!}\sum_{j:|j|=x}\binom{x}{j}\prod_{i=1}^k \int_0^\infty (nuc_i)^{j_i}(mu)^{y_i} d e^{-(nc_i+d+m)u} du \\
&= \frac{1}{x!y!}\sum_{j:|j|=x}\binom{x}{j}\prod_{i=1}^k (nc_i)^{j_i} m^{y_i} \frac{d}{nc_i + d + m}\int_0^\infty u^{j_i+y_i}(nc_i + d + m)e^{-(nc_i+d+m)u} du \\
&= \frac{1}{x!y!}\sum_{j:|j|=x}\binom{x}{j}\prod_{i=1}^k (nc_i)^{j_i} m^{y_i} \frac{d(j_i + y_i)!}{(nc_i + d + m)^{j_i+y_i+1}}.
\end{aligned}$$

Writing $\zeta(j, y) = \prod_{i=1}^k (y_i + j_i)!/j_i!$, we thus have

$$\mathbf{q}_c(x, y) = \frac{m^{|y|}}{y!}\sum_{j:|j|=x}\zeta(j, y)\prod_{i=1}^k \frac{d(nc_i)^{j_i}}{(nc_i + d + m)^{j_i+y_i+1}}.$$

On the one hand, this implies

$$\begin{aligned}
\mathbf{q}_c(x, y) &\leq \frac{m^{|y|}}{y!}\sum_{j:|j|=x}\zeta(j, y)\prod_{i=1}^k \frac{d(nc_i)^{j_i}}{(d + m)^{j_i+y_i+1}} \\
&\leq \frac{m^{|y|}}{y!}\frac{d^k}{(d + m)^{x+|y|+k}}\sum_{j:|j|=x}\zeta(j, y)\prod_{i=1}^k (nc_i)^{j_i} =: \varphi_c(x, y),
\end{aligned}$$

Notice that $\varphi_c(x, y)$ only depends on the sorted vector c. On the other hand,

$$\mathbf{q}_c(x, y) = \frac{m^{|y|}}{y!}\sum_{j:|j|=x}\chi(j, y)\prod_{i=1}^k \frac{d(nc_i)^{j_i}}{(d + m)^{j_i+y_i+1}}\left(1 + \frac{nc_i}{d + m}\right)^{-(j_i+y_i+1)}$$

$$\geq \left(1+\frac{n}{d}\right)^{-(x+|y|+k)} \varphi_c(x,y),$$

We thus have for all c,

$$\left(1+\frac{n}{d}\right)^{-(x+|y|+k)} \varphi(x,y) \leq \mathbf{q}_c(x,y) \leq \varphi_c(x,y).$$

Since $\varphi_c = \varphi_{\bar{c}}$ whenever $W(c,\bar{c}) = 0$ for $c, \bar{c} \in \Delta_k$, we have for all such $c, \bar{c}$ that

$$\left(1+\frac{n}{d}\right)^{-(x+|y|+k)} \leq \frac{\mathbf{q}_c(x,y)}{\mathbf{q}_{\bar{c}}(x,y)} \leq \left(1+\frac{n}{d}\right)^{x+|y|+k}.$$

This implies that

$$1/2 \leq \mathbf{q}_c(x,y)/\mathbf{q}_{\bar{c}}(x,y) \leq 2, \quad \text{for all } x+|y| \leq M := \frac{\log 2}{\log(1+n/d)} - k. \tag{G5}$$

Under condition **(S)**, we note that $M \leq C_0 d/n$. Now, we form the decomposition

$$\chi^2(\mathbf{Q}_c \| \mathbf{Q}_{\bar{c}}) = \underbrace{\sum_{\substack{(x,y)\in I \\ x+|y|\leq M}} \frac{\mathbf{q}_c^2(x,y)}{\mathbf{q}_{\bar{c}}(x,y)} - 1}_{A} + \underbrace{\sum_{\substack{(x,y)\in I \\ x+|y|> M}} \frac{\mathbf{q}_c^2(x,y)}{\mathbf{q}_{\bar{c}}(x,y)}}_{B}.$$

To bound A, notice first that for all $z > 0$,

$$\begin{aligned}\sum_{x+|y|>z} \mathbf{q}_c(x,y) &= \mathbb{P}(n\langle \varpi, c\rangle + m\|\varpi\|_1 \geq z) \\ &\leq \mathbb{P}((n+m)\|\varpi\|_1 \geq C_0 z) \leq k \exp\left(-\frac{dz}{k(n+m)}\right),\end{aligned} \tag{G6}$$

and in particular, since $M \asymp d/n$, we obtain

$$\sum_{x+|y|>M} \mathbf{q}_c(x,y) \lesssim \exp\left(-C_2 d\right). \tag{G7}$$

We thus have,

$$\begin{aligned} A &= \sum_{\substack{(x,y)\in I \\ x+|y|\leq M}} \frac{\mathbf{q}_c^2(x,y)}{\mathbf{q}_{\bar{c}}(x,y)} - 1 \\ &= \sum_{\substack{(x,y)\in I \\ x+|y|\leq M}} \frac{(\mathbf{q}_c^2(x,y) - \mathbf{q}_{\bar{c}}(x,y))^2}{\mathbf{q}_{\bar{c}}(x,y)} + 2 \sum_{\substack{(x,y)\in I \\ x+|y|> M}} \mathbf{q}_c(x,y) - \sum_{\substack{(x,y)\in I \\ x+|y|> M}} \mathbf{q}_{\bar{c}}(x,y) \\ &\lesssim \sum_{\substack{(x,y)\in I \\ x+|y|\leq M}} \left(\frac{\mathbf{q}_c^2(x,y) - \mathbf{q}_{\bar{c}}(x,y)}{\sqrt{\mathbf{q}_{\bar{c}}(x,y)} + \sqrt{\mathbf{q}_{\bar{c}}(x,y)}}\right)^2 + e^{-C_2 d}, \end{aligned}$$

where the final display follows from equations (G5)–(G7). We have thus shown that

$$A \lesssim H^2(\mathbf{Q}_c, \mathbf{Q}_{\bar{c}}) + e^{-C_2 d},$$

Furthermore,

$$\begin{aligned} B &= \sum_{\substack{(x,y)\in I\\ x+|y|>M}} \frac{\mathbf{q}_c^2(x,y)}{\mathbf{q}_{\bar{c}}(x,y)} \\ &\lesssim \sum_{\substack{(x,y)\in I\\ x+|y|>M}} \mathbf{q}_c(x,y)\left(1+\frac{n}{d}\right)^{x+|y|+k} \\ &\lesssim \sum_{\substack{(x,y)\in I\\ x+|y|>M}} ke^{-\frac{d}{k(n+m)}(x+|y|+k)}\left(1+\frac{n}{d}\right)^{x+|y|+k} \lesssim e^{-C_2 d}, \end{aligned}$$

for a possibly larger constant $C_2 > 0$. The claim follows from here. □

d. *Proof of Lemma 9*

The upper bound is clear. For the lower bound, notice that

$$\mathbb{E}[\langle c, \varpi^t\rangle] = \sum_{i=1}^k c_i \mathbb{E}[\varpi_i \wedge t] \geq \int_0^t x d e^{-dx} dx = \frac{1}{d}\big(1 - dte^{-dt}\big),$$

as desired. □

e. *Proof of Lemma 10*

We have,

$$\begin{aligned} &\mathbb{E}_\varpi\left[f(x;\widetilde{U}_c)\prod_{i=1}^k f(y_i;\widetilde{V}_i)\right] \\ &= \mathbb{E}_\varpi\left[\frac{e^{-(\widetilde{U}_c+\widetilde{V}_1+\cdots+\widetilde{V}_k)}(\widetilde{U}_c\widetilde{V}_1^{y_1}\ldots\widetilde{V}_k^{y_k})}{x!y_1!\ldots y_k!}\right] \\ &\geq \frac{e^{-(n+km)t}}{x!y_1!\ldots y_k!}\mathbb{E}_\varpi\left[\widetilde{U}_c^x\widetilde{V}_1^{y_1}\ldots\widetilde{V}_k^{y_k}\right] \\ &= \frac{e^{-(n+km)t}}{x!y_1!\ldots y_k!}\left(n^x m^{\sum_i y_i}\right)\mathbb{E}_\varpi\left[\left(\sum_{i=1}^k c_i(\varpi_i\wedge t)\right)^x\prod_{i=1}^k(\varpi_i\wedge t)^{y_i}\right] \\ &= \frac{e^{-(n+km)t}}{x!y_1!\ldots y_k!}\left(n^x m^{\sum_i y_i}\right)\mathbb{E}_\varpi\left[\sum_{\substack{0\leq j_1,\ldots,j_k\leq x\\ j_1+\cdots+j_k=x}}\binom{x}{j_1,\ldots,j_k}\prod_{i=1}^k c_i^{j_i}(\varpi_i\wedge t)^{j_i+y_i}\right] \\ &= \frac{e^{-(n+km)t}}{x!y_1!\ldots y_k!}\left(n^x m^{\sum_i y_i}\right)\sum_{\substack{0\leq j_1,\ldots,j_k\leq x\\ j_1+\cdots+j_k=x}}\binom{x}{j_1,\ldots,j_k}\prod_{i=1}^k c_i^{j_i}\mathbb{E}_\varpi\left[(\varpi_i\wedge t)^{j_i+y_i}\right] \\ &\geq \frac{e^{-(n+km)t}}{x!y_1!\ldots y_k!}\left(n^x m^{\sum_i y_i}\right)\sum_{\substack{0\leq j_1,\ldots,j_k\leq x\\ j_1+\cdots+j_k=x}}\binom{x}{j_1,\ldots,j_k}\prod_{i=1}^k c_i^{j_i}\lambda^{j_i+y_i} \end{aligned}$$

$$= \frac{e^{-(n+km)t}}{x!y_1!\dots y_k!}\left(n^x m^{\sum_i y_i}\right)\lambda^{x+\sum_i y_i}$$
$$= e^{-(n+km)t}\cdot f(x;n\lambda)\prod_{i=1}^k f(y_i;m\lambda).$$

The claim now follows from the fact that $nt \le (n/d)^{1-\gamma_0} \le 1$ and $mtk \le k(m/d)^{1-\gamma_0} = k(m/d)^{\frac{1}{1+\gamma}} \le 1$, by assumption on k. □

f. *Proof of Lemma 11*

The collection $\{\varphi_{\alpha,\beta}(\cdot;\boldsymbol{\lambda})\}_{\alpha,\beta}$ is dense in $L^2(g_{\boldsymbol{\lambda}})$, and satisfies the orthogonality property

$$\mathbb{E}_{(X,Y)\sim g_{\boldsymbol{\lambda}}}\left[\varphi_{\alpha,\beta}(X,Y;\boldsymbol{\lambda})\varphi_{\alpha',\beta'}(X,Y;\boldsymbol{\lambda})\right]$$
$$= \sum_{(x,y)\in I}\left(\varphi_\alpha(x;\lambda_0)\varphi_{\alpha'}(x;\lambda_0)f(x;\lambda_0)\right)\cdot\prod_{i=1}^k\left(\varphi_\beta(y_i;\lambda_i)\varphi_{\beta'}(y_i;\lambda_i)f(y_i;\lambda_i)\right)$$
$$= \left(\sum_{x=0}^\infty \varphi_\alpha(x;\lambda_0)\varphi_{\alpha'}(x;\lambda_0)f(x;\lambda_0)\right)\cdot\prod_{i=1}^k\sum_{y_i=0}^\infty\left(\varphi_\beta(y_i;\lambda_i)\varphi_{\beta'}(y_i;\lambda_i)f(y_i;\lambda_i)\right)$$
$$= \alpha!\lambda_0^\alpha I(\alpha=\alpha')\cdot\prod_{i=1}^k\beta_i!\lambda_i^{\beta_i}I(\beta_i=\beta_i'),$$

where we used the orthogonality of the univariate Charlier basis (cf. equation (K5)). We deduce that $\{\varphi_{\alpha,\beta}(\cdot;\boldsymbol{\lambda})\}_{\alpha,\beta}$ forms an orthogonal basis of $L^2(g_{\boldsymbol{\lambda}})$. To prove the second identity, recall from equation (K6) that the generating function of the Charlier polynomials with parameter λ_0 is given by $e^{-u_0}(1+u_0/\lambda)^x$ for all $u_0 > -\lambda_0$, thus we have

$$\frac{f_{\lambda_0+u_0}(x)}{f_{\lambda_0}(x)} = e^{-u_0}\left(1+\frac{u_0}{\lambda_0}\right)^x = \sum_{\ell=0}^\infty \varphi_\ell(x;\lambda_0)\frac{(u_0/\lambda_0)^\ell}{\ell!},\quad x=0,1,\dots.$$

Re-applying a similar identity, we obtain for all $(x,y)\in I$ and all $\mathbf{u}\in\mathbb{R}^{k+1}$ such that $u_j \ge -\lambda_j$, $j=0,\dots,k$,

$$\frac{g_{\boldsymbol{\lambda}+\mathbf{u}}(x,y)}{g_{\boldsymbol{\lambda}}(x,y)} = \left(\sum_{\alpha=0}^\infty\varphi_\alpha(x;\lambda_0)\frac{(u_0/\lambda_0)^\alpha}{\alpha!}\right)\prod_{i=1}^k\left(\sum_{\beta_i=0}^\infty\varphi_{\beta_i}(x_i;\lambda_i)\frac{(u_i/\lambda_i)^{\beta_i}}{\beta_i!}\right)$$
$$= \sum_{(\alpha,\beta)\in I}\varphi_{\alpha,\beta}(x,y;\boldsymbol{\lambda})\frac{(u_0/\lambda_0)^\alpha}{\alpha!}\prod_{i=1}^k\frac{(u_i/\lambda_i)^{\beta_i}}{\beta_i!}.$$

and the claim then follows. □

g. *Proof of Lemma 12*

Let $G_1,\dots,G_L \overset{iid}{\sim} \mathrm{Exp}(1)$ and let $G=\sum_{i=1}^L G_i \sim \mathrm{Gamma}(s,1)$. Let $X = d\sum_{i=1}^L X_i$. Notice that the lower tail of the rescaled Beta density f_{dX_1} is dominated by the exponential density f_{G_1}; indeed, one has for all $x\in[0,1/2]$,

$$f_{dX_1}(x) = \frac{d-1}{d}(1-x/d)^{d-2} \le e^{-x}(1-x/d)^{-2} \lesssim e^{-x} = f_{G_1}(x).$$

We thus have

$$\mathbb{E}\left[\left(d\sum_{i=1}^{L} X_i\right)^{-3/2}\right] \lesssim \mathbb{E}[1/G^{3/2}] + \mathbb{E}\left[X^{-3/2}\middle|\min_{i\in S_2} d\pi_{i1} > 1/2\right] \lesssim \mathbb{E}[1/G^{3/2}] + L^{-3/2}.$$

The remaining expectation can be computed in closed form as

$$\mathbb{E}[1/G^{3/2}] = \Gamma(L-3/2)/\Gamma(L) \lesssim L^{-3/2},$$

where Γ denotes the Gamma function. The claim follows. □

3. Proofs Deferred from Appendix E

a. *Proof of Lemma 17*

Notice first that

$$\mathbb{E}[B_i] = (d+1)\sum_{j=1}^{d}\left(\frac{2c_i}{d^2} + \sum_{r\neq i}\frac{c_r}{d^2}\right) = (d+1)d\left(\frac{c_i}{d^2} + \frac{1}{d^2}\right) = c_i + O(d^{-1}).$$

Second, notice that by equation (E6), one has with probability at least $1 - Ce^{-\sqrt{d}/C}$,

$$\left|\mathbb{E}[\widehat{c}_i|\Pi] - B_i\right| \leq (d+1)\sum_{j=1}^{d}\sum_{r=1}^{k} c_r \varpi_{ij}\varpi_{rj}\left|(1-S_r^{-1})(1-S_i^{-1})\right| \lesssim d\sum_{j=1}^{d}\sum_{r=1}^{k} c_r\varpi_{ij}\varpi_{rj}.$$

The claim now follows by re-applying the same argument as under equation (E7). □

Appendix H: Description of Synthetic Data Analysis in Section IV

In this appendix, we provide the details of our analysis on the synthetic data presented in the main text.

1. Time-dependent models

To generate the synthetic dataset that mimics possible increasing error rates in the real experiments, we consider a one-dimensional array of L qubits and construct a circuit of L layers. Each layer consists of a set of two-qubit random unitaries applied to neighboring qubits on all even bonds, and a set of following random unitaries on odd bonds. After each layer, we introduce single-qubit Pauli errors (X, Y, or Z) on every qubit. We study two types of error models:

1. **Experiment-mimicking model**: the error rate at each layer is drawn uniformly from $[0.25\epsilon, 0.75\epsilon]$, where ϵ increases linearly from 2.5×10^{-4} (first layer) to 10^{-3} (last layer).

2. **Null model**: ϵ is fixed at $\sim 6\times10^{-4}$.

For both error models, the many-body fidelity is $F \approx 0.5$.

In Regime A (with side information from classical simulation), we analyze the synthetic data using MLE, where each column of Π corresponds to one of the Pauli errors. From this estimator, we extract the error rate for each spacetime position (Fig. 2a) and the average error rate ϵ_{est} for each layer (Fig. 2b). Fitting ϵ_{est} as a linear function of depth, $\beta\cdot\text{depth}+\epsilon_0$, yields an error growth rate β. To validate the time-dependence in the error rate, we should test whether the extracted β significantly differs from zero. For this purpose, we simulate 500 instances of the null (time-independent) model and perform the above analysis, constructing a histogram of the extracted β values to determine confidence intervals and p-values (Fig. 2c). In regime B, we repeat the same analysis using variational EM.

2. Correlated error models

To generate the synthetic dataset that mimics spatially correlated errors possibly existing in the real experiments, we consider a 5×4 two-dimensional array of qubits and construct a five layer circuit. Each layer consists of four sets of two-qubit random unitaries, consecutively applied to neighboring qubits cycling among pairs in the four different orientations. After each layer, we introduce incoherent errors, which include all single-qubit Pauli errors as well as select correlated errors. All error rates are assumed to be the same across different layers. The single-site Pauli error rates are drawn from a uniform random distribution $[10^{-3}, 3 \times 10^{-3}]$. We consider two different models of correlated errors:

1. **Two-body correlated error**: correlated-XX errors that may exist for any pair of qubits. Here, we consider the situation where error rates are negligibly small except for one "bad" pair with error rate $\sim 10^{-3}$. Our goal is to identify this pair by applying our algorithm to synthetic data.

2. **Multi-body correlated error**: correlated multi-X errors can exist along any column or row of qubits. Here, we consider the situation where all these error rates are negligibly small except one "bad" row and one "bad" column with error rate $\sim 10^{-3}$. Our goal is to identify such a row and column from analyzing the synthetic data.

Error rates are chosen such that the many-body fidelity is $F \approx 0.5$ in both models. In Fig. 3, we focus on regime A (i.e. with classically computed π_i's) and use the MLE estimator.

Appendix I: Description of Real Data Analysis in Section V

In this Appendix, we describe the error model and the numerical methods used to analyze data from the experiment in Ref. [1].

1. Error Model

Our data analysis involves a model for the output probability distribution of the form

$$p_c(z|\Pi) = c_1\pi_1(z) + \sum_{i>1} c_i\pi_i(z) + c_{-1}1/d, \tag{I1}$$

where the index i contains information both about the error type and spacetime location.

As stated in Eq. (7), we generate the Π matrix from a physical model of a noisy quantum state, parameterized by noise coefficients we wish to learn. In general, these terms,

Error	Kraus op. $K_i^{(a)}$	Coef. $w_i^{(a)}$	Fid. contribution f_i.
State prep.	X	+1	0
1q dephasing	Z	+1	0
2q dephasing	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$	+1	+1/4
2q flip-flop	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	+1	+1/4
$1 \to 0$ readout error	$\|0\rangle\langle 1\|$ $\|1\rangle\langle 1\|$	+1 -1	-1/2
$0 \to 1$ readout error	$\|1\rangle\langle 0\|$ $\|0\rangle\langle 0\|$	+1 -1	-1/2
$1 \to 0, 1 \to 0$ double readout error	$\|00\rangle\langle 11\|$ $\|01\rangle\langle 11\|$ $\|10\rangle\langle 11\|$ $\|11\rangle\langle 11\|$	+1 -1 -1 +1	+1/4

Table S1. Error processes modeled in the analysis of RCS data from Ref. [1] (Fig. 4). The readout error sources have multiple terms $K_j^{(a)}$ with coefficients $w_j^{(a)}$ [Eq. (I2)], derived below. The last column indicates the fidelity contribution of the error source, necessary to obtain the many-body fidelity (App. I 2).

proportional to the unknown coefficient c_i, will be of the form

$$R_i(\rho) = \sum_a w_i^{(a)} K_i^{(a)} \rho K_i^{(a)\dagger}, \tag{I2}$$

where the sum over a indicates multiple terms which may be associated with the same error source. As examples, when the error corresponds to a unitary operator (e.g. a Pauli error), there is only one term, and K_i is said unitary. However, more complicated error channels such as asymmetric readout errors require multiple terms, e.g. $R^{\text{readout}}_{1\to 0,j}(\rho) = (|0\rangle\langle 1|)_j \rho (|1\rangle\langle 0|)_j - (|1\rangle\langle 1|)_j \rho (|1\rangle\langle 1|)_j$ describes the contribution of $1 \to 0$ readout errors on qubit j, see Table S1 and discussion below.

To benchmark a realistic error model, we include several classes of errors which have been reported in the literature. They are: 1. State preparation errors (simply as a bit-flip X_j on the initial state $|0\rangle^{\otimes N}$), 2. single-qubit errors, where for our analysis we focus simply on Pauli Z_j dephasing errors, and 3. two-qubit errors representing (a) dephasing on the 11 state and (b) flip-flop exchange between neighboring pairs of qubits. Finally, we include 4. asymmetric readout errors with different rates of $0 \to 1$ and $1 \to 0$ errors, as well as double $1 \to 0, 1 \to 0$ readout errors, which occur at non-negligible rates because of the larger $1 \to 0$ error rates. Each of these have error channels of the form Eq. I2, with parameters summarized in Table S1.

The two-qubit errors may arise from processes such as coupling to higher transmon levels [1, 23], which may appear as stochastic errors in the control angles ϕ and θ of the FSIM class of gates:

$$\text{FSIM}(\theta, \phi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta/2 & -i\sin\theta/2 & 0 \\ 0 & -i\sin\theta/2 & \cos\theta/2 & 0 \\ 0 & 0 & 0 & \exp(i\phi) \end{pmatrix} \tag{I3}$$

Integrating over Gaussian fluctuations of θ and ϕ gives a more complicated channel (of Lindblad form) proportional to the fluctuations $\Delta\theta^2, \Delta\phi^2$. However, in this work we do not assume a precise model for these two-qubit errors and we instead use a simpler unitary error channel (Table S1), taking these as representative of dephasing processes on the 11 state, or flip-flop between 01 and 10 states.

While symmetric readout errors can simply be modeled as Pauli X_j errors on qubits j, asymmetric readout errors, which capture the strongly biased readout errors reported in Ref. [1], are more involved. The simplest way to obtain the relevant operators for asymmetric readout is to linearize the amplitude damping channel (of strength γ, acting on qubit j)

$$\begin{aligned} R[\rho] &= \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}_j \rho \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}_j + \begin{pmatrix} 0 & \sqrt{\gamma} \\ 0 & 0 \end{pmatrix}_j \rho \begin{pmatrix} 0 & 0 \\ \sqrt{\gamma} & 0 \end{pmatrix}_j \\ &= \rho + \gamma \left[-|1\rangle\langle 1|_j \rho |1\rangle\langle 1|_j - \frac{1}{2}|0\rangle\langle 0|_j \rho |1\rangle\langle 1|_j - \frac{1}{2}|1\rangle\langle 1|_j \rho |0\rangle\langle 0|_j + |0\rangle\langle 1|_j \rho |1\rangle\langle 0|_j \right] + O(\gamma^2). \end{aligned} \tag{I4}$$

The middle two terms (proportional to γ) correspond to dephasing induced by amplitude damping and can be neglected for readout errors, since the system is immediately measured in the Z basis. However, if one wanted to model an amplitude damping channel in the middle of the circuit, all terms above should be kept). This gives the operators for $1 \to 0$ readout in Table S1.

The effect on the classical probability distribution can be understood as follows. Assume for simplicity that a $1 \to 0$ readout error occurs on the first bit. We write the distribution $\pi_1(z) = (\pi_1^{(0)}(z'), \pi_1^{(1)}(z'))$ in terms of the distributions on the substrings $z' = z_2 z_3 \cdots z_N$, conditioned on $z_1 = \{0, 1\}$. The readout error acts on the distribution as:

$$\pi_1(z) \mapsto \pi_1(z) + \gamma \left(\pi_1^{(1)}(z'), -\pi^{(1)}(z') \right) \equiv \pi_1(z) + \gamma \pi_{1\to 0,1}^{\text{readout}}(z), \tag{I5}$$

that is, it shifts probability mass from $z_1 = 1$ onto $z_1 = 0$. These terms precisely correspond to application of the operators in Table S1.

Double readout errors are modeled in a similar fashion: applying the amplitude damping channel with rates γ_i, γ_j on qubits i and j, and keeping terms proportional to $\gamma_i \gamma_j$, we obtain

$$\begin{aligned} \pi_1(z) \mapsto & \pi_1(z) + \gamma_i \pi_{1\to 0,i}^{\text{readout}}(z) + \gamma_j \pi_{1\to 0,j}^{\text{readout}}(z) \\ & + \gamma_i \gamma_j \left(\pi_1^{(11)}(z'), -\pi_1^{(11)}(z'), -\pi_1^{(11)}(z'), \pi_1^{(11)}(z') \right), \end{aligned} \tag{I6}$$

where the length-four vector now runs over $(z_i, z_j) \in \{00, 01, 10, 11\}$, and $z' = z \backslash (z_i, z_j)$. Surprisingly, the second order contribution *adds* probability mass to the 11 state.

Finally, in Eq.(I1) we additionally include a "white noise" term proportional to $1/d$, not assumed in our theoretical analysis, but which models the aggregate weight of errors outside of our model, which we expect to sum to such a featureless distribution [35]. Assuming a Markovian error model with a total rate of γ local error events per unit time, a simple estimate gives a many-body fidelity of $e^{-\gamma t}$, $\gamma t e^{-\gamma t}$ "single" error events, and $[(\gamma t)^2/2]e^{-\gamma t}$ "double" error events where two independent local errors occur. In this work, we neglect the vast majority of such double and higher-order errors, only including double readout events. Therefore, a simple estimate for the weight c_{-1} is $1 - \widehat{F} - \widehat{F}\log\left(1/\widehat{F}\right)$, where $\widehat{F}$ is the estimated many-body fidelity (see Appendix I 2). For the $N = 18$ dataset, this gives 0.41 for $\widehat{F} = 0.24$: we additionally estimate double readout errors to constitute 0.05 of the signal, leading to reasonable agreement with the estimated $\widehat{c}_{-1} = 0.32(1)$.

After obtaining $\pi_1(z)$ and $\pi_i(z)$ via classical simulation of the RCS circuits using the `Cirq` package, we construct a matrix Π with entries $\pi_{ij} := \pi_i(z_j)$. This matrix generally

has negative entries, even though $\Pi^\top c$ is guaranteed to have nonnegative entries for any physically sensible error vector c. We perform our fitting procedure under the modeling assumption that the bitstring histogram Y has entries drawn independently according to the distribution:

$$Y_j \sim \mathrm{Poi}(n\Pi_{\cdot j}^\top c), \quad j = 1, \ldots, d, \tag{I7}$$

for some $c \in \{x \in \mathbb{R}_+^k : \Pi_{\cdot j}^\top x \geq 0,\ j = 1, \ldots, d\}$. As we saw in Section C, this Poissonian model is statistically indistinguishable from the multinomial sampling model when the shot noise is large, the entries of Π are nonnegative, and $c \in \Delta_k$. In our more general setting here, where c may not lie in the simplex, the multinomial model is not well-defined, which is the reason we adopt the above more general Poissonian model. We use a Poisson MLE estimator to fit the c coefficients,

$$\widehat{c}^{\mathrm{MLE}} = \operatorname*{argmax}_{x \in \mathbb{R}_+^k} \sum_{j=1}^{d} \left(Y_j \log\left(\Pi_{\cdot j}^\top x\right) - \Pi_{\cdot j}^\top x \right), \tag{I8}$$

which should be contrasted to the multinomial MLE presented in equation (13) of the main text. In practice, we find the matrices Π to be poorly conditioned, and we add a ridge regularization penalty $10^{-8}\|x\|_2^2$ to the objective function (I8).

2. Converting learned error rates into physical quantities

a. *Many-body fidelity*

As alluded to in Table S1, we obtain the many-body fidelity estimate by a weighted sum of the learned c_i:

$$\widehat{F} = \widehat{c}_1 + \sum_{i>1} f_i \widehat{c}_i, \tag{I9}$$

where coefficients f_i for various sources of error are given in Table S1.

The reason why this is required is because the sources of error we consider need not result in output states orthogonal to the target output state, and hence output distributions π_i orthogonal to π_1.

For a Pauli error channel, with high probability in a RUC, the quantum state associated with each error trajectory has exponentially small overlap with the target quantum state and hence $f_i = 0$ in these cases. For a more general error channel, however, this is not the case. f_i can be computed by a simple analytical theory: one simply assumes the state is Haar random at the point the error is applied. For a single error, a Haar-average [139] reveals that the many-body fidelity is

$$f_i \approx \mathbb{E}_{\psi \sim \text{ Haar}}[\langle\psi| \sum_i w_i^{(a)} K_i^{(a)} |\psi\rangle\langle\psi| K_i^{(a)\dagger} |\psi\rangle] = \sum_i w_i^{(a)} \frac{|\mathrm{tr}(K_i^{(a)})|^2 + \mathrm{tr}(K_i^{(a)} K_i^{(a)\dagger})}{d(d+1)}, \tag{I10}$$

where the trace should be taken as over the entire $N-$qubit Hilbert space. For 2-qubit dephasing or flip-flop error (Table S1), $|\mathrm{tr}(K_i)|^2 = d^2/4$, while $\mathrm{tr}(K_i K_i^\dagger) = d$ (the second term is always sub-leading), leading to the coefficients $f_i = +1/4$. That is to say, acting with a controlled-Z or flip-flop unitary "error" produces a state which has, on-average, an fidelity of $1/4$ with the target state. This fidelity should be added back to the fidelity estimate $\widehat{F}$, in addition to $\widehat{c}_1$.

As another example, for asymmetric readout errors, one in fact has to use all the terms of the linearized amplitude damping channel Eq. (I4) (including the off-diagonal terms), this gives $f_i = -1/2$. As intuition, if there were *only* $1 \to 0$ readout errors, i.e. with the model

$$p(z) = c_1 \pi_1(z) + \sum_j c_j \pi^{\text{readout}}_{1\to 0,j}(z), \tag{I11}$$

our algorithm would learn a coefficient $\widehat{c}_1 = 1$, since for all the other terms $\sum_z \pi^{\text{readout}}_j(z) = 0$, but the sampled distribution is by definition normalized. However, the actual many-body fidelity is smaller, precisely $c_1 - \sum_j c_j/2$, with the factor of $1/2$ arising from the probability of the bit being in the 1 state. As an independent check, one can verify that the XEB fidelity between π_1 and $\pi^{\text{readout}}_{1\to 0,j}$ is $1/2$. A similar calculation yields $f_i = +1/4$ for double asymmetric readout errors: we summarize these results in Table S1.

The simple behavior of Pauli errors discussed above does not hold true near the start and end of the RUC. Near the start, the circuit depth is too low for the Haar-random assumption made above to hold, and a local error does not orthogonalize the state. Meanwhile, near the end of the RUC, a local operator *does* orthogonalize the state. However, a dephasing error does not sufficiently scramble before measurement in order to change the XEB: this is the "lag time" in the XEB that had been previously noted [22]. Both effects are evident in the correlation matrix $\Pi^T \Pi$ for 1q dephasing, 2q dephasing and 2q flip-flop errors: furthermore, these effects are largely confined to the first and last three layers. While the latter effect does not contribute to the many-body fidelity, we omit both these boundary circuit layers in order to cleanly test our fidelity coefficients f_j for errors deep in the circuit: doing so reveals close quantitative agreement between the XEB and our estimate $\widehat{F}$ [Eq. (I9)] in Fig. 4.

A more refined theory of these fidelity contributions f_j that incorporates the space-time positions of the errors would enable our method to include such boundary errors without comprimising the fidelity estimate.

b. *Correction of double readout errors on single readout error rates*

A similar effect happens between single and double readout errors: these have non-trivial overlaps, and after our fitting procedure, one must correct the estimate of the readout error rate on qubit j as

$$\widehat{c}^{\text{readout}}_{1\to 0,j} \mapsto \widehat{c}^{\text{readout}}_{1\to 0,j} - \frac{3}{14} \sum_{k\neq j} \widehat{c}^{\text{doub. readout}}_{(1\to 0)^2,jk} \tag{I12}$$

where the sum is taken over qubits $k \neq j$. The coefficient of $3/14$ comes from the following considerations: the quantum fidelity for mixed states (such as the contributions of single- and double- readout error) is less straightforward to analyze. Therefore, we use as a proxy a heuristic analysis based on the XEB: we seek to "orthogonalize" the Π matrix rows $\pi^{\text{readout}}_{1\to 0,i}$ and $\pi^{\text{doub. readout}}_{(1\to 0)^2,jk}$. We define orthogonalization with respect to the dot product:

$$\langle \pi_1, \pi_2 \rangle \equiv d \sum_z \pi_1(z)\pi_2(z) \tag{I13}$$

It is also convenient to subtract the identity component such that all vectors we consider sum to zero, that is work with $\pi_1 - 1/d$ instead of π_1. This has the feature that the XEB fidelity of a distribution $p(z)$ can be understood as the inner product $\langle p, \pi_1 - 1/d \rangle$.

This orthogonalization procedure correctly reproduces the fidelity contributions f_j: $\pi^{\text{readout}}_{1\to 0,i} + (\pi_1 - 1/d)/2$ and $\pi^{\text{doub. readout}}_{(1\to 0)^2,jk} - (\pi_1 - 1/d)/4$. Our estimation problem is equivalent

to fitting to a modified model:

$$p = 1/d + F(\pi_1 - 1/d) + \cdots + \sum_i c_{1\to 0,i}[\pi^{\text{readout}}_{1\to 0,i} + (\pi_1 - 1/d)/2] + \cdots , \tag{I14}$$

where the coefficient F is precisely the many-body fidelity (more precisely, this prescription ensures that the learned coefficient agrees with the XEB fidelity). To estimate the overlap between the double and single readout errors, we simply consider their inner product, which can be calculated to be (in our setting)

$$\langle \pi^{\text{readout}}_{1\to 0,i} + (\pi_1 - 1/d)/2, \pi^{\text{doub. readout}}_{(1\to 0)^2,jk} - (\pi_1 - 1/d)/4\rangle = \frac{3}{14} \text{ if } i = j \text{ or } i = k, \tag{I15}$$

As a reminder, we assume that the double readout errors cannot happen on the same qubit and therefore $j \neq k$. Orthogonalization the double readout term against the single readout term, and re-parametrizing the model gives the desired correction Eq. (I12), which we used in Fig. 4(e).

c. *Proportion of error sources*

Combining these results allows us to determine the proportions of each error source to the overall measurement, as plotted in Fig. 4(a).

Therefore, we assign their proportions as:

- Fidelity: estimated as in Eq. (I9).
- State preparation and 1q dephasing errors: no change to $\widehat{c}_i$.
- 2q dephasing and flip-flop errors: $(3/4)\widehat{c}_i$
- Single qubit readout errors: $(1/2)\widehat{c}^{\text{readout}}_{1\to 0,j} - (3/14)\sum_k \widehat{c}^{\text{doub. readout}}_{(1\to 0)^2,jk}$ [Eq. (I12)].
- Double qubit readout errors: contribution given by $(3/7 - 1/4)\widehat{c}^{\text{doub. readout}}_{(1\to 0)^2,jk}$.

In our problem, we have considered error sources where $\sum_z \pi_i(z) = 1$ or $\sum_z \pi_i(z) = 0$. The sum of the c_i's of the former type will be 1, in order for $p(z)$ to be normalized, while the c_i's of the latter type do not have such a constraint. One can verify that with the above prescription, the contributions over all error sources will sum to 1, as desired.

d. *Physical error rates*

Finally, we can construct estimators for the physical error rates Γ_i from the fitted coefficients c_i as well as the fidelity $\widehat{F}$. In our case, where errors correspond to the application of *only one* non-trivial Kraus operator, and under the assumption that the errors are independent, these are related by:

$$\widehat{\Gamma}_i = \frac{\widehat{c}_i}{\widehat{F} + \widehat{c}_i}. \tag{I16}$$

This relation arises as the coefficient c_i describes the probability of a specific, single event, which is the *product* of the physical error rates:

$$c_i \approx \Gamma_i \prod_{j\neq i}(1 - \Gamma_j) \approx \frac{\Gamma_i}{1 - \Gamma_i} F \ , \tag{I17}$$

where the second equality is because the many-body fidelity is given by $F \approx \prod_j (1 - \Gamma_j)$. Eq. (I16) is necessary to extract the physical error rates, as plotted in Fig. 4(c,d,e,f). In particular, this rescaling by $\widehat{F}$ is necessary for proper comparison between single- and double-readout error rates to detect correlated readout errors in Fig. 4(f).

3. Goodness-of-Fit

In this section, we conduct a goodness-of-fit analysis for model (I7) which we adopted in our real data analysis. Concretely, our goal is to test the null hypothesis

$$H_0 : Y \sim \overline{\mathbf{Q}}_c = \otimes_{j=1}^{d} \mathrm{Poi}(n \Pi_{\cdot j}^{\top} c), \quad \text{for some } c \in \mathbb{R}_+^k \text{ such that } \Pi_{\cdot j}^{\top} c \geq 0 \text{ for all } j.$$

We construct a heuristic test for this composite null hypothesis, using the following χ^2 statistic [140]:

$$\chi^2 = \sum_j \frac{(Y_j - n \Pi^{\top} \widehat{c}^{\mathrm{MLE}})^2}{n \Pi^{\top} \widehat{c}^{\mathrm{MLE}}}, \tag{I18}$$

where $\widehat{c}^{\mathrm{MLE}}$ is defined in equation (I8). We note that, under the null hypothesis, the typical magnitude of χ^2 is on the order of d.

We calibrate the χ^2 statistic heuristically, using the parametric bootstrap. Concretely, we compare the observed value of the χ^2 statistic, denoted χ^2_{obs}, to the distribution of χ^2 that would be expected if $\overline{\mathbf{Q}}_{\widehat{c}^{\mathrm{MLE}}}$ were the true data-generating distribution. We approximate this distribution by simulating $1,000$ synthetic datasets of $n = 500,000$ bitstring samples from this distribution. For each dataset, we refit the model coefficients c and compute the corresponding χ^2 values. This forms an empirical estimate of the sampling distribution of χ^2 when $Y \sim \overline{\mathbf{Q}}_{\widehat{c}^{\mathrm{MLE}}}$, which we use to compute the probability of observing a χ^2 value more extreme than χ^2_{obs}, under the hypothesis that the measurements are drawn from $\overline{\mathbf{Q}}_{\widehat{c}^{\mathrm{MLE}}}$.

For $N = 18$ and one random circuit instance, we obtain $\chi^2_{\mathrm{obs}} = 281,858$. Meanwhile, our simulated χ^2 distribution has mean $\mu = 261,818$ and standard deviation $\sigma = 770$. Our observed χ^2_{obs} is thus 26σ away from the mean, with p-value $< 10^{-3}$.

We compare this to a goodness-of-fit test using Google's two-component error model (1) as the null hypothesis—an analysis which was also conducted by [32]. We find an observed value $\chi^2_{\mathrm{obs}} = 290,342$, while the simulated χ^2 distribution has mean $\mu = 262,134$ and standard deviation $\sigma = 740$. The observed χ^2_{obs} value is 38σ away from the mean in this case, also with negligible p-value $< 10^{-3}$.

Although our analysis suggests considerable room for improvement in modeling the data of Ref. [1], our k-component model has a marked improvement of 12σ over the simplest two-component white-noise model. The overall goodness-of-fit is still poor, indicating that the dataset contains information about a host of error processes not currently in our model. Nevertheless, our fitting procedure turns out to be remarkably robust, yielding fitted error rates for quantities of interest that are comparable with all available estimates from alternative benchmarking methods (see Fig. 4). As in the spirit of cross-entropy benchmarking, we expect this robustness to be due to the fact that the remaining error sources not captured by our model lie in spaces that are roughly orthogonal to the row space of Π (up to centering).

Appendix J: Justifying the Independent Porter-Thomas Assumption

In this section, we provide some quantitative evidence for the validity of assumption **(PT)**, which we leveraged throughout our theoretical study. This assumption requires the various bitstring error distributions $\Pi_{i\cdot}$ to be mutually independent, and distributed according to

the Porter-Thomas law. In what follows, we will show that this assumption holds to second order: under mild assumptions on the errors of the circuit, we find that the rows of $\Pi_{i\cdot}$ are approximately uncorrelated, and have marginal moments which are consistent with the Porter-Thomas law.

Specifically, we shall explicitly calculate $\mathbb{E}[\Pi\Pi^T]$, where the rows $\Pi_{i\cdot}$ are probability distributions arising from the presence of a single error in a local brickwork random unitary circuit. For simplicity of analysis, we will assume here that our errors are single-qubit Pauli terms. Furthermore, we also consider errors that are within the bulk of circuit, so that we can replace the circuit before and after the signal with global Haar random unitaries R_1, R_2. More specifically, for diagonal terms, we take

$$\pi_0(z) = |\langle z|R_2 R_1|0\rangle|^2 \tag{J1}$$

$$\pi_i(z) = |\langle z|R_2 P_i R_1|0\rangle|^2, \tag{J2}$$

where P_i is a Pauli error and $R_1, R_2 \sim \text{Haar}(2^N)$ with N the system size. We treat the vectors $\pi_0(z)$ and $\pi_i(z)$ as forming the rows of Π.

Now, using Weingarten calculus [139, 141], we can compute

$$\mathbb{E}_{R_1,R_2}\left[\sum_z \pi_0(z)^2\right] = \frac{2}{d} + O\left(\frac{1}{d^2}\right), \tag{J3}$$

$$\mathbb{E}_{R_1,R_2}\left[\sum_z \pi_i(z)^2\right] = \frac{2}{d} + O\left(\frac{1}{d^2}\right). \tag{J4}$$

The leading order terms exactly match results for Porter-Thomas distributions, where $d = 2^N$. Thus, differences from Porter-Thomas are exponentially small.

For off-diagonal terms involving $\pi_0(z)$ and an error distribution $\pi_i(z)$, we can similarly use Eq. J1. This yields

$$\mathbb{E}_{R_1,R_2}\left[\sum_z \pi_i(z)\pi_0(z)\right] = \frac{1}{d} + O\left(\frac{1}{d^3}\right). \tag{J5}$$

The leading order term is again exact if assuming i.i.d. Porter-Thomas distributions.

Finally, for off-diagonal terms involving two different signals, we also consider the portion of the underlying brickwork circuit $\mathcal{R}$ that lies between the spacetime locations of these two signals. To be specific, suppose the signal P_i and P_j are t layers apart in the original circuit. Then, $\mathcal{R}$ would comprise of exactly these t layers of local random unitaries that separate the two signals. This leads to modified expressions of the form

$$\pi_i(z) = |\langle z|R_2 \mathcal{R} P_i R_1|0\rangle|^2, \tag{J6}$$

$$\pi_j(z) = |\langle z|R_2 P_j \mathcal{R} R_1|0\rangle|^2 \tag{J7}$$

for computing

$$\mathbb{E}_{R_1,R_2,\mathcal{R}}\left[\sum_z \pi_i(z)\pi_j(z)\right]. \tag{J8}$$

Now,

$$\mathbb{E}_{R_1,R_2}\left[\sum_z \pi_i(z)\pi_j(z)\right] = \frac{1 + d^{-2}\text{Tr}(P_i(t)P_j^*)\text{Tr}(P_i^*(t)P_j)}{d} + O\left(\frac{1}{d^2}\right), \tag{J9}$$

where we have defined

$$P_i(t) \equiv \mathcal{R}^\dagger P_i \mathcal{R}. \tag{J10}$$

The leading order term above differs only from the i.i.d. Porter-Thomas case by $d^{-2}\mathrm{Tr}(P_i(t)P_j^*)\mathrm{Tr}(P_i^*(t)P_j)$, where we have included d^{-2} to normalize the trace, which is $O(d^2)$.

To proceed, we can compute

$$\mathbb{E}_{\mathcal{R}}\left(d^{-2}\mathrm{Tr}(P_i(t)P_j^*)\mathrm{Tr}(P_i^*(t)P_j)\right), \tag{J11}$$

where the average is over individual two-qubit Haar random unitaries in the brickwork $\mathcal{R}$. This can be done through the standard technique of mapping the unitaries to an Ising spin model, as explained in Refs. [37, 142]. This yields

$$\mathbb{E}_{\mathcal{R}}\left(d^{-2}\mathrm{Tr}(P_i(t)P_j^*)\mathrm{Tr}(P_i^*(t)P_j)\right) = D(x,t), \tag{J12}$$

where x and t indicate how far P_j is from P_i in the space and time directions. Specifically, t is is equal to the depth of $\mathcal{R}$, and x represents how many qubits away the error P_j is from P_i. For the full expression of $D(x,t)$, see Ref. [45]—here, we only discuss relevant properties of the function. Specifically, $D(x,t)$ decays exponentially in both x and t. Thus, for signals P_i, P_j that are relatively spaced out in the random circuit, $D(x,t) \ll 1$, and the expectation $E_{R_1,R_2,\mathcal{R}} \sum_z \pi_i(z)\pi_j(z)$ also approximately satisfies the i.i.d. Porter-Thomas result of $\frac{1}{d}$.

In all cases, we see that the second moments of the rows of $\Pi_{i,\cdot}$ match those of i.i.d. Porter-Thomas distributions up to exponentially small corrections in the system size N and spacing of signals.

Appendix K: Further Technical Background

In this Appendix, we summarize several known technical results and definitions which are used throughout our proofs.

1. Technical Results

We begin by stating a few standard facts about Poisson distributions. The following is a standard upper bound on the Kullback-Leibler divergence between Poisson random variables.

Lemma 23. *For any $\mu, \nu > 0$, it holds that*

$$\mathrm{KL}\big(\mathrm{Poi}(\mu) \,\|\, \mathrm{Poi}(\nu)\big) = \mu \log \frac{\mu}{\nu} + \nu - \mu.$$

Furthermore, for all $C > 0$, there exists $K > 0$ such that if $|\nu - \mu| < C\nu$, then

$$\mathrm{KL}\big(\mathrm{Poi}(\mu) \,\|\, \mathrm{Poi}(\nu)\big) \leq K \cdot \frac{(\mu-\nu)^2}{\nu}.$$

The following is Lemma 2 of [143].

Lemma 24. *Let $X \sim \mathrm{Poi}(\lambda)$. For any integer $r \geq 1$, let*

$$\widehat{T} = X!/(X-r)!.$$

Then,

$$\mathbb{E}[\widehat{T}] = \lambda^r, \quad \mathrm{Var}[\widehat{T}] \leq \lambda^r\big((\lambda+r)^r - \lambda^r\big).$$

Next, we state a technical result about perturbations of polynomials, which is adapted from [74].

Lemma 25. *Let f be a polynomial of degree k with k real roots $x_1, \ldots, x_k \in \mathbb{R}$. Assume these roots are pairwise distinct, and let $\delta = \min_{i \neq j} |x_i - x_j|$. Then, for any $\epsilon < (\delta/2)^k$, the polynomial $f + \epsilon$ also has k real roots $x_1^\epsilon, \ldots, x_k^\epsilon \in \mathbb{R}$ satisfying*

$$W_1\left(\frac{1}{k}\sum_{i=1}^k \delta_{x_i}, \frac{1}{k}\sum_{i=1}^k \delta_{x_i^\epsilon}\right) \leq \delta/2.$$

The following Lemma collects several elementary facts about Dirichlet and exponential distributions which are used throughout our proofs.

Lemma 26. *Let $d \geq 2$, and let $\pi = (\pi_1, \ldots, \pi_d)^\top \sim \mathcal{D}_d$ be a flat Dirichlet-distributed random vector, and let $\varpi = (\varpi_1, \ldots, \varpi_d)^\top$ be a random vector consisting of i.i.d. Exp(d)-distributed random variables. Then, the following assertions hold.*

1. *$\pi_i \sim \mathrm{Beta}(1, d-1)$ for $i = 1, \ldots, d$.*
2. *$\sum_{i=1}^d \varpi_i \sim \mathrm{Gamma}(d, d)$.*
3. *$\pi \overset{d}{=} (\varpi_1, \ldots, \varpi_d)/\sum_{i=1}^d \varpi_i$.*
4. *If $G \sim \mathrm{Gamma}(d, d)$ is independent of π, then the vector $(G\pi_1, \ldots, G\pi_d)^\top$ consists of i.i.d. Exp(d) random variables.*
5. *If $G \sim \mathrm{Gamma}(\alpha, \lambda)$ with $\alpha \in \mathbb{N}$ and $\lambda > 0$, and $Y|G \sim \mathrm{Poi}(G)$, then the marginal law of Y is Negative Binomial with number of trials α and probability parameter $\lambda/(\lambda+1)$.*
6. *We have for all $\ell = 1, 2, \ldots$*

$$\mathbb{E}[\pi_1] = \mathbb{E}[\varpi_1] = \frac{1}{d}, \quad \mathrm{Var}[\pi_1] = \frac{1}{d^2}, \quad \mathbb{E}[\pi_1^\ell] = \frac{\ell!}{d^\ell}.$$

2. Classical Polynomial Families

We now recall the definitions and basic properties of three polynomial families which play an important role in our development.

a. Elementary Symmetric Polynomials

The *elementary symmetric polynomials* are a family of k-dimensional polynomials, defined for all $c_1, \ldots, c_k \in \mathbb{C}$ by

$$e_0(c_1, \ldots, c_k) := 1 \quad \text{and} \quad e_j(c_1, \ldots, c_k) := \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq k} \prod_{\ell=1}^j c_{i_\ell}. \tag{K1}$$

By *Vieta's formula*, the elementary symmetric polynomials can be used to describe the coefficients of a univariate monic polynomial $f(z) = \prod_{i=1}^{k}(z - c_i)$, with roots $c_1, \dots, c_k \in \mathbb{C}$. Concretely, one has:

$$f(z) = z^k + \sum_{j=1}^{k}(-1)^j e_j(c_1, \dots, c_k) z^{k-j}, \quad z \in \mathbb{C}. \tag{K2}$$

Since the polynomials e_j are symmetric, they admit an algebraic representation in terms of the moments $m_1(c), \dots, m_k(c)$ of the vector c, namely

$$m_j(c) = \frac{1}{k}\sum_{i=1}^{k} c_i^j, \quad j = 1, \dots, k.$$

This representation can be made explicit using *Newton's identities* which state that for any $\ell = 1, \dots, k$,

$$e_\ell(c_1, \dots, c_k) = \frac{k}{\ell}\sum_{j=1}^{\ell}(-1)^{j-1} e_{\ell-j}(c_1, \dots, c_k) m_j(c), \tag{K3}$$

In particular, equations (K2)–(K3) together imply that the vector c is uniquely determined, up to permutation of its entries, by the vector of moments $m(c) = (m_j(c) : 1 \le j \le k)$. That is, one has the following simple fact.

Lemma 27. *For any $c, c' \in \mathbb{C}^k$, it holds that*

$$m(c) = m(c') \implies \{c_1, \dots, c_k\} = \{c'_1, \dots, c'_k\}.$$

Some of our results will rely on a *quantitative* analogue of Lemma 27. Concretely, when constructing statistical estimators $\widehat{c}$ of c via moment estimation, we will be led to the question of quantifying the distance between $\widehat{c}$ and c in terms of their moment distance. It turns out that such quantitative bounds can be obtained by combining Newton's identities with existing perturbation bounds for polynomial roots, which were first developed by Refs. [144, 145]. This strategy was recently used by Hundrieser *et al.* [74], who proved the following result.

Lemma 28 ([74])**.** *There exists a constant $C = C(k) > 0$ such that for any $c, c' \in \mathbb{C}^k$,*

$$W(c, c') \le C\|m(c) - m(c')\|^{\frac{1}{k}}.$$

This Lemma shows that the sorted loss function W is $(1/k)$-Hölder continuous with respect to the ℓ_1 distance between moment vectors.

Hundrieser *et al.* [74] additionally showed that the Hölder exponent $1/k$ can be improved if the coordinates of the elements c, c' admit some separation. In order to state this refined result, recall the set Δ_{k,k_0} defined in Appendix B 3 a. Let $c^\star \in \Delta_{k,k_0}$ be given, and let $v_1 > \cdots > v_{k_0}$ denote its k_0 distinct entries. Define the Voronoi cells

$$V_0 = \emptyset, \quad V_\ell = \big\{z \in \mathbb{C}^k : \|z - v_i\| \le \|z - v_j\|, \forall i \ne j\big\} \setminus V_{\ell-1}, \quad \ell = 1, \dots, k_0.$$

Furthermore, let r_ℓ denote the multiplicity of v_ℓ among the entries of $c^\star$, for all $\ell = 1, \dots, k_0$. Given $c \in \Delta_k$, write $c_{V_\ell} = \{c_i \in V_\ell : 1 \le i \le k\}$. We then define, for all $c, c' \in \mathbb{C}^k$:

$$\overline{\mathcal{D}}_{c^\star}(c, c') = 1 \wedge \sum_{\ell=1}^{k_0} W^{r_\ell}(c_{V_\ell}, c'_{V_\ell}), \tag{K4}$$

with the convention that $W(c_{V_\ell}, c_{V_{\ell'}}) = \infty$ when $|c_{V_\ell}| \neq |c_{V_{\ell'}}|$. Finally, let

$$\delta(c^\star) = \min_{1 \le \ell < \ell' \le k} |v_\ell - v_{\ell'}|.$$

We then have the following refined stability bound.

Lemma 29 ([74]). *Let $1 \le k_0 \le k$ and $c^\star \in \Delta_{k,k_0}$. Then, there exists a constant $C = C(k, k_0, \delta(c^*)) > 0$ such that for any $c, c' \in \mathbb{C}^k$, we have*

$$\overline{\mathcal{D}}_{c^\star}(c, c') \le W^{k-k_0-1}(c, c') \le C \big\| m(c) - m(c') \big\|.$$

b. *Charlier Polynomials*

Let $f(x;\lambda) = e^{-\lambda}\lambda^x/x!$ denote the Poi(λ) density, evaluated at $x = 0, 1, \ldots$. The family of Charlier polynomials

$$\varphi_\ell(x;\lambda) := \sum_{r=0}^{\ell} (-1)^{\ell-r} \binom{\ell}{r} \frac{(x)_r}{\lambda^r}, \quad x, \ell = 0, 1, \ldots,$$

indexed by a parameter $\lambda > 0$, are a classical family of polynomials on $\mathbb{R}$ which are orthogonal with respect to the $L^2(\mathrm{Poi}(\lambda))$ norm [146]. One has the relation

$$\sum_{x=0}^{\infty} \varphi_\ell(x;\lambda)\varphi_{\ell'}(x;\lambda) = \ell!\lambda^\ell I(\ell = \ell'), \quad \ell, \ell' = 0, 1, \ldots \tag{K5}$$

The exponential generating function associated to the Charlier polynomials is

$$G(x,t) = \sum_{\ell=0}^{\infty} \varphi_\ell(x;\lambda)\frac{t^\ell}{\ell!} = e^{-t}\left(1 + \frac{t}{\lambda}\right)^x, \quad \text{for all } t \in \mathbb{R}. \tag{K6}$$

c. *Bell Polynomials*

Given an integer $p \ge 1$, the family of incomplete Bell polynomials $\{B_{\ell,p}\}_{\ell=1}^p$ consists of the set of polynomials on $\mathbb{R}^{p-\ell+1}$ defined by

$$B_{p,\ell}(\xi_1, \ldots, \xi_{p-\ell+1}) = p! \sum_{(h_1,\ldots,h_{p-\ell+1}) \in \mathcal{H}_{p,\ell}} \prod_{i=1}^{p-\ell+1} \frac{\xi_i^{h_i}}{(i!)^{h_i} h_i!}, \tag{K7}$$

for all $\xi_1, \ldots, \xi_{p-\ell+1} \in \mathbb{R}$. Here, $\mathcal{H}_{p,\ell}$ consists of all tuples $(h_1, \ldots, h_{p-\ell+1})$ of nonnegative integers such that

$$\sum_{i=1}^{p-\ell+1} h_i = \ell, \quad \sum_{i=1}^{p-\ell+1} i h_i = p.$$

Furthermore, the p-th complete Bell polynomial is defined by

$$B_p(\xi_1, \ldots, \xi_p) = \sum_{\ell=1}^{p} B_{p,\ell}(\xi_1, \ldots, \xi_{p-\ell+1}) = p! \sum_{r_1 + 2r_2 + \cdots + pr_p = p} \prod_{i=1}^{p} \frac{\xi_i^{jr_i}}{(i!)^{r_i} r_i!}.$$

One has the identity $|\mathcal{H}_{p,\ell}| = S(p,\ell)$, where

$$S(p,\ell) = \sum_{i=1}^{\ell} \frac{(-1)^{\ell-i} i^p}{(\ell-i)!\ell!} \leq \frac{\ell^p}{\ell!}. \tag{K8}$$

Furthermore, one has the basic identities

$$B_{p,\ell}(a,\ldots,a) = a^\ell S(p,\ell) \tag{K9}$$

$$B_{p,\ell}(a,a^2,\ldots,a^{p-\ell+1}) = a^p S(p,\ell) \tag{K10}$$

Given a random variable X with cumulants $\xi_p = \kappa_p(X)$ and moments $\eta_p = \mathbb{E}[X^p]$, for $p = 1,2,\ldots$, one has the relations

$$\begin{aligned} \xi_p &= \sum_{\ell=1}^{p} (-1)^{\ell-1}(\ell-1)! B_{p,\ell}(\eta_1,\ldots,\eta_{p-\ell+1}) \\ \eta_p &= \sum_{\ell=1}^{p} B_{p,\ell}(\kappa_1,\ldots,\kappa_{p-\ell+1}). \end{aligned} \tag{K11}$$