Parameterized quantum circuits as universal generative models for continuous multivariate distributions

Alice Barthe,[1,2,3] Michele Grossi,[2] Sofia Vallecorsa,[2] Jordi Tura,[1,3] and Vedran Dunjko[1,3]

[1]$\langle aQa^L \rangle$ *Applied Quantum Algorithms, Universiteit Leiden, the Netherlands*
[2]*Quantum Technology Initiative, CERN, Geneva, Switzerland*
[3]*Instituut-Lorentz, Universiteit Leiden, the Netherlands*

Parameterized quantum circuits have been extensively used as the basis for machine learning models in regression, classification, and generative tasks. For supervised learning, their expressivity has been thoroughly investigated and several universality properties have been proven. However, in the case of quantum generative modelling, much less is known, especially when the task is to model distributions over continuous variables. In this work, we elucidate expectation value sampling-based models. Such models output the expectation values of a set of fixed observables from a quantum circuit into which classical random data has been uploaded. We prove the universality of such variational quantum algorithms for the generation of multivariate distributions. We explore various architectures which allow universality and prove tight bounds connecting the minimal required qubit number, and the minimal required number of measurements needed. Our results may help guide the design of future quantum circuits in generative modelling tasks.

I. INTRODUCTION

Parameterized quantum circuits are the centrepiece of numerous approaches to machine learning on quantum computers, motivated by numerous near-term hardware limitations [BCLK+22, CAB+21]. The use of these models for supervised learning has been widely studied, for example in solving regression problems [MCLS20] where the goal is to assign continuous labels to data points. Variational algorithms have also been explored in so-called generative modelling tasks, where the objective is to generate new samples following a distribution that generated the training data [LW18].

A prominent example of distributions with discrete support is the quantum Born machine [LW18], which stores a distribution over n-bit strings in a n-qubit state. Another model for discrete distributions is the quantum Boltzmann machine [AAR+18]. Going beyond distribution with discrete support, an approach has been introduced to model distributions where the random variable can in principle take on any value within a continuous interval. In such models, the quantum circuit takes classical randomness as input and outputs expectation values, consequently, we call this model expectation value samplers. This model has been extensively used as a quantum generator in the context of quantum generative adversarial networks [RAG21, DDK18].

While for quantum Born machines, the expressivity and universality have been clarified [LW18, CMDK20], in contrast, expectation value sampling models are not so well understood. In particular, an interesting feature of expectation value sampling is that the dimension of the output is not inherently tied to the number of qubits used. In this work, we focus on the expressivity of the generators based on expectation value sampling depending on the number of qubits and the spectrum of the observables.

In the first section, we introduce and define expectation value sampling. We define formally what a universal generative model family is, and give some background on random variable transformation and some fundamental properties.

The second section addresses whether every arbitrary distribution can be modelled using expectation value sampling, i.e. whether this model is universal. We recall existing results on the universality of parameterized quantum circuits. We extend existing results and use them to constructively prove the existence of two families of universal expectation value models.

Following this, in the third section, using variants of the Holevo bound, we look into the necessary conditions for an expectation value scheme to be universal. We show that there exists a necessary condition as an upper bound on the dimensionality of the target distribution as a combination of the number of qubits and the observable spectral norm. By leveraging the scaling with the observable norm, we show that there exists a trade-off between the number of measurements and the number of qubits to reach universality.

The final fourth section provides tools to analyse how the number of qubits, the data encoding scheme, and the set of observables affect the expressivity of the expectation value sampling models, with a focus on the dependence relationship between random variables.

II. BACKGROUND AND DEFINITIONS

In this section, we give formal definitions of expectation value sampling models and universal generative model families. We also introduce random variable transformations, a core concept in this work.
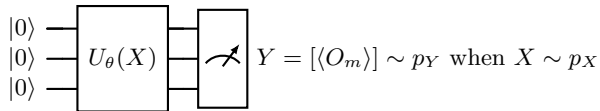
FIG. 1: Expectation value sampling model: a random vector is classically sampled. It is used to generate a random quantum state using a parameterized quantum circuit. The expectation values of fixed observables are returned as the output random vector.

A. Definition of expectation value sampling

The expectation value sampling procedure goes as follows. A random variable is classically sampled and used as input to a parameterized quantum circuit which specifies a random quantum state. The expectation values of fixed observables are measured and returned as another random variable. We illustrate this procedure in Figure 1. It is important to note that, in contrast to the quantum circuit Born machine, the randomness does not come from the measurement process, but from the classical randomness provided as an input to the quantum circuit. Formally we define an expectation value sampling model as follows.

Definition 1 (Expectation value sampling model). *An expectation value sampling model on n qubits is defined by $(U_\theta, \mathbf{O}, p_X)$, where $U_\theta : \mathcal{X} \subseteq \mathbb{R}^L \to \mathcal{U}(2^n)$ is a parameterized quantum circuit taking data as input and returning a unitary matrix, $\mathbf{O} = (O_m)_{1 \leq m \leq M}$ is a vector of M observables, and $p_X : \mathcal{X} \subseteq \mathbb{R}^L \to \mathbb{R}$ is the input distribution. We define the associated mapping f as follows:*

$$x \in \mathcal{X} \xrightarrow{f} \left(\langle 0 | U_\theta(x)^\dagger O_m U_\theta(x) | 0 \rangle \right)_{1 \leq m \leq M}. \quad (1)$$

The output of the model is a sample drawn from the distribution p_Y with $Y = f(X) \sim p_Y$ when $X \sim p_X$.

In the above, the N-dimensional unitary group is denoted $\mathcal{U}(N)$.

Notably, unlike QCBMs, expectation value samplers have continuous support (absolutely continuous random variables, see Definition 6).

The central question of this work is whether such a model can generate any multivariate distribution, more precisely whether expectation value sampling is a universal generative model, according to the definition we will give in the next section.

B. Definition of a universal generative model family

In this subsection, we define universal generative model families.

Definition 2 (Universal generative model family). *A generative model is a family of parameterized sampling*

procedures which enable the sampling from a corresponding set of M-dimensional probability density functions $\mathcal{P}(\mathcal{X})$ on $\mathcal{X} \subseteq \mathbb{R}^M$.

A generative model is called universal if for every probability density function q on $\mathcal{X}$ there exists a sequence $\{p_k | p_k \in \mathcal{P}(\mathcal{X})\}_{1 \leq k \leq \infty}$ such that the sequence of random variables $X_k \sim p_k$ converges in distribution to $X \sim q$. Equivalently, this means that the sequence of cumulative distribution functions of p_k, which we call P_k, converges pointwise to the cumulative distribution function of q, which we call Q.

$$\forall x \in \mathcal{X}, \lim_{k \to \infty} P_k(x) = Q(x). \quad (2)$$

This definition of closeness for random variables, also called convergence in distribution, is common in probability theory. For example, the central limit theorem precisely states the average of any L independent random variables with mean μ and variance Σ converges in distribution to the normal distribution $\mathcal{N}(\mu, \Sigma/L)$.

In this work, we will mostly consider distributions with finite support, because this is the case in most practical real-world problems. In particular, their probability density functions are integrable functions and convergence in distribution implies convergence in the first Wasserstein distance W_1 [Vil09]. For completeness, we recall below the definition of the Wasserstein distance, also known as the Earth Mover's Distance, that we use in the context of this work.

Definition 3 (Wasserstein distance). *The k-th Wasserstein distance between two probability density functions p and q on $[-1, 1]^M$ is defined as:*

$$W_k(p, q) = \left(\inf_{\gamma \in \Pi(p,q)} \int_{\mathbb{R}^2} \|x - y\|^k d\gamma(x, y) \right)^{\frac{1}{k}}, \quad (3)$$

where $\Pi(p, q)$ is the set of couplings of p and q, and $\| \cdot \|$ denotes the Euclidean distance. The parameter $k \geq 1$ determines the so-called order of the Wasserstein distance.

For the rest of the paper, we will only consider the first-order Wasserstein distance ($k = 1$) and therefore simply refer to it as the Wasserstein distance.

Note: Universality is defined on a given support noted as $\mathcal{X}$ in Definition 2. For this work, we choose the support to be $\mathcal{X} = [-1, 1]^M$, because the first step of most machine learning pipelines is to rescale the data to fit on a given interval.

Choosing universality on a cubic support $[-1, 1]^M$ allows the expression of fully independent variables. Restricting $\mathcal{X}$ to smaller subsets would yield constraints on the dependence relationship expressible. For example, proving universality of a family of models for Dirichlet distributions would correspond to choosing $\mathcal{X} = \mathbb{S}_1^1$ where $\mathbb{S}_1^1 := \{x \in \mathbb{R}^M \geq 0 \mid \sum_m x_m = 1\}$ is the unit sphere for the 1-norm.

C. Random Variable Transformation

A core concept in this work is that of random variable transformation. In this subsection, we introduce it and provide some of the associated fundamental properties.

The first step in our analysis is the observation that an expectation value sampling model is a process that maps an input random vector (parameterizing the quantum circuit) to an output random vector (the expectation values of a set of observables). The literature on optimal transport and measure theory [BKM05] states that for every pair of absolutely continuous random variables of the same dimension, there exists a mapping to transform one into the other.

Lemma 1. *For every pair of probability density functions $p_X \in \mathcal{P}(\mathcal{X} \subseteq \mathbb{R}^{\mathcal{M}})$ and $p_Y \in \mathcal{P}(\mathcal{Y} \subseteq \mathbb{R}^{\mathcal{M}})$, there exists a mapping $f : \mathcal{X} \to \mathcal{Y}$, that maps $Y \sim p_Y$ to $X \sim p_X$ as $Y = f(X)$.*

We give intuition on how to construct this mapping and show that it can be chosen to be bounded piece-wise continuous in Appendix A.

Many generative modelling systems including Generative Adversarial Networks, Variational Auto-Encoders and normalizing flows rely on this idea to generate arbitrary distributions. In particular, we choose the initial distribution to be something simple, and then by altering the mapping applied to this random input, we obtain a rich spectrum of possible output distributions.

Then, sufficient conditions for a family of mappings to yield a universal generative model, in the sense of Definition 2, are well-known, and expressed in terms of the universality of mappings themselves. From [KPB21], it is sufficient for a family of mappings to be dense in the set of all monotonically increasing functions in the pointwise convergence topology to yield a universal generative model. We explain the difference between pointwise topology and uniform topology in Appendix C. Since these are sufficient conditions and monotonic functions are included in all functions, the following holds.

Theorem 1. *If a family of mappings $\mathcal{G} = \{g : \mathcal{X} \subseteq \mathbb{R}^M \to \mathcal{Y} \subseteq \mathbb{R}^M\}$ is dense in the set of all functions in the pointwise convergence topology, then this family of mappings $\mathcal{G}$ together with a probability density function p_X with non zero support on $\mathcal{X}$ yields a universal generative model family on $\mathcal{Y}$ (cf Definition 2).*

Note: In conventional machine learning, this notion of universality is common. It has been proven for several families of mappings in the context of normalizing flows, which are mappings with the additional property of being invertible: generic triangular mappings [BKM05], neural networks mappings [HKLC18] and polynomial mappings [JSY19].

In the next section, to prove the universality of expectation value samplers, we make explicit the connection between universal mapping families and the known re-sults on the universality of parameterized quantum circuits as supervised learning models.

III. PARAMETERIZED QUANTUM CIRCUITS AS UNIVERSAL GENERATORS

A. Universality of the product encoding circuit

In recent literature, a number of universality properties of parameterized quantum circuits have been proven. In [PSLNGS+21], a family of single qubit quantum circuits with an increasing number of layers is proven to be universal in the uniform sense for continuous multi-dimensional input functions as complex coordinates of the quantum state in the computational basis (see Theorem 9). We extend this result to fit our needs.

In Appendix B we modify universality results from functions as coordinates in the computational basis to functions as the expectation value of an observable. In Appendix C we broaden the universality of quantum re-uploading models to some discontinuous functions by relaxing the required strength of convergence. More precisely we go from the uniform density in bounded continuous functions to the pointwise density in bounded piece-wise continuous functions. Finally, by stacking M universal circuits, we extend universality to multivariate output functions. All these extensions of [PSLNGS+21] together yield the theorem below.

Theorem 2. *For every natural number M, for every mapping $f \in \mathcal{B}([0,1]^M \to [-1,1]^M)$, there exists a sequence of sets of M single qubit quantum circuits and unit norm observables (indexed by k).*

$$\{(U_{k,m} : [0,1]^M \to \mathcal{U}(2), O_{k,m})_{1 \leq m \leq M}\}_{1 \leq k \leq \infty} \qquad (4)$$

such that the sequence of functions $\{g_k\}_{1 \leq k \leq \infty}$ defined as

$$g_{k,m}(x) = \langle 0| U_{k,m}(x)^\dagger O_{k,m} U_{k,m}(x) |0\rangle \qquad (5)$$

converges pointwise to f.
$\mathcal{B}$ is the set of piecewise continuous functions, and the norm of observable is the spectral norm.

The theorem above shows that there exists a family of M-qubit circuits with unit norm observables that yield a family of functions that is pointwise dense in the set of bounded piece-wise continuous functions. This matches the sufficient conditions of Theorem 1 for mappings to yield a universal generative model. This yields that there exists a family of expectation value sampling models as defined in Definition 1 and illustrated in Figure 2 that is universal in the sense of Definition 2.

Theorem 3. *For any M, for all M-dimensional probability density functions p_Z with support included in $[-1,1]^M$, and for all accuracy $\epsilon > 0$ there exists a M-qubit circuit U and set of M observables $\mathbf{O}$ with unit spectral norm $\|O_m\| = 1$ such that the expectation value*
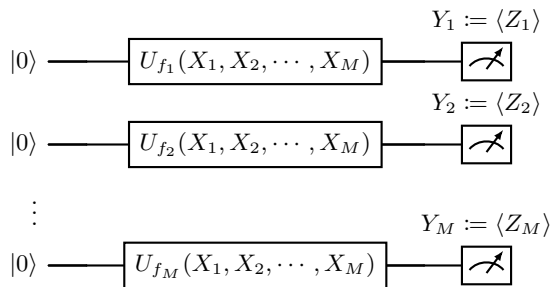
FIG. 2: Product Encoding Circuit as a universal generator, stacking circuits U_f approximating f from [PSLNGS+21]. It yields the random variable $Y = g(X), X \sim U([0,1]^M)$ with $f_m = \sqrt{(g_m + 1)/2}$.

sampling model $(U, \mathbf{O}, p_X)$ where p_X is the uniform distribution on $[0,1]^M$ yields a probability density function p_Y that is ϵ-close to p_Z in the Wasserstein distance.

This is the first central result of this work: n-qubit expectation value samplers with constant observable norm are universal for M-dimensional distributions with constant support radius, for $M = n$.

Note: We stated our result for distributions on the cube $[-1,1]^M$, but the universality straightforwardly generalizes. In particular, the size of the cube can be rescaled, by rescaling the norm of the observables. Finally, since any distribution with infinite support but finite moments can be arbitrarily approximated by a distribution with finite support, as we show in Appendix D, this means that if we allow the observable norm to scale, we can also approximate any distribution with finite moments (but perhaps infinite support).

B. Universality with less qubits

It is worth noting that the construction presented in the previous section uses the same number of qubits as the dimension of the output distribution. However, as we mentioned in the introduction, the dimension of the output does not need to be directly linked to the number of qubits used. This raises the question of the existence of a more qubit-frugal family of universal circuits, in which the output dimension is (much) larger than the qubit number. We will call such constructions "observable-dense expectation value samplers".

In Appendix E we show the existence of such a family if we allow for observables to have large norms. We formalize this in the following theorem and illustrate the corresponding circuit in figure Figure 3.

Theorem 4. *For any M, for all M-dimensional probability density functions p_Z with support included in $[-1,1]^M$, and for all accuracy $\epsilon > 0$ there exists a $n = \Theta(\log M)$-qubit circuit U taking L input variables and set of M observables $\mathbf{O}$ with spectral norm $\|O_m\| \in \Theta(M)$ such that the expectation value sampling model $(U, \mathbf{O}, p_X)$*
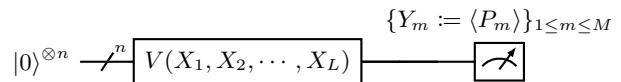

FIG. 3: Dense Encoding Circuit as a universal generator, based on a universal state preparation circuit V, with each parameterized gate replaced by a circuit from [PSLNGS+21]. $n = \log(M+1)$ and $P_m = 2M |m\rangle \langle m| - I$.

where p_X is the uniform distribution on $[0,1]^L$ yields a probability density p_Y that is ϵ-close to p_Z in the Wasserstein distance.

In this section, we have proven a number of *sufficient* universality conditions for expectation value samplers. In particular, we have shown the existence of two extremal families of parameterized quantum circuits that are universal generators on $[-1,1]^M$. There is a "product encoding" design, illustrated in Figure 2, with $n = M$ qubits and with unit norm observables (local Pauli), and a "dense encoding" design, illustrated in Figure 3 with $n = \log(M)$ qubits and M norm observables (amplified probabilities of bitstrings). The product encoding circuit has a large number of qubits for a constant observable norm, and the dense encoding circuit has a logarithmic number of qubits but large observable norms. This hints that there might be a trade-off between the number of measurements and the number of qubits. In the next section, we prove this is the case, by proving some *necessary* universality conditions.

IV. NECESSARY CONDITIONS FOR UNIVERSALITY

In this section, we use Holevo-like bounds to prove some necessary conditions on the number of qubits and the observable norm for an expectation value sampling model to be universal on $[-1,1]^M$.

A. Dimension of the observables space

As previously mentioned, an appealing feature of expectation value sampling models is that the dimension of the output vector is *a priori* independent of the number of qubits n, unlike in the case of quantum Born machines where each qubit corresponds to exactly one binary random variable. In particular, we can imagine using even just a single qubit with an arbitrary number of observables O_m to generate an M-dimensional random vector. However, it is obvious that in this case, the random variables corresponding to each observable cannot all be fully independent. Indeed, any observable can be expressed as a linear combination of the three Pauli matrices. Therefore the distribution output by a single qubit expectation value sampler will have at most three degrees of independence, and for $M > 3$ it is impossible to have

universality because it is impossible to approximate e.g. the 4-dimensional uniform distribution. Extending this reasoning to several qubits, we find the first necessary condition, that the dimension of the target dimension has to be lower or equal to the dimension of the space of observables. We formalize this in the following theorem.

Theorem 5. *For an n-qubit expectation value sampling model $(U_\theta(x), \mathbf{O}, p_X)$ to be able to approximate any distribution with support in $[-1, 1]^M$ to any accuracy $\epsilon > 0$, it is necessary that $M \leq 4^n - 1$.*

B. Holevo's bound

Another necessary condition for an expectation value sampling model to be universal on distributions with support in $[-1, 1]^M$ can be derived using a combination of Holevo's bound found in [ANTSV02] and Chernoff bound. We formalize it in the following theorem, which we prove in Appendix F.

Theorem 6. *For an n-qubit expectation value sampling model $(U_\theta, \mathbf{O}, p_X)$ to be able to approximate any distribution with support in $[-1, 1]^M$ to any accuracy $\epsilon > 0$ with respect to the Wasserstein distance, it is necessary that for every $m \leq M$:*

1. $\lambda_{\min}(O_m) \leq -1 + \epsilon$ and $\lambda_{\max}(O_m) \geq +1 - \epsilon$

2. $n \in \Omega\left(\frac{M(1-\epsilon)^2}{\|O_m\|^2}\right)$

with $\lambda_{\min/\max}(O)$ returning respectively the minimum and maximum eigenvalues of observable O.

The above necessary conditions use results which involve a more general case where we assume that prior to measurement we use a general parameterized quantum channel which can also prepare mixed states. However, we can reduce this to the special case of unitaries as well. By purification, mixed states can be mimicked by using $2n$ qubits pure states, since we are only interested in scalings, the factor 2 plays no role in the second condition of Theorem 6, and the necessary conditions also hold for pure states.

The combination of both previous necessary conditions is the second central result of this paper. It formalizes that even if expectation value sampling models may output arbitrary large dimensional distributions, in practice their expressivity is limited by the number of qubits and observables.

C. Asymptotic optimality of the product and dense encoding families

It is a natural question to ask whether the universal families we found in Section III saturate the necessary conditions found in the previous subsection. We illustrate such considerations in Figure 4.
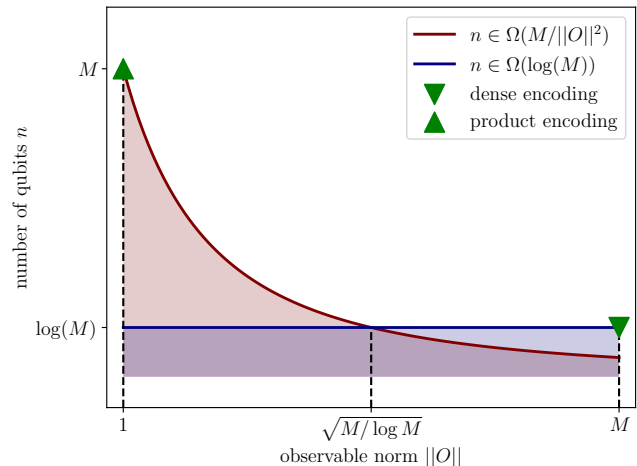

FIG. 4: Visual summary of results. We show the asymptotic use of resources for expectation value samplers to reach universality for a M-dimensional target distribution: the necessary conditions from Section IV A and Section IV B, as well as the existence of families from Section III A and Section III B.

Corollary 1. *The product encoding family defined in Section III A has a number of qubits $n \in \Theta(M)$ and an observable norm of 1. It is optimal with respect to the necessary condition from Section IV B.*

Corollary 2. *The dense encoding family defined in Section III B has a number of qubits $n \in \Theta(\log(M))$ and an observable norm of $\|O\| \in \Theta(M)$. It asymptotically saturates the condition from Section IV A. However, it is not optimal with respect to the condition from Section IV B, that yields $\|O\| \in \Omega(\sqrt{M/\log(M)})$, leaving a quasi-quadratic gap.*

We may conjecture that there exists a family of (Pareto) optimal circuits, balancing between observable norms and qubit numbers, with varying "encoding densities" in between the two extremal cases. In practice, the observable spectral norm relates to the number of measurements required to approximate it up to an additive accuracy. This highlights a trade-off between the number of qubits and the number of measurements, or space versus time complexity, which we formalize in the next subsection.

D. Observable norm and number of measurements

To estimate an expectation value up to a desired constant additive error, the number of measurements required is proportional to the norm of the observables. Therefore, large observable norms require a large number of measurements. This intuition is sharpened in the following lemma, proven in Appendix G.

Lemma 2. *An arbitrary expectation value sampler outputs an M-dimensional random vector $Y \sim p_Y$ with an*

infinite number of measurements, i.e. with access to exact expectation values. We consider the same circuit but with a finite number of measurements T that estimates expectation value by sampling and averaging for each observable and yields a random vector $\hat{Y} \sim p_{\hat{Y}}$. The number of measurements T required to guarantee that the Wasserstein distance between both distributions is smaller than ϵ satisfies

$$T \in \Theta \left(\frac{M\|O\|}{\epsilon^2} \right). \tag{6}$$

This lemma can be applied to both the product encoding circuit and the dense encoding circuit to show the trade-off between the number of qubits and the number of measurements.

E. On correlations in the target distribution

It is interesting to note that the parameter driving the necessary conditions is not the dimension of the target distribution itself M but rather the number of independent variables M'. Because we defined universality as being able to capture full independence we use $M = M'$ in the proofs of the necessary conditions. However, by restricting our interest to universality for classes of distributions with certain dependence conditions fulfilled, we can obtain much more economic bounds than for the general case, which is illustrated in Figure 4. We give two examples of universality for a restricted family of distributions.

First, consider the family of distributions that are M-dimensional but that have their support included in a 2-dimensional linear subspace. We choose the expectation value sampler as the circuit from [PSLNGS+21] and observables $\{\sqrt{2}X, \sqrt{2}Z\}$ we have a universal generator on one qubit for this 2-linear family for *any* target distribution dimension M.

Second, consider the family of M-dimensional Dirichlet distributions, for which each coordinate is positive, and such that the sum of all coordinates is unity. For this case, we can take the "dense encoding" circuit with $n = \log M$ qubits, see Figure 3, and instead of amplified probabilities we take the raw probabilities as observables. This constitutes a universal model for the Dirichlet family. In this case, we get away with exponentially fewer qubits than the general case.

These observations may be used to suggest that expectation value samplers are better suited to generate highly correlated distributions. Note that for the Dirichlet case, the result is not surprising as the normalization condition perfectly matches the normalization of quantum states, making this type of universality particularly well suited to expectation value samplers.

V. CHARACTERIZING EXPECTATION VALUE SAMPLING DISTRIBUTIONS

In this section, we propose additional tools to refine our understanding of expectation value sampling models, exploring how the choice in encoding, input distribution, and observables affects the expressivity.

A. Observable choice and primary mapping

In this subsection we explore the expressivity that comes with the choice in observables, considering expectation value models with fixed encoding $U(x)$ on n qubits and M-dimensional inputs. We are using the standard Pauli basis for the space of $2^n \times 2^n$ Hermitian operators $\mathbf{P}$. It is composed of all possible combinations of n Pauli matrices $\sigma_{0,1,2,3}$, which yields $|\mathbf{P}| = 4^n$. We formalize as follows

$$\mathbf{P} := (P_k, k \in \{0,1,2,3\}^n) \tag{7}$$
$$= (\otimes_{1 \leq i \leq n} \sigma_{k_i}, k_i \in \{0,1,2,3\}). \tag{8}$$

Any vector of M observables $\mathbf{O} = (O_m)$, can be expressed as a linear mapping applied on the vector of all Pauli strings: $\mathbf{O} = A\mathbf{P}$, where A is an $M \times 4^n$ matrix, and $\mathbf{P}$ is a 4^n dimensional vector. Therefore the distribution associated with the Pauli basis encompasses any distributions, which leads us to define the primary mapping as follows.

Definition 4 (Primary mapping). *The primary mapping g of an n-qubit encoding circuit $U_\theta(x)$ is defined as the mapping of the associated expectation value sampling model with the Pauli basis $\mathbf{P}$, defined as $(U_\theta(x), \mathbf{P}, p_X)$ according to Definition 1. It can be expressed as follows*

$$x \in [0, 2\pi)^N \xrightarrow{g} \left(\langle 0| U_\theta(x)^\dagger P_k U_\theta(x) |0\rangle \right)_{1 \leq k \leq 2^n}. \tag{9}$$

It yields the 4^n-dimensional random variable $Z = g(X)$ when $X \sim p_X$.

It is easy to see that any distribution obtained by an expectation value sampling model can always be expressed by considering an intermediary output of the 4^n set of observables, followed by the linear mapping A. We capture this idea in the following theorem.

Theorem 7. *Given an encoding circuit $U_\theta(x)$ and random variable with distribution p_X, for any the choice of observables $\mathbf{O}$, the expectation value sampling model $(U_\theta(x), \mathbf{O}, p_X)$ (according to Definition 1) is a linear transformation of the expectation value sampling model $(U_\theta(x), \mathbf{P}, p_X)$, where $\mathbf{P}$ is the Pauli basis per definition of the primary mapping in Definition 4.*

This concept of primary mapping has immediate consequences on the possible correlation of output variables of expectation value sampling models. For a given data encoding part on n qubits, the primary mapping will yield a

random variable Z with a covariance matrix C_z. Because any expectation value sampling model based on the same data encoding is a linear transformation of the primary mapping, the number of uncorrelated variables is limited by the number of non-null eigenvalues of the covariance matrix C_z, which is upper bounded in any case by $4^n - 1$.

Lemma 3. *Given an encoding circuit $U(x)$ over n qubits and random variable with distribution p_X, we note L the number of non-zero eigenvalues of the covariance matrix of the primary mapping. For any choice of observables, any expectation value model using $(U(x), p_X)$ will yield at most L uncorrelated variables. In addition $L \leq 4^n - 1$.*

B. Random variable encoding as a polynomial chaos expansion

After focusing on the choice of observables, in this subsection, we analyze the impact of the choice of the input random variable and the circuit encoding on the expressivity. We propose the polynomial chaos expansion [Xiu10] as a useful tool to analyze the expressivity of expectation value sampling models, as the analogue of the Fourier decomposition. The general polynomial chaos expansion is a representation of random variables as a vector in a Hilbert space of orthogonal functions, as defined below.

Definition 5 (Generalized Polynomial Chaos Expansion). *A generalized chaos expansion is characterized by a probability density function p_X defined on the support $\mathcal{X} \subseteq \mathbb{R}^M$ with finite moments (usually chosen as standard distributions, such as Gaussian or uniform). This choice defines an inner product for functions in $\{f : \mathcal{X} \to \mathbb{R}\}$:*

$$\langle f|g \rangle_{p_X} := \int_{\mathcal{X}} f^*(x)g(x)p_X(x)dx. \quad (10)$$

This choice of inner product comes with the choice of an ordered family of functions, usually polynomials, that are orthonormal with respect to the above inner product.

$$\Phi_{p_X} = \{\phi_l : \mathcal{X} \to \mathbb{R}, \forall(k,l), |\phi_l\rangle \langle \phi_k| = \delta_{k,l}\} \quad (11)$$

A generalized chaos expansion is a representation of a random variable Y with probability density p_Y as a vector α in this Hilbert space, such that:

$$Y = \sum_{l=0}^{\infty} \alpha_l \phi_l(X) \sim p_Y, X \sim p_X \quad (12)$$

This provides a Hilbert space as a potential structure to study random variable mappings. In particular, one of the main results of polynomial chaos expansion is that they are universal generative models.

Common pairs of distribution and associated orthogonal polynomials family can be found in Table I. In the context of expectation value sampling, we focus on the

Normal distribution	Hermite polynomials
Uniform distribution	Legendre polynomials
Exponential distribution	Laguerre polynomials
Beta distribution	Jacobi polynomials

TABLE I: Pairs of distributions and corresponding orthonormal families commonly used in General Polynomial Chaos expansion.

family of functions orthonormal with respect to the inner product associated with the uniform distribution on $\mathcal{X} = [0, 2\pi]^M$

$$\Phi = \{\prod_{1 \leq m \leq M} e^{ik_m x_m}, k \in \mathbb{Z}^M\} \quad (13)$$

Quantum reuploading circuits are a widely used class of parameterized quantum circuits that output a function of the data. They are used in a regressive context, where optimization techniques are used such that their output fits a target function. It has been widely studied and used that if they use integer-valued spectrum Hamiltonian, their output hypothesis function can be decomposed as an exact finite Fourier series [SSM21]. This means that there exists $c_{\mathbf{k},l} \in \mathbb{C}$ such that the hypothesis function f can be exactly written as a finite Fourier series:

$$f(x) = \sum_{k_0, k_1, \cdots, k_M = -K}^{+K} c_{\mathbf{k},l} \prod_{1 \leq m \leq M} e^{ik_m x_m}, . \quad (14)$$

This fact extends to a generative modelling context where expectation value sampling models using integer-valued quantum reuploading circuits yield distributions with an exact finite polynomial chaos expansion. We formalize this below.

Theorem 8. *Any expectation value sampling model using a quantum reuploading model with integer-valued spectrum $U_\theta(x)$, together with the uniform distribution on $[0, 2\pi]^M$ outputs a random variable Y that has an exact finite polynomial chaos expansion for any choice of observables.*

This subsection formalizes the tight connection between quantum circuits used in a regressive context with their use in a generative context. Therefore it is expected that, beyond the universality, many properties of such models can be transferred from a regressive context to a generative context, but we leave that for future work.

C. Choice of input random variable

We discuss the choice of input random variable, noted as p_X in Definition 1. We have shown several reasons to consider the uniform distribution as a good design choice as an input random variable, mostly based on the

fact that it has a bounded support. Indeed, it is known that the most commonly used parameterized quantum circuits are periodic, in particular when they have a finite Fourier decomposition. This is also why the universality of quantum reuploading circuits is proven for functions on bounded domains, corresponding to a half period. In contrast, let us consider an expectation value sampler $(U, \mathbf{O}, p_X)$, where the input random variable follows a Gaussian distribution $X \sim \mathcal{N}(0,1)$, which has unbounded support and for which the mapping f is 1-periodic. $X = 0$ is the highest probability event and will yield the exact same output as a very low probability event, for example, $X = 100$. This means that a very low probability input event and a very high probability input event will yield the same output sample, which is a feature rarely considered desirable. This choice of uniform distribution is in contrast to classical GANs where Gaussians are typically preferred.

VI. CONTRIBUTIONS AND DISCUSSION

The first central result of this work is that expectation value samplers, based on parameterized quantum circuits, are universal generative models. We provide constructive proofs for two extremal circuit designs. The "product encoding" design has a number of qubits linear in M and unit norm observables. In contrast, the "dense encoding" design has a number of qubits logarithmic in M and observable norm linear in M. The second central result is necessary conditions on resources such as the number of qubits and number of measurements for expectation value sampling models to be universal. This allows us to assess the optimality of the universal family we constructed. We conjecture there exists a series of universal circuits with varying "encoding densities" in between these two extremal cases, that are optimal with respect to the necessary conditions we have proven in this work. In addition, we propose additional tools to analyse the expressivity of expectation value samplers. We hope that the knowledge presented in this work will guide the choice of random variable encoding and observables in future designs.

While we characterized important properties of expectation value samplers, we did not address the question of whether it is a good idea to use them. Expectation value samplers cannot be proven to be a path for certain types of quantum advantage as easily as is the case in Quantum Circuit Born Machines [CMDK20]. Indeed, expectation value samplers, rather than requiring sampling from the full distribution of quantum measurements, are defined around expectation values only. Thus the hardness of simulating distributions from expectation value samplers does not connect straightforwardly to any hardness-of-sampling results established in the domain of quantum supremacy results [CMDK20].

Nonetheless, expectation value samplers still consist of genuinely quantum computations, and arguments for non-simulatability can be made. Assuming BQP is not in BPP, there exists no polynomial time algorithm that takes a classical description of an arbitrary expectation value sampler A as an input and outputs a sample from a distribution that is epsilon close to that of the output of A. Take as an example the case where a hard-to-simulate circuit does not depend on input data. This yields a Dirac delta distribution for which there exist classical samplers to efficiently sample from it, however, such classical samplers cannot be easily found based on the classical description of the expectation value sampler. It may be possible to construct stronger advantage arguments where we find the existence of an expectation value sampler such that its output distribution cannot be sampled by any polynomial-time randomized Turing machine (subject to standard assumptions) but we leave this for future work.

[Aar07] Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Phys-* *ical and Engineering Sciences*, 463(2088):3089–3114, September 2007. Publisher: Royal Society.

[AAR⁺18] Mohammad H. Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko. Quantum Boltzmann Machine. *Physical Review X*, 8(2):021050, May 2018. Publisher: American Physical Society.

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows*. Lectures in Mathematics ETH Zürich. Birkhäuser-Verlag, Basel, 2005.

[ANTSV02] Andris Ambainis, Ashwin Nayak, Amnon Ta-Shma, and Umesh Vazirani. Dense quantum coding and quantum finite automata. *Journal of the ACM*, 49(4):496–511, July 2002.

[Bai99] R. Baire. Sur les fonctions de variables réelles. *Annali di Matematica Pura ed Applicata (1898-1922)*, 3(1):1–123, December 1899.

[BCLK⁺22] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1):015004, February 2022. Publisher: American Physical Society.

[BKM05] V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, April 2005. Publisher: IOP Publishing.

[CAB⁺21] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, September 2021. Number: 9 Publisher: Nature Publishing Group.

[CKWD] Kevin Chen, Andrii Kurkin, Hao Wang, and Vedran Dunjko. in preparation.

[CMDK20] Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi. The Born supremacy: quantum advantage and training of an Ising Born machine. *npj Quantum Information*, 6(1):1–11, July 2020. Number: 1 Publisher: Nature Publishing Group.

[DDK18] Pierre-Luc Dallaire-Demers and Nathan Killoran. Quantum generative adversarial networks. *Physical Review A*, 98(1):012324, July 2018. Publisher: American Physical Society.

[HKLC18] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural Autoregressive Flows. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2078–2087. PMLR, July 2018. ISSN: 2640-3498.

[JGM⁺23] Sofiene Jerbi, Casper Gyurik, Simon C. Marshall, Riccardo Molteni, and Vedran Dunjko. Shadows of quantum machine learning, May 2023. arXiv:2306.00061 [quant-ph, stat].

[JSY19] Priyank Jaini, Kira A. Selby, and Yaoliang Yu. Sum of Squares Polynomial Flow, June 2019. arXiv:1905.02325 [cs, stat].

[KPB21] I. Kobyzev, S. D. Prince, and M. A. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(11):3964–3979, November 2021. Place: Los Alamitos, CA, USA Publisher: IEEE Computer Society.

[Lac84] M. Laczkovich. Baire 1 functions. *Real Analysis Exchange*, 9(1):15–28, January 1984. Publisher: Michigan State University Press.

[Leb04] H. Lebesgue. Une propriété caractéristique des fonctions de classe 1. *Bulletin de la Société Mathématique de France*, 32:229–242, 1904. Publisher: Société mathématique de France.

[LW18] Jin-Guo Liu and Lei Wang. Differentiable learning of quantum circuit Born machines. *Physical Review A*, 98(6):062324, December 2018. Publisher: American Physical Society.

[MCLS20] Antonio Macaluso, Luca Clissa, Stefano Lodi, and Claudio Sartori. A Variational Algorithm for Quantum Neural Networks. In Valeria V. Krzhizhanovskaya, Gábor Závodszky, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, Lecture Notes in Computer Science, pages 591–604, Cham, 2020. Springer International Publishing.

[PB11] Martin Plesch and Časlav Brukner. Quantum-state preparation with universal gate decompositions. *Physical Review A*, 83(3):032302, March 2011. Publisher: American Physical Society.

[PSLNGS⁺21] Adrián Pérez-Salinas, David López-Núñez, Artur García-Sáez, P. Forn-Díaz, and José I. Latorre. One qubit as a Universal Approximant. *Physical Review A*, 104(1):012405, July 2021. arXiv:2102.04032 [quant-ph].

[RAG21] Jonathan Romero and Alán Aspuru-Guzik. Variational Quantum Generators: Generative Adversarial Quantum Machine Learning for Continuous Distributions. *Advanced Quantum Technologies*, 4(1):2000003, 2021.

[SSM21] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, March 2021. Publisher: American Physical Society.

[Vil09] Cédric Villani. *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.

[Xiu10] Dongbin Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.

Appendix A: Constructive mapping between a random variable and the uniform distribution

Let us consider an absolutely continuous random variable Y with probability density function p_Y on a bounded set $[a, b]^M$. We recall a definition of an absolutely continuous variable below.

Definition 6. *A random variable X is said to be absolutely continuous if its cumulative distribution function (CDF) can be expressed as the integral of a non-negative function, known as the probability density function (PDF).*

This excludes for example Dirac deltas. We are going to construct an invertible mapping to transform the uniform random variable X into this random variable $Y = [Y_k]$. We call $G_1 : [0, 1] \to [a, b]$ the cumulative distribution function of the marginal of Y_1. It is invertible

and we define F_1 as its inverse,

$$Y_1 = F_1(X_1), X_1 = G_1(Y_1). \tag{A1}$$

Next we consider the marginal of Y_2 conditioned by Y_1, we define the cumulative distribution $G_2 : [a,b]^2 \to [0,1]$,

$$G_2(y_1, y_2) = P(Y_2 = y_2 | Y_1 = y_1). \tag{A2}$$

It is invertible with respect to Y_2,

$$G_2^{-1}(Y_1, X_2) = Y_2 \iff G_2(Y_1, Y_2) = X_2. \tag{A3}$$

We define $F_2 : [0,1]^2 \to [a,b]$ as follows,

$$F_2(X_1, X_2) = G_2^{-1}(F_1(X_1), X_2). \tag{A4}$$

We have $Y_2 = F_2(X_1, X_2)$. Continuing this process iteratively for all coordinates, it is possible to fully define the invertible mapping $F = [F_k]$ with inverse $G = [G_k]$ such that $Y = F(X)$. This is the essence of triangular mapping in [BKM05]. In addition, this mapping is bounded and piece-wise continuous, because it is composed of inverse cumulative distribution functions of absolutely continuous variables on bounded support.

Appendix B: From state coordinate universality to expectation of observable universality

We start by recalling Theorem 4 from [PSLNGS+21], which proves that the following circuit is universal. It has L layers, $\theta \in \mathbb{R}^{(M+2) \times L}$ parameters and for $x \in \mathbb{R}^M$ is defined as

$$U_\theta(x) := \prod_{l=1}^{L} R_y(\theta_{0,l}) \left(\prod_{m=1}^{M} R_z(x_m \theta_{m,l}) \right) R_z(\theta_{M+1,l}). \tag{B1}$$

Theorem 9. *For any pair of functions and real number*

$(f \in \mathcal{C}([0,1]^M \to [0,1]), \phi \in \mathcal{C}([0,1]^M \to [0, 2\pi)), \epsilon > 0)$

There exists a one qubit circuit $U : [0,1]^M \to \mathcal{U}(2)$ s.t.

$$\forall x, \left| \langle 1 | U(x) | 0 \rangle - f(x) e^{i\phi(x)} \right| < \epsilon. \tag{B2}$$

In this work, $\mathcal{C}$ is the set of continuous functions. This theorem yields the universality of functions embedded in a quantum state in the uniform sense. In the context of expectation value sampling, we are interested in the universality of function as the expectation value of a unit norm observable, captured by the following theorem.

Theorem 10. *For any function $g \in \mathcal{C}([0,1]^M \to [-1,1])$ and for any $\epsilon > 0$, there exists a one qubit circuit $U(x) : [0,1]^M \to \mathcal{U}(2)$ and an observable O with unit spectral norm $\|O\| = 1$ s.t.*

$$\forall x, \left| \langle 0 | U^\dagger(x) O U(x) | 0 \rangle - g(x) \right| < \epsilon. \tag{B3}$$

Proof: We are given an arbitrary function $g \in \mathcal{C}([0,1]^M \to [-1,1])$ and $\epsilon > 0$. We define the function $f = \sqrt{\frac{g+1}{2}}$ which is well defined on $\mathcal{C}([0,1]^M \to [0,1])$. We apply Theorem 9 to $(f, \phi = 0, \epsilon/4)$ and get a circuit U that yields a state close to

$$|x\rangle = \sqrt{1 - f(x)^2} |0\rangle + f(x) |1\rangle. \tag{B4}$$

The Z expectation value of the above state is

$$\langle x | Z | x \rangle = 2f(x)^2 - 1 = g(x). \tag{B5}$$

We are now going to prove that the expectation value is close to the target function g. First, we note that the square function is 2-Lipschitz on $[0,1]$ and therefore for every pair of real numbers $(x, y) \in [0,1]^2$

$$|x - y| < \epsilon \implies |x^2 - y^2| < 2\epsilon. \tag{B6}$$

We define $p_{0/1}$ the probabilities of measuring $U(x) |0\rangle$ in state $|0\rangle$ and $|1\rangle$ respectively. Recalling that

$$|\langle 1 | U(x) | 0 \rangle - f(x)| < \epsilon/4, \tag{B7}$$

we can write

$$\left| p_1 - f(x)^2 \right| < \epsilon/2 \tag{B8}$$

$$\left| p_0 - (1 - f(x)^2) \right| =$$
$$\left| 1 - |\langle 1 | U(x) | 0 \rangle|^2 - (1 - f(x)^2) \right| < \epsilon/2. \tag{B9}$$

Finally,

$$|\langle Z \rangle - g(x)| \leq \left| p_1 - f(x)^2 \right| + \left| p_0 - (1 - f(x)^2) \right| < \epsilon. \tag{B10}$$

This yields the uniform density of quantum functions in the set of bounded continuous functions.

Appendix C: From uniform density on continuous functions to pointwise density on discontinuous functions

We first start by highlighting the difference between uniform convergence and pointwise convergence and illustrate it with an example.

Definition 7 (Pointwise convergence). *Let $\{f_k | f_k : \mathcal{X} \to \mathbb{R}\}_{1 \leq k \leq \infty}$ be a sequence of functions, and let $f : \mathcal{X} \to \mathbb{R}$ be another function defined on the same domain $\mathcal{X}$. We say that the sequence $\{f_k\}$ converges pointwise to f if, for each $x \in \mathcal{X}$, the sequence of real numbers $\{f_k(x)\}_{1 \leq k \leq \infty}$ converges to $f(x)$ as k approaches infinity,*

$$\lim_{n \to \infty} f_k(x) = f(x) \quad \text{for all } x \in \mathcal{X}.$$

Definition 8 (Uniform convergence). *Let $\{f_k | f_k : \mathcal{X} \to \mathbb{R}\}_{1 \leq k \leq \infty}$ be a sequence of functions, and let $f : \mathcal{X} \to \mathbb{R}$ be another function defined on the same domain $\mathcal{X}$. We say that the sequence f_k converges uniformly to f if, for*
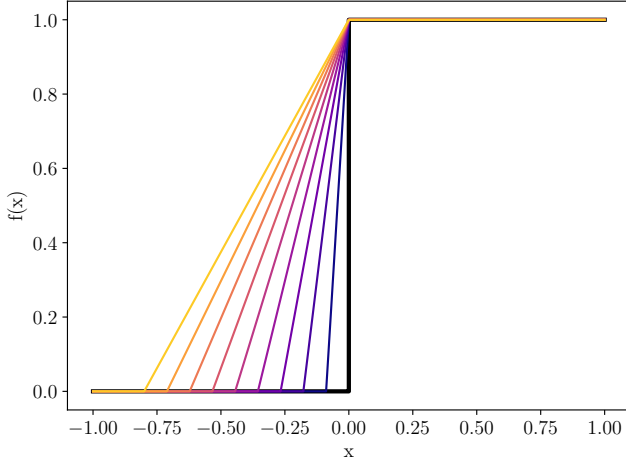
FIG. 5: A sequence of continuous functions converging pointwise but not uniformly to the step function.

any given $\epsilon > 0$, there exists an $K \in \mathbb{N}$ such that for all $k \geq K$ and for all $x \in \mathcal{X}$, the difference $|f_k(x) - f(x)|$ is less than ϵ,

$$\forall \epsilon > 0, \exists K \in \mathbb{N} : \forall k \geq K, \forall x \in \mathcal{X}, |f_k(x) - f(x)| < \epsilon.$$

Uniform convergence is stronger, it implies pointwise convergence, but the reverse is not true. For example, consider the step function f.

$$f(x) = \begin{cases} -1, & \text{if } x \in [-1, 0[\\ +1, & \text{if } x \in [0, +1] \end{cases} \tag{C1}$$

It is impossible to define a sequence of continuous functions that would uniformly converge to it, however, it is possible to have a sequence of continuous functions that converges pointwise to it, see Figure 5.

$$f_k(x) = \begin{cases} -1, & \text{if } x \in [-1, 1/k[\\ 1 + kx, & \text{if } x \in [-1/k, 0] \\ +1, & \text{if } x \in]0, +1] \end{cases} \tag{C2}$$

In [Bai99] Baire defined hierarchical pointwise convergence classes of functions. Baire class 0 is the set of continuous functions, and the class c is the set of functions that are the pointwise limit of class $c - 1$. In particular in [Lac84, Leb04], it was proven that bounded piecewise continuous functions are of class 1. This means that for any bounded piece-wise continuous function f there exists a sequence of bounded continuous functions that converges pointwise to f. This means that bounded continuous functions are dense in bounded piece-wise continuous functions in the pointwise topology. Therefore, building on Theorem 10 we have the following theorem, writing $\mathcal{B}$ as the set of piecewise continuous functions.

Theorem 11. *For any function $f : \mathcal{B}([0,1]^M \to [-1,1])$, there exists a sequence (indexed by k) of one qubit circuit*

and observables with unit spectral norm,

$$\{(U_k(x) : [0,1]^M \to \mathcal{U}(2), O_k)\}_{1 \leq k \leq \infty} \tag{C3}$$

such that the sequence of functions $\{g_k\}$ with

$$g_k(x) = \langle 0| U_k^\dagger(x) O_k U_k(x) |0\rangle \tag{C4}$$

converges pointwise to f.

Appendix D: Notes on infinite support

We consider a distribution with probability density function p_Y with infinite support $\mathbb{R}^M$ and finite moments. The goal is to find a sequence of random variables with bounded support probability density function that converges pointwise to p_y. We define the sequence of random variables $\{Y_k\}_{1 \leq k \leq \infty}$ whose probability density functions p_{Y_k} are proportional to that of Y on the hypercube $[-k - k_0, +k + k_0]^M$, where k_0 is the first integer such that Y has non-null support on the hypercube. This sequence converges in distribution to Y.

Appendix E: Construction of the "dense encoding" circuit

Note: We have used slight abuse of notations in the below proof to increase readability, specifically in the approximations.

Considering architecture such as in [PB11] and universal one qubit gate as $R_x(\alpha)R_z(\beta)R_x(\gamma)$, for any number of qubits, it is possible to design a circuit $U : [0, 2\pi)^L \to \mathcal{U}(2^n)$ made only of a finite number of fixed gates (CNOT and constant rotations) and parameterized σ_z rotations gates that can reach any pure quantum state when applied to state $|0\rangle$,

$$\forall |\psi\rangle, \exists \theta \in [0, 2\pi)^L, |\psi\rangle = U(\theta) |0\rangle. \tag{E1}$$

Let's consider a distribution p_ψ over pure states. Because the architecture above can reach any pure state, there exists a corresponding distribution over the parameters p_θ such that the distribution $V(\theta) |0\rangle, \theta \sim p_\theta$ matches perfectly p_ψ in distribution. We define $g : [0,1]^L \to [0, 2\pi)^L$ as the mapping that transforms the uniform distribution over $[0,1]^L$ into p_θ. We have $V(f(X)) |0\rangle, X \sim U([0,1]^L)$ matches perfectly p_ψ in distribution, where V is composed of a finite number of fixed gates and L σ_z rotation gates parameterized by $g_l(X)$.

Lemma 4. *For any distribution over pure states p_ψ, there exists a circuit architecture V made of constant gates and parameterized σ_z rotations, and a mapping g such that*

$$X \sim U([0,1]^L), V(g(X)) |0\rangle \sim p_\psi \tag{E2}$$

Next, we decompose $R_z \circ g$ gates into a sequence of constant gates and parameterized σ_z rotations.

From [PSLNGS$^+$21], in the proof in the appendix, it is shown that there exists a quantum circuit W taking a multidimensional input and approximating the following parameterized quantum gate,

$$\forall f : [0,1]^L \to [0,1], \phi : [0,1]^L \to [0,2\pi], \exists W$$

$$W(x) \equiv_\epsilon \begin{bmatrix} \sqrt{1 - f(x)^2}e^{+i\phi(x)} & -f(x)e^{+i\phi(x)} \\ f(x)e^{-i\phi(x)} & \sqrt{1 - f(x)^2}e^{-i\phi(x)} \end{bmatrix}. \tag{E3}$$

Choosing $\phi = 0$, and $f = \sin g$, we have $\sqrt{1 - f^2} = \cos g$, and we note that $W(x) = R_z \circ g$. We can conclude the following.

Lemma 5.

$$\forall f : [0,1]^L \to [0,1], \exists W, R_z \circ f \equiv_\epsilon W, \tag{E4}$$

where W is a quantum circuit made of constant gates and R_z gates applied to individual components of x.

Combining both above lemmas we get the following.

Lemma 6. For any distribution over pure states p_ψ, there exists a circuit architecture V made of constant gates and parameterized z rotations such that

$$X \sim U([0,1]^L), V(X)|0\rangle \sim_\epsilon p_\psi. \tag{E5}$$

We are now going to use that lemma to prove that n-qubits expectation value samplers are universal for $\exp(n)$-dimensional distributions with constant support if the observables are allowed to have $\exp(n)$ norms.

We are given an arbitrary random variable Y following a M-dimensional distribution p_Y with support $[-1,1]^M$. We define the following state over n qubits with $M = 2^n - 1$.

$$|\psi(Y)\rangle = \sum_{m \leq M} Z_m |m\rangle + \sqrt{1 - \sum_{m \leq M} Z_m^2} |M+1\rangle \tag{E6}$$

$$Z_m = \sqrt{\frac{Y_m + 1}{2M}} \tag{E7}$$

We define as p_ψ as the probability density functions over states when $Y \sim p_Y$, using the previous lemma we get a circuit W composed only of constant gates and z rotations gates with one of the L parameters as input that approximates p_ψ. We define the observables $\forall m \leq M, O_m = 2M |m\rangle\langle m| - I$. They have spectral norm $\|O_m\| = 2M - 1 = \Theta(2^n)$. In addition, $\langle\psi(Y)| O_m |\psi(Y)\rangle = Y_m$. Therefore, the expectation value sampler $(W, O, U([0,1]^L))$ approximates p_Y. This concludes the proof to Theorem 4.

Appendix F: Proof of theorem in Section IV B

We start by proving the Theorem 6, which we recall below.

Theorem. For an n-qubit expectation value sampling model $(U_\theta, \mathbf{O}, p_X)$ to be able to approximate any distribution with support in $[-1,1]^M$ to any accuracy $\epsilon > 0$ with respect to the Wasserstein distance, it is necessary that for every $m \leq M$:

1. $\lambda_{\min}(O_m) \leq -1 + \epsilon$ and $\lambda_{\max}(O_m) \geq +1 - \epsilon$

2. $n \in \Omega\left(\frac{M(1-\epsilon)^2}{\Lambda(O_m)}\right) \subseteq \Omega\left(\frac{M(1-\epsilon)^2}{\|O_m\|^2}\right)$

with $\lambda_{\min/\max}(O)$ returning respectively the minimum and maximum eigenvalues of observable O, and $\Lambda(O) := -\lambda_{\min}(O)\lambda_{\max}(O)$.

Proof: Let's suppose there is an n-qubit expectation value scheme with M observables $\mathbf{O}$ that is able to approximate any distributions with support included in $[-1,1]^M$ to ϵ with respect to the first Wasserstein distance W_1.

Because the expectation value model is universal, it means that for any vertex of the hypercube $c \in \{-1,1\}^M$, it can approximate the Dirac delta at c. We then use the lemma below.

Lemma 7. Any distribution that is ϵ-close (in the Wasserstein distance) to the Dirac delta at a given point must have nonzero support within the Euclidean distance sphere centred in that point with radius ϵ.

This means that there is non-zero support on the ϵ sphere around c, and therefore there exists a quantum state ρ_c whose list of expectations is ϵ-close to that point. This yields the following result.

$$\forall c \in \{-1,1\}^M, \exists \rho_c, \sum_{m=1}^M (\text{Tr}(O_m\rho_c) - c_m)^2 \leq \epsilon^2. \tag{F1}$$

Because the above is a sum of positive components, we can write $\forall m$

(a) if $c_m = 0$, then $-1 - \epsilon \leq \text{Tr}(O_m\rho_c) \leq -1 + \epsilon$

(b) if $c_m = 1$, then $+1 - \epsilon \leq \text{Tr}(O_m\rho_c) \leq +1 + \epsilon$

The above yields conditions on the spectrum of O_m. We note $\lambda_{\min}$ and $\lambda_{\max}$ respectively the minimum and maximum eigenvalues of O_m. We define $\gamma := 1 - \epsilon$. We know that $\forall \rho, \text{Tr}(O\rho) \geq \lambda_{\min}$ therefore, $-\gamma \geq \lambda_{\min}$, the same reasoning applies for the maximum eigenvalue, yielding:

(a) $\lambda_{\min} \leq -\gamma$,

(b) $\lambda_{\max} \geq +\gamma$.

For the rest of the proof, we combine approaches from the proof of theorem 2.6 in [Aar07] and that of B.1 in [JGM+23]. We define the two-outcome POVMs $\{E_m, I - E_m\}$ with

$$E_m = \frac{O_m - \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \qquad (F2)$$

We define $\beta := \frac{-\lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}$. The spectral inequalities yield $\beta \geq \frac{1}{2}$. With this definition, the two conditions above translate to:

(a) $c_m = 0, p_0 := \text{Tr}(E_m \rho_c) \leq \beta - \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}$

(b) $c_m = 1, p_1 := \text{Tr}(E_m \rho_c) \geq \beta + \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}$

We define the amplified Positive Operator-Valued Measures (POVMs) that apply $\{E_m, I - E_m\}$ to $L \geq 1$ copies of ρ and return 1 if and only if at least βL copies of the original POVMs return 1.

From Holevo's bound (found as Theorem 5.1 in [ANTSV02]), for the amplified scheme to correctly identify the corresponding bit $b_m = (c_m + 1)/2$ with probability q it is necessary that

$$nL \geq (1 - H(q))M, \qquad (F3)$$

where H is the binary entropy function.

We define the random variable $X_{m,l}$ which takes the value of the output of the POVM of the m-th observable on the l-th copy. We define $X_m^{(L)} := \frac{1}{L} \sum_l X_{m,l}$.

In the case $c_m = -1$, we have $\mathbb{E}[X_m^{(L)}] = p_0$. The probability of the amplified POVMs yielding the wrong output is

$$P(X_m^{(L)} > \beta) \leq P\left(X_m^{(L)} > p_0 + \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}\right) \quad (F4)$$

Recalling that $\beta \geq 1/2$, we can use the Chernoff bound on the Bernoulli variable $X_m^{(L)}$ and we get

$$P(X_m^{(L)} > p_0 + \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}) \leq \exp\left\{-\frac{\gamma^2 L}{2(-\lambda_{\min})\lambda_{\max}}\right\}$$
$$(F5)$$

We define $\Lambda = (-\lambda_{\min})\lambda_{\max}$, we have $\gamma^2 \leq \Lambda \leq \|O\|^2$.

For the probability of the amplified POVMs to yield the correct output with probability q it is necessary that $P(X_m^{(L)} \leq \beta) \geq q$.

Finally, we get

$$\log(1/q) \leq \frac{\gamma^2 L}{2\Lambda} \qquad (F6)$$

Combining Chernoff's and Holevo's inequalities, we conclude the proof of Theorem 6:

$$\Lambda \geq \frac{1 - H(q)}{\log(1/q)} \frac{\gamma^2 M}{n}. \qquad (F7)$$

Note: The above is a tighter condition than in Theorem 2.6 in [JGM+23] but falls back to it, when $\Lambda = \|O\|^2$,

which corresponds to $\lambda_{\min} = -\|O\|$ and $\lambda_{\max} = \|O\|$. In the opposite scenario, we have $\Lambda = \gamma^2$, which corresponds to constant norm observables, yielding $n \in \Omega(M)$. The norm of observables affects the number of measurements to reach a desired additive accuracy.

Appendix G: Approximation of expectation values

In practice, we do not have access to exact expectation values and we have to estimate them through sampling. This creates a distribution $p_{\hat{Y}}$ (n dimensional) slightly different from the distribution with exact expectation values p_Y. For simplicity, we assume that shot noise $\rho_\epsilon \sim \mathcal{N}(0, \epsilon I_n)$ (n dimensional) acts like an additional Gaussian noise with zero mean and standard deviation ϵ. ρ_ϵ and Y are independent and we consider the random variable $\hat{Y} = Y + \rho_\epsilon$. Therefore the density $p_{\hat{Y}}$ is the convolution of p_Y and p_{ρ_ϵ}. Using Lemma 7.1.10 from [AGS05], we know that the Wasserstein distance between p_Y and $p_{\hat{Y}}$, $W_p(p_Y, p_{\hat{Y}}) \in O(\epsilon)$.

In practice, different techniques exist to estimate expectation values of observables with different degrees of measurement efficiency, with shadow tomography techniques [CKWD] surpassing the "vanilla estimation". For simplicity, we consider a vanilla estimation where each observable with norm $\|O\|$ is measured t times and the average is returned. This yields a shot noise close to the Gaussian model above with $\epsilon^2 \in \Theta(\|O\|/t)$. The total number of measurements T is then $T = tM$, which yield $T \in \Theta(M\|O\|/\epsilon^2)$

Note that the first result cannot be trivially applied to techniques such as shadow tomography [CKWD]. Indeed the corresponding shot noise ρ cannot in general be modelled by a Gaussian independent noise, at the least the covariance matrix will in general not be proportional to the identity.