LIMITS OF QUANTUM GENERATIVE MODELS WITH CLASSICAL SAMPLING HARDNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling tasks have been successful in establishing quantum advantages both in theory and experiments. This has fueled the use of quantum computers for generative modeling to create samples following the probability distribution underlying a given dataset. In particular, the potential to build generative models on classically hard distributions would immediately preclude classical simulability, due to theoretical separations. In this work, we study quantum generative models from the perspective of output distributions, showing that models that anticoncentrate are not trainable on average, including those exhibiting quantum advantage. In contrast, models following sparse distributions can be trained. We consider special cases to enhance trainability, and observe that this opens the path for classical algorithms for surrogate sampling. This observed trade-off is linked to verification of quantum processes. We conclude that quantum advantage can still be found in generative models, although its source must be distinct from anticoncentration.

1 INTRODUCTION

Quantum computers are known to have extended capabilities with respect to their classical counterparts, which has been established through examples such as Shor factorization (Shor, 1997), Grover search (Grover, 1996), or experimental demonstrations of quantum advantages (Arute et al., 2019; Broome et al., 2013). Inspired by the success of classical Machine Learning (ML), quantum machine learning (QML) explores recipes to process data through quantum resources and surpass current computational capabilities. The presence of data, however, incorporates subtleties for showing formal separations, since simulating a computational process and learning from data are fundamentally different tasks (Schreiber et al., 2023; Gil-Fuster et al., 2025; Huang et al., 2021). Still, in certain cases, it is possible to show robust separations (Jerbi et al., 2024; Gyurik et al., 2024; Sweke et al., 2021; Coyle et al., 2020), e.g., by planting the seed of a quantumly easy and classically hard problem inside a learning algorithm (Liu et al., 2021). For that, strong assumptions on data are necessary. However, realistic data might be noisy, incomplete, imperfect or redundant, reducing the scope of applicability of these separations. The central difficulty thus lies in finding practically useful algorithms that perform better or more efficiently than classical counterparts.

Quantum computers suit a natural interpretation as generative models, i.e., ML models learning data distributions to subsequently sample from them. Sampling is inherent to quantum processes, since measurements are the only way to retrieve information from quantum states. In particular, the measurement process probabilistically generates samples based on the underlying state through Born's rule. This property is exploited in generative QML (GQML) models, such as parameterized quantum circuits, Boltzmann machines or Gibbs states (Barthe et al., 2025; Shen et al., 2025; Kurkin et al., 2025; Zoufal, 2021). We recall that sampling tasks were the first framework where quantum advantage was found, with Boson sampling (Aaronson & Arkhipov, 2011), Instantaneous Quantum Polynomial (IQP) (Bremner et al., 2010; 2016) circuits or random quantum circuits (RQC) (Boixo et al., 2018; Arute et al., 2019). Using these circuits as subroutines opens the potential for quantum-advantageous generative models (Kasture et al., 2023; Recio-Armengol & Bowles, 2025).

In this work, we show that classical hardness of circuits has its reflections in GQML. Shortly, classical hardness requires *anticoncentration*, that is, the underlying probability distribution is on average *almost but not exactly* uniform. As a consequence, most loss functions will strongly concentrate, preventing any training. The result shows reminiscence of barren plateaus (McClean et al., 2018;

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

Larocca et al., 2025; Cerezo et al., 2025). This observation is tightly related to verification of a quantum sampling task, and deeply rooted in the assumptions allowing for quantum advantage.

To show these results, we analytically study three abstractions of quantum circuits; (1) a circuit description corresponding to product states, (2) a family yielding pseudo-independent probability distributions, which we connect to quantumly-advantageous circuits, and (3) pseudo-independent sparse distributions, which are related in spirit to classically verifiable circuits with quantum advantage (Aaronson & Zhang, 2024; Zhang, 2025). We show that the first two families lead to exponential concentration of the square distance, although only pseudo-independent states can potentially show classical hardness. For the sparse case, loss functions do not concentrate, however, due to missing anticoncentration, classical sampling hardness is not established. We further provide numerical experiments to interpolate between analytical results and practically relevant circuits. We note that our findings do not incorporate hardness constraints, but rather an analysis of commonly found underlying distributions.

The concentration results we find have direct implications for GQML, targeting the subtle difference between sampling hardness and the utility of the model. The lack of trainability, implied by the on-average impossibility of distinguishing between the target function and the produced one, precludes the on-average usage of hard-to-sample generative models. While it is possible to find trainable instances, the sampling hardness guaranties are severely compromised in this case.

The paper is organized as follows. Section 2 provides relevant background on QGML and classical sampling hardness. Section 3 states our theoretical framework and proves concentration for loss functions. Section 4 numerically extends the theoretical statements. Connection to verification and implications for QML are detailed in Section 5 and Section 6. We conclude in Section 7.

2 Background

2.1 Quantum computers as generative models

Quantum computers are easily interpreted as discrete sampling machines, where the probability of sampling a possible outcome follows Born's rule. For a circuit C, this probability is

$$p_C(x) = |\langle x| C |0\rangle|^2, \quad x \in \{0,1\}^n. \tag{1}$$

Interestingly, the task of sampling bitstrings has been fertile in showing quantum advantages, both theoretically and experimentally, through, e.g., IQP (Bremner et al., 2016), RQC (Boixo et al., 2018; Aaronson & Chen, 2016; Arute et al., 2019) or Boson sampling (Aaronson & Arkhipov, 2011).

This perspective opens a path for using quantum computers as a generative ML model, where the goal is to learn the underlying distribution q of a dataset of samples (Benedetti et al., 2019; Amin et al., 2018; Lloyd & Weedbrook, 2018). Any generative model based on a sampling task with provable quantum advantage has the potential to inherit the classical hardness, thus, no classical algorithm could - even approximately - substitute the quantum machine. This has, in particular, been explored for IQP circuits (Kasture et al., 2023; Recio-Armengol et al., 2025; Kurkin et al., 2025). In this scenario, classical sampling hardness is, to the best of our knowledge, an open problem.

From a practical point of view, a G(Q)ML model must be trainable, that is, there must exist an efficient algorithm that finds an instance of the model that accurately mimics the training data (Gil-Fuster et al., 2025). The notion of trainability has been widely addressed in quantum variational models, for instance in the context of barren plateaus (BP). For a generative model, training requires estimating the difference between the current candidate and the target probability distribution, typically through a loss function. In a quantum setting, losses can be encoded into observables. If their expectation values can be efficiently estimated classically, training can be delegated to classical surrogates (Rudolph et al., 2024; Van Den Nest, 2011), thus avoiding pitfalls of current hardware.

2.2 Sampling hardness

As mentioned before, sampling has proven to be a fruitful ground for quantum advantage (Bremner et al., 2016; Boixo et al., 2018; Aaronson & Chen, 2016; Arute et al., 2019; Aaronson & Arkhipov, 2011). The hardness proofs all follow a common pattern, rooted in hard problems from complexity

theory. In particular, the quantum circuit must encode the solution of a #P-hard problem into the probability of sampling a given bitstring. By Stockmeyer's theorem (Stockmeyer, 1985), sampling from the circuit then allows a multiplicatively accurate estimation of said probability, and the collapse of the polynomial hierarchy via Toda's theorem (Toda, 1991). This last event is widely believed to be false under common assumptions in complexity theory, thus proving classical hardness[1].

Crucially, the #P-hard requirement previously considered is only compatible with estimating $p_C(x)$ to multiplicative precision. Further, sampling hardness in the considered examples is an average statement established through anticoncentration, which reads as (Dalzell et al., 2022)

$$2^{2n}\mathbb{E}\left[p_c(x)^2\right] \geq \beta' \qquad \forall x;\ \beta' > 1. \tag{2}$$

Intuitively speaking, anticoncentration means that the typical probability distribution of a quantum model is *almost but not exactly* uniformly distributed. Hence, with high probability, $p_C(x) \in \Theta(2^{-n})$. Under this assumption, additive precision,

$$|p_C(x) - \hat{p}_C(x)| < \epsilon, \quad \epsilon^{-1} \in \mathcal{O}\left(\text{poly}\left(n\right)\right) \tag{3}$$

becomes meaningless. In fact, the diagonal structure of IQP circuits allows for the existence of an efficient classical algorithm for estimating $p_C(x)$ to $1/\text{ poly}\left(n\right)$ additive precision (Van Den Nest, 2011). However, the estimated $p_C(x)$ does not solve the problem at the core. A paradigmatic example of such a distribution is the Porter-Thomas distribution (Porter & Thomas, 1956).

2.3 Connection of this paper with barren plateaus

Barren plateaus (BPs) are a concentration phenomenon extensively studied specially in the recent literature of variational quantum algorithms (Larocca et al., 2025; McClean et al., 2018; Larocca et al., 2022; Cerezo et al., 2025; Ragone et al., 2024). In this section, we attempt to provide the minimal possible description of properties of BPs relevant to this work. Originally, BPs were defined as a concentration property of loss functions of variational quantum algorithms (Peruzzo et al., 2014; Farhi et al., 2014). Let $U(\theta)$ be a unitary operator, and let $|\psi_0\rangle$ be an initial state, then BPs appear if

$$\text{Var}_\theta\left[\langle\psi_0| U^\dagger(\theta)\hat{H}U(\theta)\,|\psi_0\rangle\right] \in \exp\left(-\mathcal{O}(n)\right), \tag{4}$$

for n the number of qubits of the system. This definition also holds for the derivatives of the previous loss function, although equivalence of both cases has been established (Arrasmith et al., 2021). At the core of this result lies the existence of 2-designs, that is, ensembles of states matching the first two moments of randomly-distributed states. BPs have been useful to navigate the frontier between classical and quantum devices. In particular, Cerezo et al. (2025) showed that trainability of variational models is incompatible with quantum advantages relying on exponentially large spaces[2].

The reader familiar with the literature of barren plateaus might find resonances of this phenomenon in our work. The tension of trainability and classical tractability is common to both topics. However, some of the technicalities for barren plateaus are not met in this work and are left unused on purpose. As an example, one can consider IQP circuits, which by construction can never be 2-designs (Bonet-Monroig et al.) and have a polynomially large Lie-algebra (Larocca et al., 2022). In addition, training of generative models cannot be easily described by one observable.

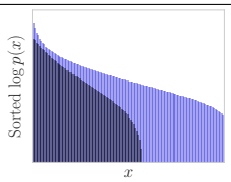
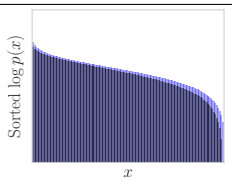
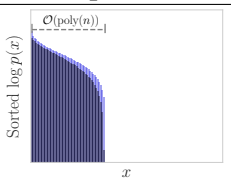
3 Concentration of loss functions

In this section, we model three relevant QGML architectures with different families of probability distributions and analyze statistical properties of their loss. Our results, summarized in Table 1, show that losses computed on anticoncentrated distributions will exponentially concentrate. This aims to illustrate that quantum models with sampling hardness rooted in anticoncentration will likely be untrainable.

[1] Several subtleties in these proves are hidden due to space and narrative constraints. We refer the interested reader to the original sources for an in-depth discussion.

[2] Again, many subtleties accompanying these results are left out for readability purposes. We recommend the interested reader to refer to the original source.

Table 1: Properties of the three considered distributions in this work. A typical example of each distribution lies within the blue area. μ and σ are element-wise mean and standard deviations of the sorted probability vectors $p_C(x)$. Classical sampling and estimation of $p_C(x)$ correspond to the existence of efficient classical algorithms to perform these tasks. Product circuits allow for classical sampling by (Bravyi et al., 2021) and estimation of $p_C(x)$ by Cirac et al. (2021). Simulation in both definitions for IQP (Bremner et al., 2010) and RQC (Boixo et al., 2018) is prohibited by (Bremner et al., 2016) and (Boixo et al., 2018) respectively. Peaked circuits (Aaronson & Zhang, 2024) do not allow for estimation of $p_C(x)$ (Zhang, 2025), however, classical sampling is not established.

Distribution	**Product**	**Pseudo-ind.**	**Peaked pseudo-ind.**
Avg behavior: Black: $\mu - \sigma$ Blue: $\mu + \sigma$			
Circuit	Product circuits	IQP, RQC	Peaked circuits
Anticonc.	×, prop. 1	✓, prop. 2	×, prop. 3
Loss conc.	✓, prop. 4	✓, prop. 6	×, prop. 8
Class. sampl.	✓	×	?, see sec. 3.1.3
Class. est. $p_C(x)$	✓	×	×

3.1 Three analytical families of distributions

3.1.1 Product distributions

The first distribution we consider is the *product distribution*, where we sample n uniformly-random real numbers $\boldsymbol{a} \in [0,1]^n$. Then, bitstrings are sampled bit by bit; 0 with probability a_i, 1 with probability $1 - a_i$. Evidently, we can access $p_{\boldsymbol{a}}(x)$ classically and sample from it. Formally,

Definition 1 (Product distribution)**.** *Let $\boldsymbol{a} \in [0,1]^n$ be uniformly distributed on its domain. A bitstring $x \in \{0,1\}^n$ is randomly distributed according to the product distribution if each bit is independently sampled from a binomial distribution with parameter a_i.*

This family corresponds to quantum circuits with a layer of single-qubit Haar-random unitaries (Diestel & Spalsbury, 2014), and also matrix product states (MPS) (Cirac et al., 2021) with bond dimension one, which admit classical sampling (Bravyi et al., 2021). The symmetry over x, allows showing a bias towards few bitstrings, becoming more prevalent with large n. Formally, the following establishes the concentration, with the proof provided in Appendix E.1.

Proposition 1 (Concentration of product distribution)**.** *Let $p_{\boldsymbol{a}}$ be a product distribution. This probability distribution concentrates, as*

$$\forall x: \quad \text{Prob}\left(p_{\boldsymbol{a}}(x) \geq \frac{y}{2^n}\right) \leq y^{-1/2} \exp\left(-\mathcal{O}(n)\right). \tag{5}$$

3.1.2 Pseudo-independent distributions

The second family of probability distribution is what we call *pseudo-independent distributions*. These arise from i.i.d. sampling positive random variables following some distribution, and normalizing by the sum. $p_C(x)$ then corresponds to the normalized random variables.

Definition 2 (Pseudo-independent distributions)**.** *Let $Y_1, Y_2, \ldots Y_N \in [0,\infty)$, be a set of $N = 2^n$ independent and identically distributed random variables. A bitstring $x \in \{0,1\}^n$ is randomly distributed according to a pseudo-independent distribution if*

$$p(x) = \left(\sum_{j=1}^{N} Y_j\right)^{-1} Y_x. \tag{6}$$

We call the family pseudo-independent because all correlations come from the normalization and decrease as N increases, which we establish more formally in Lemma 5 in Appendix E.2. The

pseudo-independent distributions are chosen as the representative of all quantum circuits whose output are *almost but not exactly* uniformly distributed, such as IQP and RQC. It is commonly known that RQC follow a Porter-Thomas distribution (Boixo et al., 2018), that is

$$\text{Prob}\left(p_C(x) \geq y\right) \approx \exp\left(-Ny\right), \tag{7}$$

which clearly anticoncentrates. A similar behavior can be observed for IQP circuits. In more detail, the Porter-Thomas distribution is the marginal of a Dirichlet distribution (Kotz et al., 2005), i.e., a Beta distribution (Bailey, 1992). Definition 2 is matched for this distribution, see Appendix B.2. Other similar models, such as Boson and Fermion sampling, are explored in Appendix C.

Very general assumptions suffice to establish anticoncentration. Since the correlations appear only through normalization, each $p_C(x)$ approximately follows the underlying independent probability distribution, with proper re-scaling. The proof is provided in Appendix E.2.

Proposition 2 (Anticoncentration of pseudo-independent distribution)**.** *Let p be an instance of the pseudo-independent distributions. This probability distribution anticoncentrates, as*

$$\forall x: \quad \text{Prob}\left(p(x) \geq \frac{y}{2^n}\right) \in \Omega(1). \tag{8}$$

3.1.3 Pseudo-independent peaked distributions

Lastly, we consider *pseudo-independent peaked distributions*. Those are similar to pseudo-independent distributions, but restricted to have sparse outcomes. We define them as follows.

Definition 3 (Pseudo-independent peaked distributions)**.** *Let $\bar{p}(j)$ be the underlying probability distribution of a set of outcomes, $j \in \{0, 1, \ldots, K\}$ as given in Definition 2. A bitstring $x \in \{0,1\}^n$ is randomly distributed according to the peaked pseudo-independent distribution if*

$$p(x) = \begin{cases} \bar{p}(j) & j \in S \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

for S, a random subset of $\{0, 1, \ldots, N\}$ of cardinality K, $K \leq N$.

The circuits are inspired by peaked quantum circuits (Aaronson & Zhang, 2024), where there must exist one x so that $p_C(x) \in \Omega(\text{poly}^{-1}(n))$. We adapted this approach to have circuits whose outcomes concentrate in a polynomial number of bitstrings. Peaked circuits were proposed as efficiently verifiable quantumly advantageous circuits, in contrast to anticoncentrated ones (Hangleiter et al., 2019). Still, estimating $p_C(x)$ remains classically hard (Zhang, 2025; Gharibyan et al., 2025). It is immediate that these distributions concentrate, which we formalize in the following.

Proposition 3 (Concentration of peaked pseudo-independent distribution)**.** *Let p be a peaked pseudo-independent distribution. This probability distribution concentrates, as*

$$\forall x: \quad \text{Prob}\left(p_{\boldsymbol{a}}(x) \geq \frac{y}{2^n}\right) \leq \frac{K}{N} \in \mathcal{O}\left(\frac{\text{poly}(n)}{2^n}\right). \tag{10}$$

The proof is a counting argument. Even in the assumption that $p_C(x) = 1$ for some C and x, the fraction of instances contributing to the bitstring x is only K/N.

3.2 Loss functions

In any training procedure, computing losses to a reasonable level of accuracy is a necessary requirement. This section, however, establishes that the most commonly used loss functions exhibit strong concentration property, prohibiting efficient and informative estimation. For that matter, we focus our analysis on the square distance (SD) $\Delta(p, q)$ and maximum mean discrepancy (MMD) (Gretton et al., 2012) loss, which we introduce in Appendix D. The first is a simple, yet effective measure for similarity and inherently connected to the definition of anticoncentration. The latter has been actively employed in GQML (Rudolph et al., 2024; Herrero-Gonzalez et al., 2025).

In the remainder of the section, we will provide bounds to the SD and MMD between p and the uniform distribution u. These results will be enough to show concentration of the real loss function by virtue of the triangle inequality, which equivalently applies for the MMD^2.

$$\left|\Delta(q,u)^{1/2} - \Delta(p,u)^{1/2}\right| \leq \Delta(p,q)^{1/2} \leq \Delta(q,u)^{1/2} + \Delta(p,u)^{1/2}. \tag{11}$$

3.2.1 PRODUCT DISTRIBUTION

We now return to the product distributions from Definition 1. Intuively, the SD is computing squares of differences of several quantities smaller than 1, thus, generally allowing large SDs. However, concentration implies that a very small fraction of $p_C(x)$ will be considerably larger than 2^{-n}. This fraction decreases exponentially with the number of qubits, translating into very small values of SD.

Proposition 4 (SD of product distribution)**.** *Let $p_{\boldsymbol{a}}$ be a product distribution depending on a random variable $\boldsymbol{a}$, and let u be the uniform distribution. Then, with probability at least $1-\delta$,*

$$\Delta(p_{\boldsymbol{a}}, u) \leq \frac{\exp\left(-\ln(3/2)n\right)}{\delta}. \tag{12}$$

For the proof, it suffices to apply a triangle inequality with the uniform distribution u, and then compute the second moment of $p_C(x)$, which decays exponentially. This is enough to invoke Chebychev's inequality and state this proposition.

The previous result states that a product distribution is not trainable using the SD as a loss function, however, it is by no means classically hard to sample from. There exists a compact representation of the circuit, implying a loss function sensitive to it. A loss function fulfilling this requirement is the MMD^2, which does not concentrate, as shown in the following, with the proof in Appendix E.

Proposition 5 (MMD of product distribution)**.** *Let $p_{\boldsymbol{a}}$ be a product distribution depending on a random variable $\boldsymbol{a}$, and let u be the uniform distribution. Consider a kernel k_ς with $\varsigma \in \mathcal{O}(n)$. Then, with probability $1-\mathcal{O}\left(\mathrm{poly}(1/n)\right)$,*

$$\mathrm{MMD}^2\left(p_{\boldsymbol{a}}, u\right) \in \Theta\left(\mathrm{poly}(1/n)\right). \tag{13}$$

3.2.2 PSEUDO-INDEPENDENT DISTRIBUTION

As seen in Section 3.1.2, pseudo-independent distributions anticoncentrate, which suffices to set trivial yet easy bounds on the values of the SD. In particular, for a given x

$$\mathbb{E}\left[(p(x)-u(x))^2\right] = \mathrm{Var}\left[p(x)-u(x)\right] \geq \frac{(\beta'-1)}{N^2} \implies \mathbb{E}\left[\Delta(p,u)\right] \in \Omega\left(\frac{1}{N}\right). \tag{14}$$

We can extend this into a full concentration inequality, with the proof in Appendix E.

Proposition 6 (SD of pseudo-independent distributions)**.** *Let p be an instance of the pseudo-independent distributions. Then, with probability at least $1-\Theta(\delta)$,*

$$\forall x: \quad \Delta(p,u) \leq \frac{\exp\left(-\mathcal{O}(n)\right)}{\delta}. \tag{15}$$

These results imply concentration of expectation values, i.e., for any diagonal Pauli string $\mathcal{Z}$,

$$\mathrm{Var}\left[\langle 0| C\mathcal{Z}C |0\rangle\right] \leq \mathcal{O}(\exp(-n)). \tag{16}$$

Further details are provided in Corollary 3 in Appendix E.2. To connect SD and MMD, consider the Fourier description of the distribution, see (44). By the Parseval-Plancherel identity (Plancherel & Leffler, 1910)

$$\sum_{k=0}^{n} \frac{1}{2^n} \sum_{\substack{S\subseteq[n] \\ |S|=k}} \left(\hat{P}(S) - \hat{Q}(S)\right)^2 = \Delta(p,q). \tag{17}$$

In particular, for $\varsigma \to 0$, the MMD transforms into the SD, admitting the interpretation that the MMD is a weighted average over different contributions, but on average equal to the SD. The implicit symmetry of the distributions implies that, on average, these quantities will be similar, and only strong biases on the target distribution or the generative model allow for a significant deviation. We can thus formulate an analogous result to Proposition 6, which is proven in Appendix E.2.

Proposition 7 (SD and Fourier-SD concentrate equally)**.** *With probability at least $1-\Theta(\delta)$*

$$\mathrm{MMD}^2\left(p,u\right) \leq \frac{\exp\left(-\mathcal{O}(n)\right)}{\delta}. \tag{18}$$

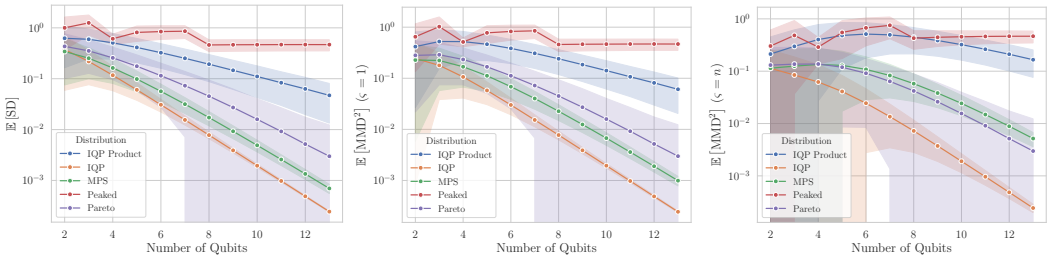


Figure 1: (left) Average of the SD. As expected, all the averages decay exponentially with the number of qubits, except for peaked distributions. The analytical bounds are satisfied in all the considered examples. For the Pareto distribution, exponential decay is also observable, even though the analytical bound is meaningless. (middle) Average of the MMD^2 with $\varsigma \in O(1)$. The behavior is similar to that of SD, with exponential decay in all cases, highlighting the randomness of the probability distributions of interest. (right) Average of MMD^2 with $\varsigma \in O(n)$. IQP product does not decay as strongly as other approaches. This behavior is due to the special properties of MMD^2, which is highly sensitive to low-order correlations. Those are dominant for product circuits.

3.2.3 Peaked distribution

The peaked pseudo-independent distributions offer clear advantages. The SD measures distances between each x by adding all $p(x) - q(x)$. If the distributions are K-sparse, with $K \ll N$, two typical distributions will have supports with no or very little overlap. Hence, the number of bitstrings of interest will grow as $\mathcal{O}(K)$, which leads to a much softer upper bound on the averages of the SD. We formulate it as follows, with the proof in Appendix E.3.

Proposition 8 (SD of peaked pseudo-independent distributions)**.** *Let $p_{\boldsymbol{a}}, q_{\boldsymbol{b}}$ be two instances of the peaked pseudo-independent distributions. Then, with probability at least $1 - \Theta(\delta)$,*

$$\forall x: \quad \Delta(p_{\boldsymbol{a}}, q_{\boldsymbol{b}}) \leq \frac{1}{K\delta}. \tag{19}$$

4 Numerical exploration

In this section, we delve into numerical extensions of the considered models to extend the scope of the analytically tractable examples to practical instances of circuits. In our numerical results, we compute the averages of SD and MMD^2 of different distributions for number of qubits $n \in [2, \ldots, 13]$ over 10^5 pairs of randomly generated distributions. We consider the following distributions. We consider Product IQP circuits, which are IQPs with only single-qubit operations, and MPS with bond dimension $\chi = n$, as a practical extension of product states with low entanglement. Further, we consider IQP circuits with two-qubit entangling gates and uniformly generated angles, as the typical quantumly advantageous instance. As our bounds for the pseudo-independent distributions get vacuous for undefined variance, we employ them with an underlying Pareto distribution with $\alpha = 2$, to study the behaviour where analytical guarantees fail. Finally, we design our peaked circuits by considering IQP circuits on $\log(n)$ qubits and distribute them in the n-qubit Hilbert space, resulting in $K \in \Theta(n)$. For further details, we refer to Appendix I.

4.1 Results

We first analyze the behavior of the SD for the considered distributions in the left of Figure 1. As expected, the IQP and IQP-product states show an exponential decay in the expectation value of the SD. Further, the MPS distribution serves as an intermediate step between IQP and IQP-product. For the case of peaked circuits, the scaling follows the expected trend, with artificial steps due to the rounding of $\log(n)$. The Pareto distribution we consider does not yield any behavior significantly different from the other considered ones. In spite of the inexistence of a bound in $\mathbb{E}\left[\Delta(p, q)\right]$, the numerical experiments show a similar exponential behavior as IQP or IQP product. Interestingly,

the IQP circuits do not only show an exponential decay on the average, but also strong concentration properties as the number of qubit increases, in agreement with the pseudo-independence assumption.

In alignment with previous works (Rudolph et al., 2024), we explore the MMD^2 for $\varsigma \in O(1)$ in the middle, and for $\varsigma \in O(n)$ on the right of Figure 1. For $\varsigma = 1$, we observe similar trends as for the MSD. Their average behavior is identical, thus, MMD^2 will on average not be trainable. For $\varsigma = n$, there is a change for IQP product and MPS states. As discussed in Section 3.2, our MMD^2 is sensitive to low-order correlations, present in states with low entanglement. It hints that these models might be easier to train than IQP, but also classical methods exists for sampling.

5 VERIFICATION

Broadly, trainability means finding a candidate distribution p that best approximates q, thus, a bare minimum for training is distinguishing p from q, a task that is known as *verification*. Usually, a full description of a q is given, together with sampling access to p, with the goal of correctly signaling $\|p-q\|_1 \geq \epsilon$ with probability $\geq 2/3$,[3], where $\|\cdot\|_1$ is the Total Variation Distance (TVD). The minimal sample complexity to distinguish two distributions is tightly related to the norm of its truncated probability vector (Valiant & Valiant, 2017). For quantum distributions with classical sampling hardness, anticoncentration is a key ingredient for advantages, but also an insurmountable obstacle for verification (Hangleiter et al., 2019). In short, exponentially many samples are needed due to the output distributions being *almost but not exactly* uniformly distributed.

To bridge the gap between training and verification, we need to circumvent the need for a complete description of q, which is typically available in the form of samples in any ML setting. However, for any distance, the triangle inequality implies upper and lower bounds by comparing to the uniform distribution u. Therefore, independent verification of the distributions p and q suffices to bound the distances between them, which is needed for training, with constant overhead. Notice that $\|p-u\|$ (resp. q) is related to the computation of $\mathbb{E}\left[p(x)^2\right]$, and thus connected to anticoncentration.

To complete the connection, we can consider the TVD as a loss function. In particular, $\|p-q\|_1 \leq \sqrt{N\Delta(p,q)}$, which could open the possibility of using the TVD as a loss function for generative models due to its exponentially larger values. This approach, however, does not seem to solve any core issue, as the concentration on $\Delta(p,q)$ previously seen imposes only constant bounds in TVD. From a more numerical perspective, while $\mathbb{E}\left[\|p-q\|_1\right] \in \mathcal{O}(1)$, the variance vanishes exponentially, as $\mathrm{Var}\left[\|p-q\|_1\right] \in \exp\left(\mathcal{O}(-n)\right)$ (Hangleiter et al., 2019), as detailed in Lemma 11. We also provide a numerical exploration of the TVD for our experiments in Appendix J.

To avoid these problems, we can look into circuits whose outputs can be verified, such as peaked circuits (Zhang, 2025; Gharibyan et al., 2025), which to the best of our knowledge, however, are hard to find (Aaronson & Zhang, 2024). Note that peaked circuits do not admit estimation of $p(x)$ to multiplicative precision in general. Otherwise, Shor's algorithm, an example of a peaked circuit, could be dequantized to factorize numbers with classical machines. Regarding sampling, classical algorithms that produce samples from the distribution of any peaked circuit can exist, as peaked circuits concentrate their outputs in a small support that can be stored in memory. However, as estimating $p_C(x)$ is hard, no classical recipe to find this classical sampling surrogate seems possible.

5.1 DISTINGUISHING DISTRIBUTIONS WITH MMD^2

Finally, it is possible to show that similar constraints to the ones found in verification apply to estimating a loss function and distinguishing between distributions, which the MMD^2 can be used for (Gretton et al., 2012). While the true MMD^2 is not accessible, it is possible to compute an unbiased estimator using l and m samples from p and q respectively. This, in turn, allows us to devise a statistical test for equality, which will signal as *true* as long as $\widehat{\mathrm{MMD}}^2 \leq (m+l)^{-1/2}$. As we saw, the MMD^2 vanishes exponentially, and then the necessary sampling budget to distinguish between p and q must increase accordingly. This is the typical scenario for a distribution probability showing anticoncentration. We refer to Appendix G for further details on these results.

[3] As in many examples, this success probability only needs to be bounded away from $1/2$, and it can be boosted to arbitrarily close to 1 using repetition arguments.

6 IMPLICATIONS FOR QML

Our results have obvious implications for GQML. Certain models with quantum advantage, such as IQP or RQC, are not trainable, and suffer from the same problems as many other variational models (McClean et al., 2018; Cerezo et al., 2025). The model only becomes useful if it exhibits a strong bias, allowing efficient and informative training. Otherwise, training will be seriously compromised, as most of the possible output distributions are *almost but not exactly* uniform.

In contrast, one could consider a q which is an outlier of the generative model, hence $\Delta(q,u) \gg N^{-1}$, implying that q is easily distinguishable from u (the uniform distribution). Still, $\Delta(p,q)$ inherits the concentration on $\Delta(p,u)$, even though the average is increased to $\approx \Delta(q,u)$, see (11).

Biased models have been suggested (Recio-Armengol et al., 2025; Rudolph et al., 2024), where data from probability distributions with sparse Fourier spectrum could be trained using MMD^2. This nicely aligns with the Gaussian kernel, and the strong bias will make these models trainable. In fact,

$$\varsigma = n \to \mathbb{E}\left[\Delta(p,q)\right] \geq 2e^{-1/6} \tag{20}$$

We highlight, however, that these instances do not belong to the average case of models with quantum advantage rooted in anticoncentration, as they are classically representable (cf. Section 3.1.1).

Finally, ML has lately investigated the *simplicity bias* (Shah et al., 2020) as a feature that allows for generalization. This bias is necessary to ensure that a small training dataset is representative enough to allow for learning the entire distribution. Probability distributions exhibiting anticoncentration do not comply with this simplicity bias. In light of our results, it is impossible to use $\mathrm{poly}\,(n)$ many samples to learn the underlying distribution, assuming a typical output from a generative model with anticoncentration, that is $p_C(x) \approx 2^{-n}$. Thus, sampling will only be useful if the underlying probability distribution is sufficiently peaked, that is, concentrated, at the expense of losing the sampling hardness that motivates the use of such a quantum model. In that spirit, Raj et al. (2026) showed[4] that moment-matching of low-order terms does not guaranty any generalization. MMD^2 is just a weighted average of these quantities, thus no generalization can be *a priori* expected.

7 CONCLUSION

In this paper, we have analyzed the statistical properties of typical loss functions of quantum generative models, in particular those exhibiting anticoncentration and quantum advantage. To this end, we gave three models of families of probability distribution, and studied their statistical properties and the behaviour of typcial loss functions. Our analytical results are complemented with numerical experiments to bridge the gap for a broader set of quantum circuits.

The immediate consequence is that quantumly advantageous generative models rooted in anticoncentration will not be usable in practice. The reason for our observation lies at the foundation of this kind of quantum advantage, where multiplicative errors on approximating certain exponentially-vanishing quantities are required. As a result, the typical outcome of such a model will be *close but not exactly* uniformly distributed, which questions the applicability of such models.

Our results prevent the use of such circuits for GQML, unfortunately leading to the loss of a candidate for advantageous generative models. However, there are other possible sources for advantages that survive our analysis. A clear candidate are peaked circuits, whose distributions strongly concentrate on small supports whose probability densities $p_C(x)$ are known to not be classical estimable. Nevertheless, the properties of peaked circuits might make their output distributions classically learnable from data. Another possible candidate for useful and, potentially, trainable models are models targeting datasets with a strong bias. However, several examples throughout the manuscript show that keeping the sampling advantage while biasing the model is a non-trivial task.

The results here presented encourage us to reflect on the applicability of GQML. Classical hardness as stated in anticoncentrated models necessarily imply trainability issues. However, interesting generated data should obey a specific purpose, and not be approximately uniform. In line with this discussion, generative models should be addressed from an angle complementary to classical hardness from multiplicative errors.

[4] In a work output after the first version of this manuscript.

AI USE STATEMENT

In this work, we used generative AI tools for

- implementing methods.

We have not used generative AI tools for

- helping in the development of theoretical models or conceptual frameworks
- formulating mathematical claims
- providing critical ingredient for proving mathematical claims
- assisting in the writing of proofs
- proposing or refining hypotheses
- designing or providing feedback on research methodology or experiments
- assisting with translation
- supporting qualitative and thematic data analysis
- interpreting results,

and generation of synthetic data sets and clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools for creating and modifying scientific figures, creating and editing of software code, brainstorming and editing the paper to improve readability. We have reviewed all AI-assisted work by manually verifying the source code for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code and results for the experiments are available in the following anonymized GitHub repository (`https://anonymous.4open.science/r/sampling-hardness-anonym-6979/`) and will be made public upon acceptance. The theoretical results in the main section are complemented with rigorous explanations and proofs in the appendix.

REFERENCES

Scott Aaronson and Alex Arkhipov. The computational complexity of linear optics. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 333–342, 2011.

Scott Aaronson and Lijie Chen. Complexity-Theoretic Foundations of Quantum Supremacy Experiments, 2016.

Scott Aaronson and Yuxuan Zhang. On verifiable quantum advantage with peaked circuit sampling, 2024. URL `https://arxiv.org/abs/2404.14493`.

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

Mohammad H. Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko. Quantum boltzmann machine. *Phys. Rev. X*, 8:021050, May 2018. doi: 10.1103/PhysRevX.8.021050. URL `https://link.aps.org/doi/10.1103/PhysRevX.8.021050`.

George Arfken, Hans-Jurgen Weber, and Frank E. Harris. *Mathematical Methods for Physicists: A Comprehensive Guide*. Elsevier, 7th ed edition, 2013. ISBN 978-0-12-384654-9.

Andrew Arrasmith, M. Cerezo, Piotr Czarnik, Lukasz Cincio, and Patrick J. Coles. Effect of barren plateaus on gradient-free optimization. *Quantum*, 5:558, 2021. ISSN 2521-327X. doi: 10.22331/q-2021-10-05-558.

Frank Arute, Kunal Arya, Ryan Babbush, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, October 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1666-5. URL http://dx.doi.org/10.1038/s41586-019-1666-5.

Ralph W. Bailey. Distributional Identities of Beta and Chi-Squared Variates: A Geometrical Interpretation. *The American Statistician*, 46(2):117–120, 1992. ISSN 0003-1305. doi: 10.2307/2684178.

A. Barthe, M. Grossi, S. Vallecorsa, et al. Parameterized quantum circuits as universal generative models for continuous multivariate distributions. *npj Quantum Inf*, 2025. URL https://doi.org/10.1038/s41534-025-01064-3.

Marcello Benedetti, Delfina Garcia-Pintos, Oscar Perdomo, Vicente Leyton-Ortega, Yunseong Nam, and Alejandro Perdomo-Ortiz. A generative modeling approach for benchmarking and training shallow quantum circuits. *npj Quantum Information*, 5(1), May 2019. ISSN 2056-6387. doi: 10.1038/s41534-019-0157-8. URL http://dx.doi.org/10.1038/s41534-019-0157-8.

Ville Bergholm, Josh Izaac, Maria Schuld, et al. Pennylane: Automatic differentiation of hybrid quantum-classical computations, 2022. URL https://arxiv.org/abs/1811.04968.

S. Boixo, S.V. Isakov, V.N. Smelyanskiy, et al. Characterizing quantum supremacy in near-term devices. *Nature Phys*, pp. 595–600, 2018. doi: 10.1038/s41567-018-0124-x.

Xavier Bonet-Monroig, Hao Wang, and Adrián Pérez-Salinas. Hierarchically discriminating Haar-randomness in quantum states from a black-box device. 28(7):074512. ISSN 1367-2630. doi: 10.1088/1367-2630/ae868b.

A. Bouland, B. Fefferman, C. Nirkhe, and U. Vazirani. On the complexity and verification of quantum random circuit sampling. *Nature Phys*, pp. 159–163, 2019. doi: 10.1038/s41567-018-0318-2.

Paolo Braccia, N. L. Diaz, Martin Larocca, M. Cerezo, and Diego García-Martín. Optimal Haar random fermionic linear optics circuits, 2025.

Sergey Bravyi, David Gosset, and Ramis Movassagh. Classical algorithms for quantum mean values. *Nature Physics*, 17(3):337–341, 2021. ISSN 1745-2473, 1745-2481. doi: 10.1038/s41567-020-01109-8.

Michael J. Bremner, Richard Jozsa, and Dan J. Shepherd. Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2126):459–472, August 2010. ISSN 1471-2946. doi: 10.1098/rspa.2010.0301. URL http://dx.doi.org/10.1098/rspa.2010.0301.

Michael J Bremner, Ashley Montanaro, and Dan J Shepherd. Average-case complexity versus approximate simulation of commuting quantum computations. *Physical review letters*, 117(8):080501, 2016.

Matthew A. Broome, Alessandro Fedrizzi, Saleh Rahimi-Keshari, Justin Dove, Scott Aaronson, Timothy C. Ralph, and Andrew G. White. Photonic boson sampling in a tunable circuit. *Science*, 339(6121):794–798, February 2013. ISSN 1095-9203. doi: 10.1126/science.1231440. URL http://dx.doi.org/10.1126/science.1231440.

M. Cerezo, Martin Larocca, Diego García-Martín, et al. Does provable absence of barren plateaus imply classical simulability? *Nature Communications*, 16(1):7907, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-63099-6.

Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling, 2023.

J. Ignacio Cirac, David Pérez-García, Norbert Schuch, and Frank Verstraete. Matrix product states and projected entangled pair states: Concepts, symmetries, theorems. *Reviews of Modern Physics*, 93(4):045003, 2021. doi: 10.1103/RevModPhys.93.045003.

B. Coyle, D. Mills, V. Danos, et al. The born supremacy: quantum advantage and training of an ising born machine. *npj Quantum Inf*, 2020. doi: 10.1038/s41534-020-00288-9.

Alexander M Dalzell, Nicholas Hunter-Jones, and Fernando GSL Brandão. Random quantum circuits anticoncentrate in log depth. *PRX Quantum*, 3(1):010333, 2022.

Joe Diestel and Angela Spalsbury. *The Joys of Haar Measure*. Number volume 150 in Graduate Studies in Mathematics. American Mathematical Society, Providence, Rhode Island, 2014. ISBN 978-1-4704-0935-7.

Adam Ehrenberg, Joseph T. Iosue, Abhinav Deshpande, Dominik Hangleiter, and Alexey V. Gorshkov. Transition of Anticoncentration in Gaussian Boson Sampling. *Physical Review Letters*, 134(14):140601, 2025. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.134.140601.

Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm, 2014. URL https://arxiv.org/abs/1411.4028.

Andrew J. Ferris and Guifre Vidal. Perfect sampling with unitary tensor networks. *Physical Review B*, 85(16):165146, 2012. doi: 10.1103/PhysRevB.85.165146.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/3a0772443a0739141292a5429b952fe6-Paper.pdf.

Hrant Gharibyan, Mohammed Zuhair Mullath, Nicholas E. Sherman, Vincent P. Su, Hayk Tepanyan, and Yuxuan Zhang. Heuristic Quantum Advantage with Peaked Circuits, 2025.

Elies Gil-Fuster, Casper Gyurik, Adrian Perez-Salinas, and Vedran Dunjko. On the relation between trainability and dequantization of variational quantum learning models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Representation Learning*, volume 2025, pp. 24069–24093, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3c215225324f9988858602dc92219615-Paper-Conference.pdf.

Johnnie Gray. quimb: a python library for quantum information and many-body calculations. *Journal of Open Source Software*, 3(29):819, 2018. doi: 10.21105/joss.00819.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Lov K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pp. 212–219, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917855. doi: 10.1145/237814.237866. URL https://doi.org/10.1145/237814.237866.

Casper Gyurik, Alexander Schmidhuber, Robbie King, Vedran Dunjko, and Ryu Hayakawa. Quantum computing and persistence in topological data analysis, 2024.

Dominik Hangleiter, Martin Kliesch, Jens Eisert, and Christian Gogolin. Sample complexity of device-independently certified "quantum supremacy". *Phys. Rev. Lett.*, 122:210502, May 2019. doi: 10.1103/PhysRevLett.122.210502. URL https://link.aps.org/doi/10.1103/PhysRevLett.122.210502.

Mario Herrero-Gonzalez, Brian Coyle, Kieran McDowall, Ross Grassie, Sjoerd Beentjes, Ava Khamseh, and Elham Kashefi. The Born Ultimatum: Conditions for Classical Surrogation of Quantum Generative Models with Correlators, November 2025.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 0162-1459. doi: 10.1080/01621459.1963.10500830.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1):2631, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22539-9.

S. Jerbi, C. Gyurik, S.C. Marshall, R. Molteni, and V. Dunjko. Shadows of quantum machine learning. *Nat Commun 15*, 2024. doi: 10.1038/s41467-024-49877-8.

Sachin Kasture, Oleksandr Kyriienko, and Vincent E. Elfving. Protocols for classically training quantum generative models on probability distributions. *Phys. Rev. A*, 108:042406, Oct 2023. doi: 10.1103/PhysRevA.108.042406. URL https://link.aps.org/doi/10.1103/PhysRevA.108.042406.

S. Kotz, N. Balakrishnan, and N.L. Johnson. *Continuous Multivariate Distributions, Models and Applications: Second Edition*. John wiley & sons, 01 2005. doi: 10.1002/9780471722069.

Andrii Kurkin, Kevin Shen, Susanne Pielawa, Hao Wang, and Vedran Dunjko. Universality and kernel-adaptive training for classically trained, quantum-deployed generative models, 2025. URL https://arxiv.org/abs/2510.08476.

M. Larocca, S. Thanasilp, S. Wang, et al. Barren plateaus in variational quantum computing. *Nat Rev Phy*, pp. 174–189, 2025. doi: 10.1038/s42254-025-00813-9.

Martin Larocca, Piotr Czarnik, Kunal Sharma, Gopikrishnan Muraleedharan, Patrick J. Coles, and M. Cerezo. Diagnosing Barren Plateaus with Tools from Quantum Optimal Control. *Quantum*, 6:824, 2022. ISSN 2521-327X. doi: 10.22331/q-2022-09-29-824.

Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 17(9):1013–1017, September 2021. ISSN 1745-2481. doi: 10.1038/s41567-021-01287-z.

Seth Lloyd and Christian Weedbrook. Quantum generative adversarial learning. *Phys. Rev. Lett.*, 121:040502, Jul 2018. doi: 10.1103/PhysRevLett.121.040502. URL https://link.aps.org/doi/10.1103/PhysRevLett.121.040502.

Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9(1), November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07090-4. URL http://dx.doi.org/10.1038/s41467-018-07090-4.

Colin McDiarmid. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

Ingram Olkin and Herman Rubin. Multivariate Beta Distributions and Independence Properties of the Wishart Distribution. *The Annals of Mathematical Statistics*, 35(1):261–269, 1964. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177703748.

Michał Oszmaniec, Ninnat Dangniam, Mauro E. S. Morales, and Zoltán Zimborás. Fermion Sampling: A robust quantum computational advantage scheme using fermionic linear optics and magic input states. *PRX Quantum*, 3(2):020328, 2022. ISSN 2691-3399. doi: 10.1103/PRXQuantum.3.020328.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

R. E. A. C. Paley and A. Zygmund. A note on analytic functions in the unit circle. *Math. Proc. Camb. Phil. Soc.*, 28(3):266–272, 1932. ISSN 0305-0041, 1469-8064. doi: 10.1017/S0305004100010112.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1):4213, 2014. ISSN 2041-1723. doi: 10.1038/ncomms5213.

Michel Plancherel and Mittag Leffler. Contribution à l'étude de la représentation d'une fonction arbitraire par des intégrales définies. *Rendiconti del Circolo Matematico di Palermo*, 30(1):289–335, 1910. ISSN 0009-725X, 1973-4409. doi: 10.1007/BF03014877.

C. E. Porter and R. G. Thomas. Fluctuations of Nuclear Reaction Widths. *Physical Review*, 104(2): 483–491, 1956. doi: 10.1103/PhysRev.104.483.

Michael Ragone, Bojko N. Bakalov, Frédéric Sauvage, Alexander F. Kemper, Carlos Ortiz Marrero, Martin Larocca, and M. Cerezo. A lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nat Commun*, 15, 7172, 2024. doi: 10.1038/s41467-024-49909-3.

Snehal Raj, Natansh Mathur, and Alejandro Perdomo-Ortiz. "Train classical, deploy quantum" requires rethinking generalization, September 2026.

Erik Recio-Armengol and Joseph Bowles. Iqpopt: Fast optimization of instantaneous quantum polynomial circuits in jax, 2025. URL https://arxiv.org/abs/2501.04776.

Erik Recio-Armengol, Shahnawaz Ahmed, and Joseph Bowles. Train on classical, deploy on quantum: scaling generative quantum machine learning to a thousand qubits, 2025. URL https://arxiv.org/abs/2503.02934.

M.S. Rudolph, S. Lerch, S. Thanasilp, et al. Trainability barriers and opportunities in quantum generative modeling. *npj Quantum Inf*, 2024. URL https://doi.org/10.1038/s41534-024-00902-0.

Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *Physical Review Letters*, 131(10):100803, September 2023. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.131.100803.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

K. Shen, A. Kurkin, A. Pérez-Salinas, et al. Variational quantum generative modeling by sampling expectation values of tunable observables. *npj Quantum Inf*, 2025. URL https://doi.org/10.1038/s41534-025-01121-x.

Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5):1484–1509, October 1997. ISSN 1095-7111. doi: 10.1137/s0097539795293172. URL http://dx.doi.org/10.1137/S0097539795293172.

Larry Stockmeyer. On approximation algorithms for # p. *SIAM Journal on Computing*, 14(4): 849–861, 1985. doi: 10.1137/0214060. URL https://doi.org/10.1137/0214060.

Ryan Sweke, Jean-Pierre Seifert, Dominik Hangleiter, and Jens Eisert. On the Quantum versus Classical Learnability of Discrete Distributions. *Quantum*, 5:417, March 2021. ISSN 2521-327X. doi: 10.22331/q-2021-03-23-417. URL https://doi.org/10.22331/q-2021-03-23-417.

Seinosuke Toda. Pp is as hard as the polynomial-time hierarchy. *SIAM Journal on Computing*, 20 (5):865–877, 1991. doi: 10.1137/0220053. URL https://doi.org/10.1137/0220053.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. doi: 10.1137/151002526. URL https://doi.org/10.1137/151002526.

Maarten Van Den Nest. Simulating quantum computers with probabilistic methods. *Quantum Info. Comput.*, 11(9–10):784–812, September 2011. ISSN 1533-7146.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. URL https://doi.org/10.1038/s41592-019-0686-2.

Yuxuan Zhang. Complexity and hardness of random peaked circuits, 2025. URL https://arxiv.org/abs/2510.00132.

Christa Zoufal. *Generative Quantum Machine Learning*. PhD thesis, ETH Zürich, 2021. URL https://arxiv.org/abs/2111.12738.

A PRELIMINARIES

We begin with several definitions, which are used in subsequent proofs.

Definition 4 (Gamma function (Abramowitz & Stegun, 1965))**.** *The Gamma function is defined as*

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{21}$$

for any real $z > 0$. For positive integers, it takes the form

$$\Gamma(z) = (z-1)! \tag{22}$$

Definition 5 (Incomplete lower Gamma function (Abramowitz & Stegun, 1965))**.** *The incomplete lower Gamma function is defined as*

$$\gamma(a, x) = \int_0^x e^{-t} t^{a-1} dt \tag{23}$$

with $a \in \mathbb{C}, \mathrm{Re}(a) > 0$.

Definition 6 (Gamma distribution)**.** *The probability density function of a Gamma distribution with shape $\alpha > 0$ and scale $\theta > 0$ is defined as*

$$f_{\alpha,\theta}(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}. \tag{24}$$

Its cumulative distribution is proportional to the incomplete lower Gamma function.

Definition 7 (Dirichlet distribution (Kotz et al., 2005))**.** *An N-dimensional random variable $\boldsymbol{x}$ follows a Dirichlet distribution if it follows a probability density function given by*

$$f(\boldsymbol{x}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_{i=1}^N \Gamma(\alpha_i)} \prod_{i=1}^N x_i^{\alpha_i - 1}, \tag{25}$$

for $\alpha_i \geq 0$. The marginal probabilities follow Beta distributions with $\alpha_0 = \sum_{i=N} \alpha_i$.

Definition 8 (Beta distribution (Bailey, 1992))**.** *A random variable x follows a Beta distribution if it follows a probability density function given by*

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{26}$$

where $\Gamma(\cdot)$ is the Gamma function. Beta distributions are the marginals of Dirichlet distributions.

Definition 9 (Symmetric Dirichlet distribution)**.** *A Dirichlet distribution is symmetric if $\boldsymbol{\alpha} = \alpha \mathbf{1}$.*

Definition 10 (Poisson distribution)**.** *A random variable X is said to follow a Poisson distribution with $\lambda > 0$ if its probability mass function is given by*

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{27}$$

Definition 11 (Pareto distribution)**.** *A random variable X is said to follow a Pareto distribution with $\alpha > 1$ if its probability mass function is given by*

$$\Pr(X = p) = \frac{\alpha}{(1+x)^{\alpha+1}}. \tag{28}$$

B CONNECTION BETWEEN PSEUDO-INDEPENDENT DISTRIBUTIONS AND HAAR-RANDOM STATES

B.1 CONNECTION BETWEEN DIRICHLET DISTRIBUTIONS AND HAAR-RANDOM STATES

It is a known fact that the probabilities $p(x)$ of measuring a bitstring x after preparing a Haar-random state follow a symmetric Dirichlet distribution (Boixo et al., 2018; Olkin & Rubin, 1964), with $\alpha = 1$. This implies that the marginal probability of sampling each x follows a probability distribution given by

$$\text{Prob}\left(p(x) = y\right) = N(1-y)^{N-2}, \tag{29}$$

which is a Beta distribution (Bailey, 1992). The Porter-Thomas distribution commonly discussed in the supremacy experiments is just an approximation of this distribution

$$\text{Prob}\left(p_C(x) \geq y\right) = N\int_y^1 (1-t)^{N-1}dt = (1-y)^N \lesssim \exp\left(-Ny\right). \tag{30}$$

Notice that the Porter-Thomas distribution is by definition incorrect since y is unbounded. However, it still returns accurate descriptions.

It is important to give relevance to the marginals of the distributions with respect to the distribution itself. It is somewhat accepted in the folklore that a Porter-Thomas shape is the signature for *quantumness* (Boixo et al., 2018), however, sampling a constant number of elements from any distribution admitting an analytical description is efficient (Bouland et al., 2019). We remark that this behavior does not suffice to identify the probability distribution as sourced from a quantum device. In fact, sampled elements from marginals do not even constitute a probability distribution because they are not constrained to sum up to one. We conjecture that the true source of *quantumness* is not visible in the marginals, but in the correlations of the sampling probabilities. These correlations must be exponentially small, and thus only visible when exponentially many samples are obtained. Along these lines, recent works on lower bounds for the query complexity of distributions might open interesting directions to shed light on classical sampling methods (Chewi et al., 2023).

B.2 CONNECTION BETWEEN PSEUDO-INDEPENDENT AND DIRICHLET DISTRIBUTIONS

Theorem 1. *Drawing X_i i.i.d. according to a Gamma distribution with shape α and rate 1 from Definition 6 results in the pseudo-independent random variables distributed according to a symmetric Dirichlet distribution from Definition 9.*

Proof. We consider i.i.d. random variables drawn from a Gamma distribution with shape α and rate 1, as $X_1, \ldots, X_N \sim^{i.i.d.} \Gamma(\alpha, 1)$ with their densities given as in Definition 6. We consider the normalized vector

$$Y_i = \frac{X_i}{\sum_i X_i}, \tag{31}$$

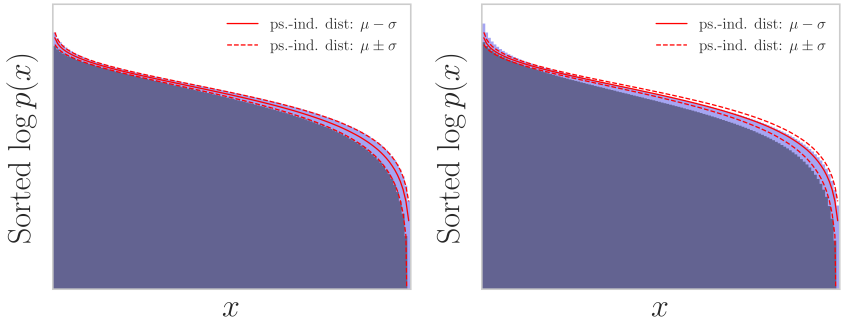
and let $S = \sum_i X_i$. The joint probability distribution of the random variables is given as

$$p(X_1, \ldots, X_N) = \prod_i p(X_i) = \prod_i \frac{1}{\Gamma(\alpha)} X_i^{\alpha-1} e^{-X_i} \tag{32}$$

We now shift the perspective and consider $X_i = Y_i S$. For $i = N$, we instead look at $X_N = S(1 - \sum_{i\in[1,N-1]} Y_i)$. We shift from $p(X_1, \ldots, X_N)$ to $p(Y_1 S, \ldots, Y_{N-1}S, (\sum_{i\in[1,N-1]} Y_i)S$, and consider the Jacobian as

$$J = \frac{\partial(X_1, \ldots, X_n)}{\partial(Y_1, \ldots, Y_{N-1}, S}. \tag{33}$$

Its determinant will encode information on how to transform variables from one to another description. We can now derive the partial derivatives for $i \in [1, N-1]$ as $\frac{\partial X_i}{\partial Y_i} = S$, $\frac{\partial X_j}{\partial Y_i} = 0$, for $i \neq j$ and $\frac{\partial X_i}{\partial S} = Y_i$. For X_N we see $\frac{\partial X_N}{\partial Y_i} = -S$ and $\frac{\partial X_N}{\partial S} = Y_n$. By standard properties of the determinant, we can pull out factor of S from one row, and multiply it onto the determinant afterward. We



(a) IQP (b) RQC

Figure 2: Comparison of output distributions of supremacy circuits and pseudo-independent distributions. We sample 10,000 circuits on 7 qubits and plot $\mu \pm \sigma$, as in Table 1. The red lines are μ and $\mu \pm \sigma$ of a pseudo-independent distribution with an underlying Gamma distribution, showing that the output distribution of supremacy circuits indeed coincides with our distribution family.

do this for all N columns, leaving the Jacobian as

$$J = \begin{bmatrix} 1 & 0 & \dots & 0 & \frac{Y_1}{S} \\ 0 & 1 & \dots & 0 & \frac{Y_2}{S} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \frac{Y_n}{S} \\ -1 & -1 & \dots & -1 & \frac{Y_n}{S} \end{bmatrix} \tag{34}$$

We add all rows to the last row, resulting in an upper triangular matrix, whose determinant is equal to the product of diagonal entries, which is $1/S$. Multiplying by the previous factor of S^N reveals the determinant as

$$|\det(J)| = S^{N-1}. \tag{35}$$

We now turn to the joint probability distribution as

$$p(Y_1, \dots, Y_N, S) = \prod \frac{1}{\Gamma(\alpha)} (SY_i)^{\alpha-1} e^{-SY_i} S^{N-1} \tag{36}$$

$$= \frac{1}{\prod_{i=1}^{N} \Gamma(\alpha)} S^{N-1} S^{N(\alpha-1)} e^S \prod Y_i^{\alpha-1} \tag{37}$$

We can now split this into two parts, one dependent only on S and one on Y.

$$p(S) = \frac{1}{\Gamma(N\alpha)} S^{N\alpha-1} e^S \tag{38}$$

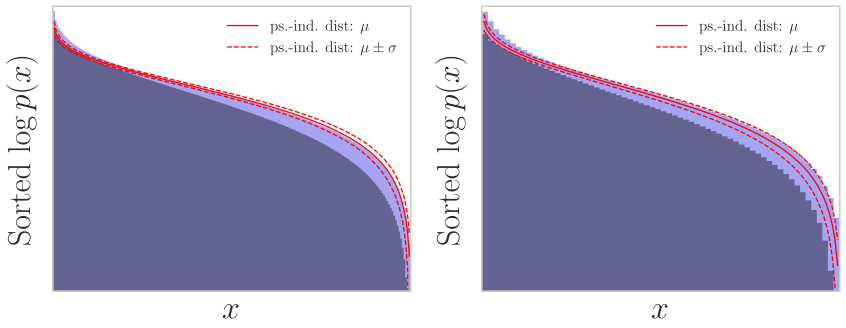
$$p(Y_1, \dots, Y_n) = \frac{\Gamma(N\alpha)}{\prod_{i=1}^{N} \Gamma(\alpha)} \prod Y_i^{\alpha-1} \tag{39}$$

It follows that the joint distribution of the pseudo-independent random variables follows a symmetric Dirichlet distribution $\{Y_1, \dots, Y_N\} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, and S follows the Gamma distribution $S \sim \Gamma(N\alpha, 1)$. □

C Similarity between quantum supremacy and pseudo-independent distributions

The sampling tasks of RQC and IQP have a direct connection to the pseudo-independent distributions, as supported by Appendix B.2. Figure 2 provides additional numerical results supporting the similarity between these distributions and pseudo-independent ones.

However, those are not the only tasks that exhibit sampling advantages. In this section, we briefly explore two other computational approaches with provable sampling advantages, and verify whether they are qualitatively captured by the previous description.



(a) Boson sampling (b) Fermion sampling

Figure 3: Comparison of output distributions of boson and fermion sampling circuits and pseudo-independent distributions. We sample 10,000 circuits on 7 (4) available (active) modes (for bosons), and 8 qubits (for fermion) and plot $\mu \pm \sigma$, as in Table 1. The boson sampling takes as input a Haar-random interferometers, while fermions use a linear optics circuit of length n. The red lines are μ and $\mu \pm \sigma$ of a pseudo-independent distribution with an underlying Gamma distribution, showing that the output distribution of supremacy circuits approximately coincides with our distribution family.

Boson sampling (Aaronson & Arkhipov, 2011) is a non-universal model of computation arguably easier to implement than universal quantum computing. Despite this limitation, sampling hardness is possible by relating the problem to the computation of the permanent of a (large) random matrix. The hardness is, again, related to anticoncentration. In this case, anticoncentration suffers a transition as a function of active vs. available modes established at the input (Ehrenberg et al., 2025). A similar story can be told about fermion sampling. In this case, a linear-optics circuit is applied to a tensor-product of magic input states, which also fulfills sampling hardness (Oszmaniec et al., 2022), by similar hardness arguments. While discussing the specifics of the sources of hardness is out of scope for this manuscript, we encourage the interested reader to refer to the original sources.

In Figure 3 we numerically compute a small example of the previous problems to assess the similarity of those models with respect to pseudo-independent distributions, as an analogous result to Figure 2, and obtain similar results to the one described in the main text. As a note, the numerical simulation of boson sampling utilizes a Haar-random interferometer. For fermion sampling, an active fermionic linear optics circuits with nearest-neighbour connectivity and n layers is enough to approximate the Haar-random distribution in the corresponding subset (Braccia et al., 2025). These results suggest that a variety of models with sampling hardness exhibit a behavior that admits at the very least an approximate description using pseudo-independent variables, from numerical evidences. Theoretical support to this statement is however more challenging than in the case of Haar-random states, and is left for future research.

D Loss Functions

To compute losses, w will consider a candidate distribution p, created by the quantum model, and a target distribution q.

Definition 12 (Squared distance (SD) over probability distributions)**.** *Let p, q be probability distributions over the domain $\mathcal{X} = \{0,1\}^n$. The square distance of the probability distributions is given by*

$$\Delta(p,q) = \sum_{x \in \mathcal{X}} \left(p(x) - q(x)\right)^2 \tag{40}$$

There exists a fundamental motivation to focus on this distance. First, the expectation value of this distance is related to the second statistical moment of the probability distribution, which immediately connects to the definition of anticoncentration from Equation (2). Hence, the SD contains a direct link to classical hardness of sampling. Second, the SD is a simple measure for the similarity of two probability distributions on average, and is a simple, yet effective, tool to analyze states with randomness, such as the ones in RQC or IQP, lying at the core of the considered proofs for hardness.

While useful for generic comparisons between probability distributions, the SD is not sensitive to details in differences that could provide more insightful information for a practitioner. To this end, other distances focusing on certain regions of the probability distribution have been proposed, e.g., by utilizing a kernel $k(x, y) \geq 0$ that weights pairs of data points for a more tailored loss function. The Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) distance is a commonly used kernel-based loss, and it has already been employed in the context of GQML (Rudolph et al., 2024; Recio-Armengol et al., 2025; Herrero-Gonzalez et al., 2025).

Definition 13 (Squared population MMD^2 (Gretton et al., 2012))**.** *Consider independent random variables x, x' and y, y', distributed according to their respective probability distributions p and q, and an arbitrary kernel function $k(\cdot, \cdot) \geq 0$. The squared population MMD^2 is defined as*

$$\mathrm{MMD}_k^2(p, q) = \mathbb{E}_{x,x' \sim p}[k(x, x')] + \mathbb{E}_{y,y' \sim q}[k(y, y')] - 2\mathbb{E}_{x \sim p, y \sim q}[k(x, y)] \tag{41}$$

The MMD is a distance on its own for characteristic kernels, (Gretton et al., 2012; Fukumizu et al., 2007), such as the Gaussian or Laplace kernels, thus, vanishes for equal distributions, i.e.,

$$\mathrm{MMD}_k^2(p, q) = 0 \Leftrightarrow p = q \quad \text{if } k \text{ is characteristic.} \tag{42}$$

In the case of generative models defined over the domain of bitstrings, previous literature (Rudolph et al., 2024; Recio-Armengol et al., 2025) has employed the Gaussian kernel shown below. Here, the 2-norm reduces to a simple Hamming distance, thus, counting distinct bits between x and y.

$$k_\varsigma(x, y) = \exp\left(-\frac{d_H(x, y)}{2\varsigma^2}\right) \tag{43}$$

Especially for large ς, this leads to low-order correlations dominating the loss, which becomes apparent when considering the MMD^2 in its alternative form via the Fourier characters of p and q.

$$\mathrm{MMD}^2(p, q) = \sum_{k=0}^{n} \frac{(1+\rho)^{n-k}(1-\rho)^k}{2^n} \sum_{S \subseteq [n], |S|=k} \left(\hat{P}(S) - \hat{Q}(S)\right)^2 \tag{44}$$

In this representation, we define $\rho = \exp\left(-1/2\varsigma^2\right)$, and $\hat{P}(S)$ as the S-Fourier character of the probability distribution p. The details on how to transform the Hamming-distance into a Fourier-based representation are available in Appendix F. This representation immediately establishes a hierarchy in the Fourier spectrum, since the relative weights differ by a factor $(1-\rho)^k(1+\rho)^{-k} \leq \exp(-\rho k)$. Fourier components with low order become much more relevant, which also has its reflections on trainability (Rudolph et al., 2024). Thus, the MMD^2 is particularly useful in the case where the distributions of interest have strong components of low-frequency Fourier weights.

E Loss function concentration

E.1 Product Distribution

We consider the product distributions from Definition 1.

Lemma 1 (Marginals of the product distribution)**.** *Let $p_{\boldsymbol{a}}$ be a product distribution, with $\boldsymbol{a} \sim U^{(n)}$. The probability distribution of $p_{\boldsymbol{a}}(x)$ is given by*

$$\mathrm{Prob}\left(p(x) = y\right) = \frac{\ln(1/y)^{n-1}}{(n-1)!} \tag{45}$$

for any $x \in \{0, 1\}^n$ and $0 \leq y \leq 1$.

Proof. First, we note that the probability distribution has a strong symmetry by exchanging 0 and 1 in the variable x and elements in $\boldsymbol{a}$. That is, the probability distribution for a given value of a_i perfectly matches that of $(1 - a_i)$ under a transformation $x_i \to x_i \oplus 1$. Since we assume a uniform distribution, both elements are equally probable, and the average over $\boldsymbol{a}$ implicitly performs an average over x. Hence, all marginals are the same, and we just need to focus on one of them, for instance $\boldsymbol{x} = 0 \ldots 0$. The condition is achieved by

$$\prod_{i=1}^{n} a_i = y. \tag{46}$$

We perform the change of variables

$$a_i = \exp(-z_i), \qquad z_i \geq 0, \tag{47}$$

which transforms the condition into

$$\sum_i z_i = \ln(1/y). \tag{48}$$

This condition on $\mathbb{R}^n$ describes a simplex over n dimensions, with volume

$$\mathrm{Vol}_{\sum_i z_i = \ln(1/y)} = \frac{\ln(1/y)^{n-1}}{(n-1)!}. \tag{49}$$

□

Lemma 2 (Marginals of the product distribution - II)**.** *In the assumptions of Lemma 1,*

$$\mathrm{Prob}\left(p(x) \geq \frac{y}{2^n}\right) = \frac{\gamma(n, n\ln 2 - \ln(y))}{\Gamma(n)}, \tag{50}$$

with $\Gamma(\cdot)$ the Gamma function from Definition 4, and $\gamma(\cdot,\cdot)$ the lower incomplete gamma function from Definition 5.

Proof. Following the procedure from the previous lemma, the new condition is given by

$$\prod_{i=1}^n a_i \geq \frac{y}{2^n}, \tag{51}$$

hence

$$\mathrm{Prob}\left(\prod_{i=1}^n a_i \geq \frac{y}{2^n}\right) = \int_{\prod_{i=1}^n a_i \geq \frac{y}{2^n}} \prod da_i \tag{52}$$

$$= \int_{\sum z_i \leq \ln(2^n/y)} \exp\left(-\sum_i z_i\right) \prod_{i=1}^n dz_i \tag{53}$$

$$= \int_0^{n\ln(2)-\ln(y)} \exp(-s)\, \frac{s^{n-1}}{(n-1)!} ds. \tag{54}$$

By definition, this is the lower incomplete gamma function from Definition 5. □

Proposition 1 (Concentration of product distribution)**.** *Let $p_{\boldsymbol{a}}$ be a product distribution. This probability distribution concentrates, as*

$$\forall x: \quad \mathrm{Prob}\left(p_{\boldsymbol{a}}(x) \geq \frac{y}{2^n}\right) \leq y^{-1/2} \exp\left(-\mathcal{O}(n)\right). \tag{5}$$

Proof. We begin from Lemma 2 with the definition of the incomplete gamma function, and a rewriting for integer numbers in the form (Arfken et al., 2013)

$$\frac{\gamma(n,x)}{\Gamma(n)} = \exp(-x) \sum_{k=n}^{\infty} \frac{x^k}{k!}, \tag{55}$$

which matches the probability of obtaining more than n events in a Poisson distribution from Definition 10, with $\lambda = x$. The Chernoff bound for the Poisson distribution is given by (Vershynin, 2018, p.50)

$$\mathrm{Prob}(X \geq a) \leq \left(\frac{\lambda}{a}\right)^a \exp(a - \lambda). \tag{56}$$

In our case,

$$\lambda = n\ln(2) - \ln(y) \tag{57}$$

$$a = n, \tag{58}$$

hence

$$\text{Prob}\left(p(x) \geq \frac{y}{2^n}\right) \leq \left(\frac{n\ln(2) - \ln(y)}{n}\right)^n \exp\left(n - \ln(2)n + \ln(y)\right). \tag{59}$$

Simplifying the first term on the right hand side, we obtain

$$\left(\frac{n\ln(2) - \ln(y)}{n}\right)^n = \left(\ln(2) - \frac{\ln(y)}{n}\right)^n \tag{60}$$

$$= \ln(2)^n \left(1 - \frac{\ln y}{n\ln 2}\right)^n \tag{61}$$

We combine the $\ln(2)^n$ with the second term, obtaining

$$\ln(2)^n \exp\left(n - \ln(2)n + \ln(y)\right) = \exp\left(n\ln(\ln(2)) + n(1 - \ln(2)) + \ln(y)\right) \tag{62}$$

$$= \exp\left(n(1 - \ln(2) + \ln(\ln(2)))\exp\left(\ln(y)\right)\right. \tag{63}$$

$$\approx \exp\left(-\frac{n}{20}\right) y \tag{64}$$

For large n, we use $(1 - \frac{\ln y}{n\ln 2})^n \approx \exp\left(-\frac{\ln(y)}{\ln(2)}\right)$ to find

$$y\exp\left(-\frac{\ln(y)}{\ln(2)}\right)\exp\left(-\frac{n}{20}\right) = \exp\left(\ln(y) - \frac{\ln(y)}{\ln(2)}\right)\exp\left(-\frac{n}{20}\right) \tag{65}$$

$$= \exp\left(\ln(y)\left(1 - \frac{1}{\ln(2)}\right)\right)\exp\left(-\frac{n}{20}\right) \tag{66}$$

$$\approx \exp\left(\ln\left(y^{-0.44}\right)\right)\exp\left(-\frac{n}{20}\right) \approx \frac{\exp\left(-\frac{n}{20}\right)}{\sqrt{y}} \tag{67}$$

which concludes the proof. □

Lemma 3 (Square distance between product distributions)**.** *Let $p_{\boldsymbol{a}}$ be an instance of the product distributions, and let u be the uniform distribution. Then*

$$\mathbb{E}_{\boldsymbol{a}}\left[\Delta(p_{\boldsymbol{a}}, u)\right] \in \exp\left(-\mathcal{O}(n)\right) \tag{68}$$

Proof. The proof is achieved by explicitly computing mean and variance of this probability distribution. We recall that the probability distributions are symmetric, hence

$$\mathbb{E}_{\boldsymbol{a},\boldsymbol{b}}\left[\Delta(p_{\boldsymbol{a}}, u)\right] = 2^n \mathbb{E}_{\boldsymbol{a}}\left[\left(p(0) - u(0)\right)^2\right], \tag{69}$$

and again by symmetry

$$\mathbb{E}_{\boldsymbol{a}}\left[\left(p(0) - u(0)\right)^2\right] = \mathbb{E}_{\boldsymbol{a}}\left[p(0)^2\right] - \mathbb{E}_{\boldsymbol{a}}\left[p(0)\right]^2. \tag{70}$$

By employing the same tools as in Lemma 1, both quantities can be easily computed, as

$$\mathbb{E}_{\boldsymbol{a}}\left[p(0)\right] = \int \prod_{i=1}^n a_i \prod_{i=1}^n da_i = \frac{1}{(n-1)!}\int_0^\infty ds \exp\left(-2s\right) s^{n-1} = \\ \frac{1}{(n-1)!}\frac{1}{2^n}\underbrace{\int_0^\infty (2ds)\exp\left(-2s\right)(2s)^{n-1}}_{\Gamma(n)=(n-1)!} = \frac{1}{2^n} \tag{71}$$

The attentive reader might have seen a simpler way of computing this, since $\mathbb{E}\left[a_i\right] = 1/2$. We choose to keep it in the form of the gamma function for consistency with the rest of the proof.

$$\mathbb{E}_{\boldsymbol{a}}\left[p(0)^2\right] = \int \prod_{i=1}^{n} a_i^2 \prod_{i=1}^{n} da_i = \frac{1}{(n-1)!}\int_0^\infty ds \exp\left(-3s\right) s^{n-1} = \\ \frac{1}{(n-1)!}\frac{1}{3^n}\underbrace{\int_0^\infty (3ds)\exp\left(-3s\right)(3s)^{n-1}}_{\Gamma(n)=(n-1)!} = \frac{1}{3^n}. \tag{72}$$

Joining all results, we find

$$\mathbb{E}_{\boldsymbol{a}}\left[\Delta(p_{\boldsymbol{a}}, u)\right] = \left(\frac{2^n}{3^n} - \frac{1}{2^n}\right) \leq \exp\left(-\ln(3/2)n\right) \in \exp\left(-\mathcal{O}(n)\right). \tag{73}$$

□

Lemma 4 (Maximum mean discrepancy between product distributions)**.** *Let $p_{\boldsymbol{a}}$ be a product distribution depending on a random variable $\boldsymbol{a}$, and let u be the uniform distribution. Consider a kernel k_ς with $\varsigma \in \mathcal{O}(n)$. Then,*

$$\mathrm{MMD}^2\left(p_{\boldsymbol{a}}, u\right) \in \Theta\left(\mathrm{poly}(1/n)\right) \tag{74}$$

with probability $1 - \mathcal{O}\left(\mathrm{poly}(1/n)\right)$.

Proof. The proof is achieved by explicitly computing mean and variance of this probability distribution. From Equation (44),

$$\mathrm{MMD}^2\left(p, u\right) = \sum_{k=1}^{n} \frac{(1+\rho)^{n-k}(1-\rho)^k}{2^n} \sum_{\substack{S\subseteq[n] \\ |S|=k}} \hat{P}^2(S), \tag{75}$$

Since p is a product distribution, each element of the Fourier transform is also a product, given by

$$\hat{P}(S) = \prod_{i=1}^{n}(1-2a_i)^{S_i}. \tag{76}$$

Plugging this result into the definition of the MMD (with Hamming-weight kernel), returns

$$\mathrm{MMD}^2\left(p, u\right) = \left(\frac{1+\rho}{2} + \sum_{i=1}^{n}\frac{1-\rho}{2}(1-2a_i)^2\right)^n - \frac{(1+\rho)^n}{2^n}. \tag{77}$$

We note that the only random element comes from the sum of terms $(1-2a)$, for the independent variables. We use Paley-Zygmund inequality to find a bound on the value of this element. Each term is an independent variable with average $1/3$ and second moment $1/5$. Thus,

$$S = \sum (1-2a_i)^2 \tag{78}$$

$$\mathbb{E}\left[S\right] = n/3 \tag{79}$$

$$\mathbb{E}\left[S^2\right] = n/5 + n(n-1)/9 \tag{80}$$

By the Paley-Zygmund inequality,

$$\mathrm{Prob}\left(S \geq \theta\right) \geq \left(1 - \frac{3\theta}{n}\right)^2 \frac{1}{1+\frac{4}{5}\frac{1}{n}}, \tag{81}$$

and thus

$$\mathrm{MMD}^2\left(p, u\right) \geq \left(\frac{1+\rho}{2} + \theta\left(\frac{1-\rho}{2}\right)\right)^n - \left(\frac{1+\rho}{2}\right)^n, \tag{82}$$

with the probability specified by Paley-Zygmund inequality. Hence, if $\rho = \exp\left(-1/2n^2\right)$, the $\mathrm{MMD}^2\left(p, u\right)$ decays as $\theta/4n$, while for ρ constant, it converges to be exponentially small. □

E.2 Pseudo-Independent Distributions

Consider the pseudo-independent distributions as defined in Definition 2. Useful symbols in this section will be

$$\mu = \mathbb{E}[Y] \tag{83}$$

$$\hat{\mu} = \frac{1}{N}\sum_x Y_x \tag{84}$$

$$\mu_2 = \mathbb{E}\left[Y^2\right] \tag{85}$$

$$\sigma^2 = \mu_2 - \mu^2 \tag{86}$$

Lemma 5 (Difference to average)**.** *Let X be a N-dimensional random variable as defined in Definition 2 and $\{Y_i\}$ its associated i.i.d. random variables. Then*

$$\text{Prob}\left(\frac{|\mu - \hat{\mu}|}{\mu} \geq k\right) \leq \frac{\sigma^2}{Nk^2\mu^2}. \tag{87}$$

Proof. Consider the random variable $\hat{\mu}$. Its variance is given by

$$\text{Var}[\hat{\mu}] = \frac{\sigma^2}{N}, \tag{88}$$

because the variance of a sum of weighted independent random variables is given by

$$\text{Var}\left[\sum_j a_j x_j\right] = \sum_j a_j^2 \,\text{Var}[x_j]\,. \tag{89}$$

We recall now Chebyshev's inequality. For any random variable,

$$\text{Prob}\left(|x - \mathbb{E}[x]| \geq t\right) \leq \frac{\text{Var}[x]}{t^2}. \tag{90}$$

This implies that

$$\text{Prob}\left(|\hat{\mu} - \mu| \geq t\right) \leq \frac{\sigma^2}{Nt^2} \tag{91}$$

The final result is obtained by normalisation of the distance $t \to k\mathbb{E}[Y]$. □

Corollary 1. *Let X be a N-dimensional random variable as defined in Definition 2 with associated i.i.d. random variables $\{Y_i\}$, and let μ and $\hat{\mu}_N$ as in Lemma 5. Then*

$$\text{Prob}\left(\mu\left|\frac{1}{\mu} - \frac{1}{\hat{\mu}_N}\right| \geq k\right) \leq \frac{\sigma^2}{Nk^2\mu^2}. \tag{92}$$

Proof. This corollary follows immediately from Lemma 5. By error propagation

$$\mu\left|\frac{1}{\mu} - \frac{1}{\hat{\mu}_N}\right| \approx \frac{|\mu - \hat{\mu}|}{\mu}\left(1 + \mathcal{O}\left(\frac{|\mu - \hat{\mu}|}{\mu}\right)\right), \tag{93}$$

which yields a first-order equivalence between this result and that of Lemma 5. □

To address the anticoncentration, we can find the following lemma.

Lemma 6 (Anticoncentration of pseudo-independent distributions)**.** *Let X be a N-dimensional pseudo-independent random variable as in Definition 2. Then*

$$\text{Prob}\left(X_i \geq \frac{\alpha}{N}\right) \geq \left(1 - \alpha\left(1 + \frac{1}{k}\right)\right)^2 \left(1 - \frac{\sigma^2 k^2}{N\mu^2}\right)\frac{\mu^2}{\sigma^2} \tag{94}$$

For constants α, k. For this bound to be informative, $\alpha(1 - k^{-1}) \leq 1$.

Proof. The anticoncentration constraint is achieved by bounding

$$\text{Prob}\left(X_i \geq \frac{\alpha}{N}\right) = \text{Prob}\left(Y_i \geq \alpha\hat{\mu}\right). \tag{95}$$

From Lemma 5, and with a slight modification,

$$\text{Prob}\left(\frac{|\hat{\mu}-\mu|}{\mu} \geq k^{-1}\right) \leq \frac{\sigma^2 k^2}{N\mu^2}. \tag{96}$$

This implies

$$\begin{aligned}\text{Prob}\left(Y_i \geq \alpha\hat{\mu}\right) &\geq \text{Prob}\left(Y_i \geq \alpha\mu\left(1+\frac{|\hat{\mu}-\mu|}{\mu}\right)\right) \\ &\geq \text{Prob}\left(Y_i \geq \alpha\mu\left(1+\frac{1}{k}\right)\right)\left(1-\frac{\sigma^2 k^2}{N\mu^2}\right).\end{aligned} \tag{97}$$

We apply now Paley-Zygmund inequality (Paley & Zygmund, 1932) to obtain

$$\text{Prob}\left(Y_i \geq \alpha\mu\left(1+\frac{1}{k}\right)\right) \geq \left(1-\alpha\left(1+\frac{1}{k}\right)\right)^2 \frac{\mu^2}{\sigma^2}. \tag{98}$$

Joining both results

$$\text{Prob}\left(Y_i \geq \alpha\hat{\mu}\right) \geq \left(1-\frac{\sigma^2 k^2}{N\mu^2}\right)\left(1-\alpha\left(1+\frac{1}{k}\right)\right)^2 \frac{\mu^2}{\sigma^2}. \tag{99}$$

To achieve the proposition of interest, we just have to choose adequate values of α and k, e. g. $\alpha = 1/3, k = 1/2$, and see that this bound approximates a constant as N increases. □

It is now straightforward to state:

Proposition 2 (Anticoncentration of pseudo-independent distribution)**.** *Let p be an instance of the pseudo-independent distributions. This probability distribution anticoncentrates, as*

$$\forall x: \quad \text{Prob}\left(p(x) \geq \frac{y}{2^n}\right) \in \Omega(1). \tag{8}$$

To prove Proposition 6, we give the following preliminary result.

Lemma 7 (Concentration of square distance for pseudo-independent distributions)**.** *Let $p_{\theta_p}(x)$ be a probability distributions following Definition 2, and u the uniform distribution. Then,*

$$\text{Prob}\left(\Delta(p_{\theta_p}, u) \geq \frac{k^2}{N}\right) \leq 2\frac{\sigma^2}{\mu^2 k^2} + 2\frac{\sigma^2\mu_2}{\mu^4 k^2 N} = 2\frac{\sigma^2}{\mu^2 k^2}\left(1+\frac{\mu_2}{\mu^2 N}\right), \tag{100}$$

Proof. We begin with the definition

$$\sum_{x\in\{0,1\}^n}\left(p_{\theta_p}(x)-u(x)\right)^2 = \sum_x\left(\frac{Y_x}{N\hat{\mu}}-\frac{1}{N}\right)^2. \tag{101}$$

We note that

$$\begin{aligned}\left(\frac{Y_x}{N\hat{\mu}}-\frac{1}{N}\right)^2 &= \left(\frac{Y_x}{N\mu}-\frac{Y_x}{N\mu}+\frac{Y_x}{N\hat{\mu}}-\frac{1}{N}\right)^2 \leq \\ &2\left(\left(\frac{Y_x}{N\mu}-\frac{1}{N}\right)^2+\left(\frac{Y_x}{N\mu}-\frac{Y_x}{N\hat{\mu}}\right)^2\right)\end{aligned} \tag{102}$$

due to convexity of the square function. We can now bound each term separately and calculate its expectation. The first term is simply

$$\sum_x\left(\frac{Y_x}{N\mu}-\frac{1}{N}\right)^2 = \frac{1}{N^2\mu^2}\sum_x(Y_x-\mu)^2, \tag{103}$$

with average is given by

$$\mathbb{E}\left[(Y_x - \mu)^2\right] = \mathrm{Var}\left[Y_x\right] = \sigma^2, \tag{104}$$

and hence

$$\mathbb{E}\left[\sum_x \left(\frac{Y_x}{N\mu} - \frac{1}{N}\right)^2\right] = \frac{\sigma^2}{N\mu^2}. \tag{105}$$

Following Markov's inequality,

$$\mathrm{Prob}\left(\sum_x \left(\frac{Y_x}{N\mu} - \frac{Y_x}{N\mu}\right)^2 \geq \frac{k^2}{N}\right) \leq \frac{\sigma^2}{\mu^2 k^2} \tag{106}$$

For the second and third term, we can just upper bound the differences between μ and $\hat{\mu}$ to obtain

$$\sum_x \mathbb{E}\left[\left(\frac{Y_x}{N\mu} - \frac{Y_x}{N\hat{\mu}}\right)^2\right] = \frac{1}{\mu^2 N^2}\left(\mu^2\left(\frac{1}{\mu} - \frac{1}{\hat{\mu}}\right)^2\right)\sum_x \mathbb{E}\left[Y_x^2\right] = \frac{\mu_2}{N\mu^2}\left(\mu^2\left(\frac{1}{\mu} - \frac{1}{\hat{\mu}}\right)^2\right). \tag{107}$$

By Corollary 1,

$$\mathrm{Prob}\left(\mu^2\left(\frac{1}{\mu} - \frac{1}{\hat{\mu}}\right)^2 \geq \frac{k^2\mu^2}{\mu_2}\right) = \mathrm{Prob}\left(\frac{\mu_2}{N}\left(\frac{1}{\mu} - \frac{1}{\hat{\mu}}\right)^2 \geq \frac{k^2}{N}\right) \leq \frac{\sigma^2\mu_2}{\mu^4 k^2 N}. \tag{108}$$

Using the union bound,

$$\mathrm{Prob}\left(\Delta(p,u) \geq \frac{k^2}{N}\right) \leq 2\frac{\sigma^2}{\mu^2 k^2} + 2\frac{\sigma^2\mu_2}{\mu^4 k^2 N} = 2\frac{\sigma^2}{\mu^2 k^2}\left(1 + \frac{\mu_2}{\mu^2 N}\right), \tag{109}$$

which yields the final result. □

With the support of the previous result, it is straightforward to state the original proposition:

Proposition 6 (SD of pseudo-independent distributions)**.** *Let p be an instance of the pseudo-independent distributions. Then, with probability at least $1 - \Theta(\delta)$,*

$$\forall x: \quad \Delta(p,u) \leq \frac{\exp\left(-\mathcal{O}(n)\right)}{\delta}. \tag{15}$$

Corollary 2. *Let $p_{\theta_p}(x)$ be an instance of the pseudo-independent probability distributions, and let u be the uniform distribution. Then,*

$$\mathrm{Prob}\left(\Delta(p_{\theta_p}, u) \geq \frac{k^2}{N}\right) \leq \frac{2\sigma^2}{\mu^2 k^2}\left(1 + \frac{\sigma^2}{\mu^2 N}\right). \tag{110}$$

Proof. The proof is analogous to the previous lemma, just substituting q_{θ_q} with a fixed variable $u(x) = N^{-1}$. The difference $|\hat{\mu} - \mu|$ is controlled as before, and we just need to take into account

$$\mathbb{E}\left[\left(\frac{Y_x}{\mu} - 1\right)^2\right] = \frac{\sigma^2}{\mu^2}, \tag{111}$$

hence yielding

$$\mathbb{E}\left[\sum_x \left(\frac{Y_x}{\mu N} - \frac{1}{N}\right)^2\right] \leq \frac{\sigma^2}{N\mu^2}\left(1 + \frac{\sigma^2}{N^2\mu}\right). \tag{112}$$

□

Corollary 3. *Let $p_C(x)$ be an instance of the pseudo-independent probability distributions generated by a quantum circuit C. Let Z be any diagonal Pauli string. Then,*

$$\mathrm{Var}\left[\langle 0| CZC |0\rangle\right] \leq \frac{\sigma^2}{N}\,\mathrm{Prob}\left(\Delta(p_{\theta_p}, u) \geq \frac{k^2}{N}\right) \leq \frac{2\sigma^2}{\mu^2 k^2}\left(1 + \frac{\mu_2}{\mu^2 N}\right). \tag{113}$$

Proof. The proof is analogous to the previous lemma. Any Pauli string of this kind splits the x into those associated with ± 1. This implies

$$\langle 0| CZC |0\rangle = \frac{\sum_{x_+} Y_{x_+} - \sum_{x_-} Y_{x_-}}{N\hat{\mu}}. \tag{114}$$

As before,

$$\left| \langle 0| CZC |0\rangle - \frac{\sum_{x_+} Y_{x_+} - \sum_{x_-} Y_{x_-}}{N\mu} \right| \leq \frac{\sum_{x_+} Y_{x_+} - \sum_{x_-} Y_{x_-}}{N} \left| \frac{1}{\hat{\mu}} - \frac{1}{\mu} \right| \leq \frac{\sum_{x_+} Y_{x_+} - \sum_{x_-} Y_{x_-}}{N\mu} \frac{\sigma^2}{Nk^2\mu^2}, \tag{115}$$

where the last inequality is given with probability at least $1 - k$. The sum of independent random variables has mean 0, and its variance is given by

$$\mathrm{Var}\left[\frac{\sum_{x_+} Y_{x_+} - \sum_{x_-} Y_{x_-}}{N\mu} \right] = \frac{\sigma^2}{N\mu^2}. \tag{116}$$

Thus,

$$\mathrm{Var}\left[\langle 0| CZC |0\rangle \right] \leq \frac{\sigma^2}{N\mu^2} \left(1 + \frac{\sigma^2}{N\mu^2} \right) \tag{117}$$

□

Proposition 7 (SD and Fourier-SD concentrate equally)**.** *With probability at least* $1 - \Theta(\delta)$

$$\mathrm{MMD}^2\left(p, u\right) \leq \frac{\exp\left(-\mathcal{O}(n)\right)}{\delta}. \tag{18}$$

Proof. We just need to see that

$$\mathbb{E}\left[\mathrm{MMD}^2\left(p, u\right) \right] = \sum_{k=0}^{n} \frac{(1+\rho)^{n-k}(1-\rho)^k}{2^n} \sum_{\substack{S \subseteq [n] \\ |S|=k}} \frac{\sigma^2}{N\mu^2}. \tag{118}$$

Identifying that there are $\binom{n}{k}$ elements of length k, this is the binomial theorem, and

$$\sum_{k=0}^{n} \frac{(1+\rho)^{n-k}(1-\rho)^k}{2^n} \binom{n}{k} = 1. \tag{119}$$

The probabilistic statement is a direct consequence of Markov's inequality. □

E.3 Peaked Distributions

Proposition 8 (SD of peaked pseudo-independent distributions)**.** *Let* $p_{\boldsymbol{a}}, q_{\boldsymbol{b}}$ *be two instances of the peaked pseudo-independent distributions. Then, with probability at least* $1 - \Theta(\delta)$*,*

$$\forall x: \quad \Delta(p_{\boldsymbol{a}}, q_{\boldsymbol{b}}) \leq \frac{1}{K\delta}. \tag{19}$$

Proof. The proof follows the same strategy as that of Lemma 7, just changing the varible N with K. For the bounds of $\hat{\mu}$ and μ, the same results hold. Changes apply in the case of the cross term, that is

$$\sum_i \mathbb{E}\left[\left(Y_i - Y_i' \right)^2 \right] = 2K\mathbb{E}\left[Y^2 \right] - 2J\mathbb{E}\left[Y \right]^2, \tag{120}$$

with J being the number of coincidences of the K out of N elements found in each sample. The number of coincidences J follows a probability distribution

$$\mathrm{Prob}\left(J\right) = \frac{\binom{N}{K}\binom{K}{J}\binom{N-K}{N-J}}{\binom{N}{K}^2}. \tag{121}$$

The numerator counts for (a) the number of combinatorial options in choosing K elements out of N in the first sample, (b) number of possible elements J coinciding from the K original, and (c) number of possibilities to rearrange the remaining terms. The denominator captures the total number of samples. This is known as a hypergeometric distribution (Hoeffding, 1963), with moments

$$\mathbb{E}[J] = \frac{K^2}{N} \tag{122}$$

$$\mathrm{Var}[J] = \frac{K^2}{N}\frac{N-K}{N}\frac{N-K}{N-1}. \tag{123}$$

In the limit of $K \ll N$, this probability distribution can be approximated by a Poisson distribution with parameter $\lambda = K^2 N^{-1}$. The Chernoff bound used in Lemma 2 shows that $J = 0$ with probability at least $1 - K^2 N^{-1}$. Therefore,

$$\mathrm{Prob}\left(\Delta(p,q) \geq \frac{k^2}{K}\right) \leq 6\frac{\sigma^2}{\mu^2 k^2} + 6\frac{\sigma^2 \mu_2}{\mu^4 k^2 K} = 6\frac{\sigma^2}{\mu^2 k^2}\left(1 + \frac{\mu_2}{\mu^2 N}\right) + \frac{K^2}{N}. \tag{124}$$

□

F Relation between MMD and Fourier moments

In this section we derive the details for the comparisons between MMD^2 and MSD. As a preliminary, we define the Fourier parity moments of a distribution. The domain of all distributions is $\mathcal{X} = \{0,1\}^n$, with n as the number of qubits. This domains has a direct relation with the group $\mathbb{F}_2^n$, whose Fourier transform is given as follows.

Definition 14 (Fourier characters of a distribution O'Donnell (2014))**.** *For any probability distribution $P(x)$ defined on $\mathcal{X}$, its Fourier decomposition over $\mathbb{F}_2^n$ is*

$$\hat{P}(S) = \sum_{x \in \mathcal{X}} P(x)\,\chi_S(x) \equiv \mathbb{E}_{x\sim P}[\chi_S(x)], \tag{125}$$

with $\chi_S(x) = (-1)^{\sum_{i\in S} x_i}$.

Notice that a set S of small cardinality will capture few-body correlations, while larger $|S|$ captures more global structures. This Fourier description is matched by the Gaussian kernel.

Definition 15 (Gaussian kernel)**.** *The Gaussian kernel is given by*

$$K_\varsigma(x,y) = \exp\left(-\frac{d_H(x,y)}{2\varsigma^2}\right), \tag{126}$$

with ς as bandwidth and $d_H(x,y)$ the Hamming distance between bitstrings x and y.

Note that the usage of Hamming distance as the internal structure of the kernel allows us to phrase it as a product kernel over individual bits, namely

$$K_\varsigma(x,y) = \exp\left(-\frac{d_H(x,y)}{2\varsigma^2}\right) \tag{127}$$

$$= \exp\left(-\frac{\sum_{i=1}^n |x_i - y_i|}{2\varsigma^2}\right) \tag{128}$$

$$= \prod_{i=1}^n \exp\left(-\frac{|x_i - y_i|}{2\varsigma^2}\right) \tag{129}$$

$$= \prod_{i=1}^n \exp\left(-\frac{1}{2\varsigma^2}\mathbb{1}_{x_i \neq y_i}\right) \equiv \prod_{i=1}^n k(x_i, y_i). \tag{130}$$

Under this rewriting, it is clear that the kernel is a positive semidefinite matrix, composed by units of the form

$$K^{(1)} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{with } \rho = \exp\left(-1/2\varsigma^2\right) \tag{131}$$

The product structure of the kernel admits a direct interpretation in the form of the Fourier basis previously stated.

Lemma 8 (Diagonalizing the kernel through Fourier basis)**.** *Consider the Gaussian kernel from Definition 15. It can be rewritten in a diagonal basis as*

$$K_\varsigma(x,y) = \frac{1}{2^n}\sum_{S\subseteq[n]}(1-\rho)^{|S|}(1+\rho)^{n-|S|}\chi_S(x)\chi_S(y), \tag{132}$$

with $\chi_S(x)$ being the Fourier characters of a bitstring x.

Proof. We start by taking the kernel

$$K^{(1)} = \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}, \tag{133}$$

which can be diagonalized as

$$K_\varsigma = \sum_{k=0}^{d}\lambda_k\phi_k\phi_k^T. \tag{134}$$

In this case

$$\lambda_k = 1+(-1)^k\rho \tag{135}$$

$$\phi_k^T = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\(-1)^k\end{pmatrix}\begin{pmatrix}1 & (-1)^k\end{pmatrix} \tag{136}$$

Then, a concrete entry for the kernel is recovered as

$$K(x,y) = \frac{1+\rho}{2} + \frac{1-\rho}{2}\chi_{\{1\}}(x)\chi_{\{1\}}(y) \tag{137}$$

The kernel is a product kernel, hence

$$K_\varsigma(x,y) = \prod_{i=1}^{n}K_\varsigma^{(1)}(x_i,y_i) \tag{138}$$

$$= \prod_{i=1}^{n}\left(\frac{1+\rho}{2} + \chi_{\{1\}}(x_i)\chi_{\{1\}}(y_i)\frac{1-\rho}{2}\right) \tag{139}$$

$$= \frac{1}{2^n}\prod_{i=1}^{n}\left(1+\rho+\chi_{\{1\}}(x)\chi_{\{1\}}(y)(1-\rho)\right) \tag{140}$$

$$= \frac{1}{2^n}\sum_{S\subseteq[n]}\left(\prod_{i\in S}(1-\rho)\chi_{\{1\}}(x_i)\chi_{\{1\}}(y_i)\prod_{i\notin S}(1+\rho)\right) \tag{141}$$

$$= \frac{1}{2^n}\sum_{S\subseteq[n]}(1-\rho)^{|S|}(1+\rho)^{n-|S|}\chi_S(x)\chi_S(y). \tag{142}$$

□

With this preliminary results, it is then straightforward to write the MMD in the Fourier basis.

Lemma 9 (MMD in the Fourier basis)**.** *Consider the Gaussian kernel applied to two probability distributions P, Q, then*

$$\mathrm{MMD}^2(P,Q) = \frac{1}{2^n}\sum_{k=0}^{n}(1+\rho)^{n-k}(1-\rho)^k\sum_{\substack{S\subseteq[n]\\|S|=k}}\left(\hat{P}(S)-\hat{Q}(S)\right)^2, \tag{143}$$

with $\hat{P}(s) = \mathbb{E}_{x\sim P}[\chi_S(x)]$.

Corollary 4 (Bounding the MMD with mean squared distance)**.** *For a Gaussian kernel,*

$$\mathrm{MMD}^2(P,Q) \leq \max_S\left(\hat{P}(S)-\hat{Q}(S)\right)^2. \tag{144}$$

Proof. The proof is immediate from Lemma 8 by using Hölder's inequality, namely

$$\|fg\|_1 \leq \|f\|_1 \|g\|_\infty. \tag{145}$$

We identify f with the eigenvalues and g with the Fourier coefficients. Then, it is clear that

$$\frac{1}{2^n}\sum_{k=0}(1+\rho)^{n-k}(1-\rho)^k \sum_{\substack{S\subseteq[n]\\|S|=k}} 1 = \sum_{k=0}^{n}\binom{n}{k}\frac{(1+\rho)^{n-k}(1-\rho)^k}{2^n} = 1, \tag{146}$$

and the ∞-norm is associated with the largest Fourier coefficient. □

G Bounds on the MMD^2 estimator and statistical test

The following bounds are an extension for $m \neq n$ of the bounds derived for the unbiased estimator for the MMD^2 and associated statistical test derived in Gretton et al. (2012). First, we recall McDiarmid's inequality, which we will use to bound the convergence of the estimator to its true value.

Theorem 2 (McDiarmid's InequalityMcDiarmid (1989))**.** *We consider random variables $X_1, \ldots, X_m$ on the domain $\mathcal{X}$, and a function $f : \mathcal{X}^m \to \mathbb{R}$, where $\forall i \in [m]$ and $\forall x_1, \ldots, x_m, x_i' \in \mathcal{X}$, there exists a $c_i < \infty$ satisfying*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i \tag{147}$$

Then, $\forall t > 0$

$$\Pr_X\left[f(X) - \mathbb{E}_X(f(X)) > t\right] < \exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2}\right) \tag{148}$$

For a bounded kernel, we can obtain the bounded differences property as follows.

Lemma 10 (Bounded differences in the $\widehat{\mathrm{MMD}}^2$)**.** *Changing a single sample from X or Y, will result in changes of at most $\frac{4K}{m}$ or $\frac{4K}{l}$ respectively.*

Proof. A sample in the within-X block of the $\widehat{\mathrm{MMD}}^2$ will occur exactly $2(m-1)$ times and from the boundedness of the kernel, each component will change by less than K. Factoring in the normalization reveals a change by at most $\frac{2K}{m}$. Analogously, a change of sample in Y will result in a change of at most $\frac{2K}{l}$.

For the cross term, we observe that a sample from X occurs exactly l times, resulting in a change of at most $\frac{2K}{m}$, and $\frac{2K}{l}$ for a sample from Y. Summing up concludes the proof. □

Inserting the bounds into McDiarmid's inequality then reveals the desired bound on the convergence.

Theorem 3 (Convergence of the $\widehat{\mathrm{MMD}}^2$ to its true value)**.** *Given a bounded kernel $0 \leq k(\cdot,\cdot) \leq K$, the unbiased statistic converges to the true MMD^2 as*

$$Pr_{X,Y}\left[|\widehat{\mathrm{MMD}}^2 - MMD^2| > t\right] < \exp\left(-\frac{2t^2(m+l)}{8K^2}\right) \tag{149}$$

Proof. The proof follows directly from applying McDiarmid's inequality with the bounded differences obtained in Lemma 10 □

Corollary 5 (Statistical test for $\widehat{\mathrm{MMD}}^2$)**.** *A hypothesis test of level α for the null hypothesis $p = q$ has the acceptance region*

$$\widehat{\mathrm{MMD}}^2 \leq K\sqrt{\frac{8}{m+l}\log(\alpha^{-1})} \tag{150}$$

Proof. To obtain an α-level test, we consider

$$\Pr_{X,Y}\left[|\widehat{\mathrm{MMD}}^2| > t\right] \leq \alpha \tag{151}$$

Inserting the bound from Theorem 3 and solving for t concludes the proof. □

H COMPUTING THE TVD FOR PSEUDO-INDEPENDENT DISTRIBUTIONS

Lemma 11 (Absolute Difference)**.** *Let p, q be two distributions from the pseudo-indepdendent family. Then, the 1-norm, defined as*

$$\|p-u\|_1 = \sum_x |p(x)-u(x)| \tag{152}$$

has statistical properties

$$\mathbb{E}\left[\|p-u\|_1\right] = M(Y) + \mathcal{O}(N^{-1}) \tag{153}$$

$$\mathrm{Var}\left[\|p-u\|_1\right] = \frac{4}{N}\left(\frac{\sigma^2}{\mu^2} - M^2(Y)\right) + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{154}$$

with $M(Y)$ being the absolute mean deviation of the underlying probability distribution.

Proof. We consider the random variable from Definition 2. The first step is to compute one element of the distance

$$\sum_x \left|\frac{Y_x}{N\hat{\mu}} - \frac{1}{N}\right| = \sum_x \left|\frac{1}{N\mu}(Y_x-\mu) + \frac{Y_x}{N}\left(\frac{1}{\hat{\mu}} - \frac{1}{\mu}\right)\right|, \tag{155}$$

and therefore

$$\left|\sum_x \left|\frac{Y_x}{N\hat{\mu}} - \frac{1}{N}\right| - \sum_x \frac{1}{N\mu}|Y_x-\mu|\right| \leq \sum_x \frac{Y_x}{N}\left|\frac{1}{\mu} - \frac{1}{\hat{\mu}}\right|. \tag{156}$$

Following the steps from Appendix E, we can now compute averages to obtain

$$\left|\sum_x \mathbb{E}\left[\left|\frac{Y_x}{N\hat{\mu}} - \frac{1}{N}\right|\right] - \sum_x \frac{1}{N\mu}\mathbb{E}\left[|Y_x-\mu|\right]\right| \leq \mu\left|\frac{1}{\mu} - \frac{1}{\hat{\mu}}\right|. \tag{157}$$

The approximation to the average can be easily computed as the normalized mean absolute deviation of the independent probability distribution, namely

$$M(Y) = \frac{1}{\mu}\mathbb{E}\left[|Y-Y'|\right]. \tag{158}$$

Hence, the average follows

$$\sum_x |p(x)-u(x)| = M(Y) + \mathcal{O}(N^{-1}). \tag{159}$$

The second part of the proof is related to the variance. We can set bounds on each independent term to state

$$\mathrm{Var}\left[|p(x)-u(x)|\right] = \mathbb{E}\left[\left|\frac{Y_x}{\hat{\mu}N} - \frac{1}{N}\right|^2\right] - \mathbb{E}\left[\left|\frac{Y_x}{\hat{\mu}N} - \frac{1}{N}\right|\right]^2. \tag{160}$$

The first term is analogous to the average of the MSD discussed in Proposition 6. Since the average without the absolute value is zero, then,

$$\mathbb{E}\left[\left|\frac{Y_x}{\hat{\mu}N} - \frac{1}{N}\right|^2\right] = \mathrm{Var}\left[\left|\frac{Y_x}{\hat{\mu}N} - \frac{1}{N}\right|\right] = \mathrm{Var}\left[\frac{Y_x}{\hat{\mu}N}\right]. \tag{161}$$

This quantity has already been computed in Corollary 2.

For the second term we can use the triangle bound previously found. Joining all terms together,

$$\mathrm{Var}\left[\sum_x |p(x)-u(x)|\right] \leq \frac{4}{N}\left(\frac{\sigma^2}{\mu^2} - M^2(Y)\right) + \mathcal{O}(N^{-2}). \tag{162}$$

□

As a reference, the Gini coefficient for the Gamma distribution, which is associated to the Haar-random distribution of states, is $2/e$.

I EXPERIMENTAL SETUP

Product IQP We will use IQP circuits with only single-qubit operations as a proxy for these states. Those circuits give rise to probabilities of the form

$$p_{\boldsymbol{\theta}}(x) = \prod_{i=1}^{n} |\langle x_i| R_x(\theta_i) |0\rangle|^2, \tag{163}$$

yielding $p_{\boldsymbol{\theta}}(x)$ as a product of sines and cosines, which admit a transformation to match Definition 1.

Matrix Product States (MPS) MPS are a family of classically tractable states with low entanglement, whose complexity is governed by its bond dimension χ, which controls qubit correlations. While MPS were originally designed to compute expectation values, they also allow for a sampling procedure (Ferris & Vidal, 2012). Thus, we employ them as an extension of our product distribution, incorporating slightly entangled states. Per definition, setting $\chi = 1$ corresponds to the product distribution from Definition 1, thus, it is the same as the product IQP outlined above. To extend the analytic results, we choose $\chi = n$, and generate random MPS using `quimb` (Gray, 2018).

IQP Circuits In the context of this paper, IQP circuits correspond to pseudo-independent distributions, where each variable follows a Gamma distribution (see Appendix B.1). Practically, IQP circuits are composed by 1) a layer of Hadamard gates, 2) an arbitrary diagonal gate, typically composed by several diagonal operations, and 3) another layer of Hadamard gates. For our numerical calculations, we generate random IQP circuits using `Pennylane` (Bergholm et al., 2022) and the `IQPopt` (Recio-Armengol & Bowles, 2025) package. We set the `max_weight` parameter to 2, corresponding to the maximum number of qubits on which each individual diagonal operation acts. The circuits are composed by all possible such pairings without geometrical restrictions. The employed operations are rotations around the x-axis with angles θ sampled uniformly over $[0, 2\pi]$.

We highlight that this family of IQP does not correspond to a fully pseudo-independent distribution, as this would require an exponential number of independent angles. Despite that, this restricted version already suffices to prove classical hardness on sampling Bremner et al. (2016), which is why we employ the model as a practically useful and considerably cheaper approximation of the full family.

Pareto Distribution To extend our analysis of pseudo-independent distributions, we additionally consider a Pareto distribution, that is, a random distribution whose density function decays as $\mathcal{O}(Y^{-\alpha})$ (see Definition 11). This distribution is interesting as an example of a pseudo-independent distribution that does not decay as quickly as that of Haar-random states. In our numerical example we choose the value $\alpha = 2$, where the variance becomes undefined, thus eliminating any tightness in the analytical bounds, while maintaining a well-behaved probability distribution for unbounded values of Y. The interest in analyzing this distribution is to gain insight into the behavior of the loss functions, where analytical guarantees fail.

To the best of our knowledge, there is no known corresponding quantum circuit. However, this is not strictly required per our formalism as it suffices to have a description of the probability distributions. We use `scipy` (Virtanen et al., 2020) to generate the random variables accordingly.

Peaked Circuits Finally, we generate peaked probability distributions by considering IQP circuits on a subset of $\log(n)$-qubits, and randomly distribute them in the n-qubit Hilbert space. This setting corresponds to the pseudo-independent peaked distributions from Definition 3, with $K \in \Theta(n)$.

J EXTENSION OF NUMERICS

As a comparison to the MMD^2, we show the TVD in Appendix J. While the supremacy distributions also show significantly smaller TVD, it remains approximately the same across system sizes (≈ 0.5). The convergence value of the loss function thus depends jointly on the loss function and the considered distributions. However, the convergence of the variance is dominated by the output distributions.

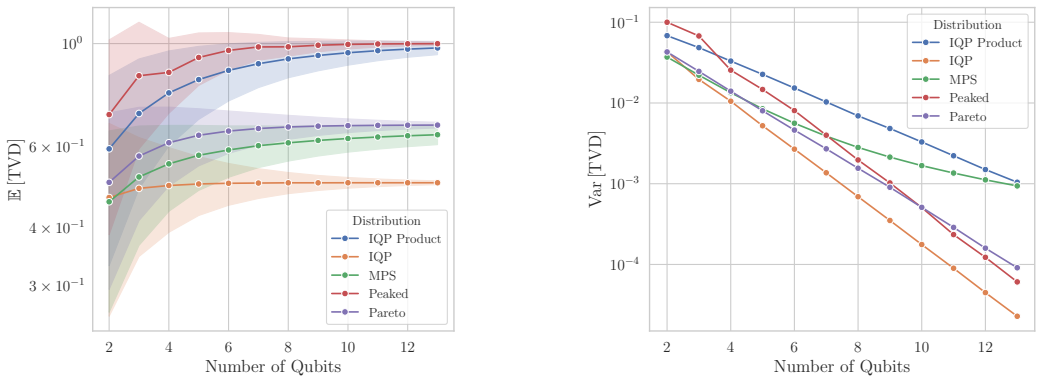


Figure 4: Average and variance of the TVD for the considered models. Untrainability in this case is not a result of exponentially vanishing losses, but rather, of exponentially decaying loss variance.

The differences in the convergence value of the TVD are rooted in the definitions of the model. Top to bottom, we reasons are as follows. In the case of peaked circuits, the TVD measures differences between two probability distributions with disjoint support (up to a small correction captured by the hypergeometric distribution). Hence, the TVD takes similar values as the TVD of any distribution with respect to 0, which is always 1. For IQP Product circuits, the reason for a large average of the TVD is similar. In Table 1 it is seen that the product distributions have a profile of probability that decays exponentially faster. That is, the relevant distribution is concentrated in a small set of entries, which leads to large TVD by the same argument as before. For MPS and Pareto distributions, the argument is more subtle. As the underlying probability distributions decay only polynomially, then there are more entries that contribute with changing weights. This leads to a TVD in between peaked / IQP products and full IQP. In the case of IQP, the TVD approaching $1/2$ is a consequence of the *almost but not exactly* uniform distribution.