

Learning anisotropic kernels for quantum generative modeling with the maximum mean discrepancy

Oriol Balló-Gimbernat^{1,2,3}, Emiliano Tolotti⁴, Laia Domingo^{2,3}, Adan Garriga¹, Fernando Vilariño^{2,3}

¹Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain

²Centre de Visió per Computador (CVC), Barcelona, Spain

³Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

⁴Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Quantum generative models, such as quantum circuit Born machines (QCBMs), are often trained via implicit loss functions, as explicit likelihood-based objectives such as KL divergence can suffer from severe trainability barriers [1]. Among such losses, the maximum mean discrepancy (MMD) is one of the most commonly used [2]. It compares a target distribution p and a model distribution q based entirely on samples from each. The most common way to compute an unbiased estimate of the squared MMD is to sample batches of vectors $\mathcal{X} \sim p$ and $\mathcal{Y} \sim q$ and use the estimator from Gretton et al. [3], a standard tool in optimized-MMD generative modeling [4]:

$$\widehat{\text{MMD}}^2(\mathcal{X}, \mathcal{Y}) = \frac{1}{|\mathcal{X}|(|\mathcal{X}|-1)} \sum_{i \neq j} k(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{|\mathcal{X}||\mathcal{Y}|} \sum_{i,j} k(\mathbf{x}_i, \mathbf{y}_j) + \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)} \sum_{i \neq j} k(\mathbf{y}_i, \mathbf{y}_j). \quad (1)$$

This estimator can be shown to be unbiased, meaning its expected value coincides exactly with the true MMD². The MMD evaluates these discrepancies by computing the expected value of the kernel k across the sampled distributions. Consequently, the choice of kernel determines which features of the distributions contribute to the measured discrepancy. A standard choice is the Gaussian kernel,

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right), \quad (2)$$

which evaluates similarity based on the squared Euclidean distance between samples. Since the outputs of a n -qubit quantum circuit measured in the computational basis are bitstrings $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$, the data naturally resides on an n -dimensional Boolean hypercube. In this discrete space, the squared Euclidean distance reduces to the Hamming distance

$$\|\mathbf{x} - \mathbf{y}\|^2 = \sum_{i=1}^n (x_i - y_i)^2 = d_H(\mathbf{x}, \mathbf{y}). \quad (3)$$

Consequently, the Gaussian kernel evaluated on the Boolean hypercube takes the form

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{d_H(\mathbf{x}, \mathbf{y})}{2\sigma^2}\right). \quad (4)$$

Because this kernel depends on the Hamming distance, it is isotropic: it sees how many coordinates differ, but not which coordinates differ. Its Walsh-Fourier expansion makes this limitation precise. Notably, the MMD with the Gaussian kernel can be represented as a classical mixture of Pauli-Z expectation values, which makes this loss suitable to train QCBMs [1]. Defining $\beta = 1/(4\sigma^2)$ and $\gamma = \tanh(\beta)$, one obtains

$$k(\mathbf{x}, \mathbf{y}) = C_{\beta,n} \sum_{S \subseteq [n]} \gamma^{|S|} \prod_{i \in S} z_i, \quad (5)$$

where $z_i = (-1)^{x_i \oplus y_i}$ and $C_{\beta,n} = (e^{-\beta} \cosh \beta)^n = (1 + \gamma)^{-n}$ is a positive normalization constant. Equation (5) highlights two important concepts. First, the Gaussian kernel is isotropic; all ℓ -order differences are weighted the same since they share the same γ^ℓ factor, regardless of which specific coordinates differ. Second, the σ parameter determines the value of the weighting factor $\gamma = \tanh(\frac{1}{4\sigma^2})$. A large σ induces a small γ , with higher-order terms vanishing. A small σ induces a γ close to 1;

higher-order terms remain close in weight to lower-order ones, with the total spectral mass across all subsets of a given degree concentrating around $\ell = n/2$.

Instantaneous quantum polynomial-time (IQP) circuits occupy a prime position for near-term applications, they are conjectured to be hard to sample from classically [5–7], yet their specific algebraic structure permits the efficient classical calculation of certain output statistics that can be used for training, such as Pauli-Z observables [8–10]. For IQP families, this enables MMD training through classically accessible Pauli-Z observables. Because the Gaussian kernel treats all same-degree parities identically, this finite operator budget is spread uniformly across parity directions regardless of their relevance to the target distribution.

We propose Sparse Parity Search (SPS), a method for allocating MMD operator budget to relevant observables. The starting point is the Walsh-Fourier form of kernels on $\{0, 1\}^n$. For a parity mask $S \subseteq [n]$, define $\chi_S(x) = (-1)^{\sum_{i \in S} x_i}$. Any nonnegative Fourier spectrum gives a positive semidefinite parity kernel

$$k_A(x, y) = \sum_{S \subseteq [n]} a_S \chi_S(x) \chi_S(y), \quad a_S \geq 0, \quad (6)$$

and its squared MMD has the exact observable decomposition

$$\text{MMD}_A^2(p, q) = \sum_{S \subseteq [n]} a_S (\hat{p}(S) - \hat{q}(S))^2. \quad (7)$$

Equation (10) turns kernel design into observable allocation. The practical question becomes whether the finite operator budget measures the gaps between target and model Fourier coefficients often enough to guide optimization, rather than only whether the kernel is characteristic in the infinite-sample limit. This finite-spectrum view aligns with classical random-feature and sparse-spectrum kernel approximations [11, 12].

SPS discovers candidate target-relevant parities by using a Goldreich–Levin-inspired search over the subset lattice [13]. Given target samples $x_1, \dots, x_m \sim p$, we estimate empirical parity coefficients

$$\hat{c}_p(S) = \frac{1}{m} \sum_{j=1}^m \chi_S(x_j). \quad (8)$$

A parity mask S is considered relevant when its empirical coefficient has large magnitude, operationally when $|\hat{c}_p(S)|$ exceeds a threshold τ . Finite-sample uncertainty is used to score and prune branches during the search. The search starts from one-bit masks and expands candidate prefixes in best-first order. Each branch is scored using a noise-corrected lower bound on the largest coefficient found among a bounded set of sampled descendants, while branches whose confidence-adjusted upper bound falls below τ are pruned. With an expansion cap E and at most B descendant probes per branch, SPS performs at most $\mathcal{O}(EBm)$ parity evaluations, far below exhaustive enumeration over 2^n masks when $EB \ll 2^n$, and adds no quantum sampling overhead.

The retained support $\mathcal{C}_\tau$ is converted into a characteristic anisotropic MMD kernel by mixing a focused target proposal with an isotropic spectral floor. The discovery primitive follows the sparse Fourier learning literature [14–16], but we use it for kernel construction rather than for learning a Boolean predictor. On the discovered support, SPS forms

$$r_{\text{SPS}}(S) = \frac{\hat{c}_p(S)^2}{\sum_{T \in \mathcal{C}_\tau} \hat{c}_p(T)^2}, \quad S \in \mathcal{C}_\tau, \quad (9)$$

and sets $r_{\text{SPS}}(S) = 0$ for $S \notin \mathcal{C}_\tau$. This proposal assigns more mass to discovered parities with larger empirical target bias, including deterministic or nearly deterministic constraints. The final spectrum is

$$a_S = (1 - \lambda)\pi_\sigma(S) + \lambda r_{\text{SPS}}(S), \quad 0 \leq \lambda < 1, \quad (10)$$

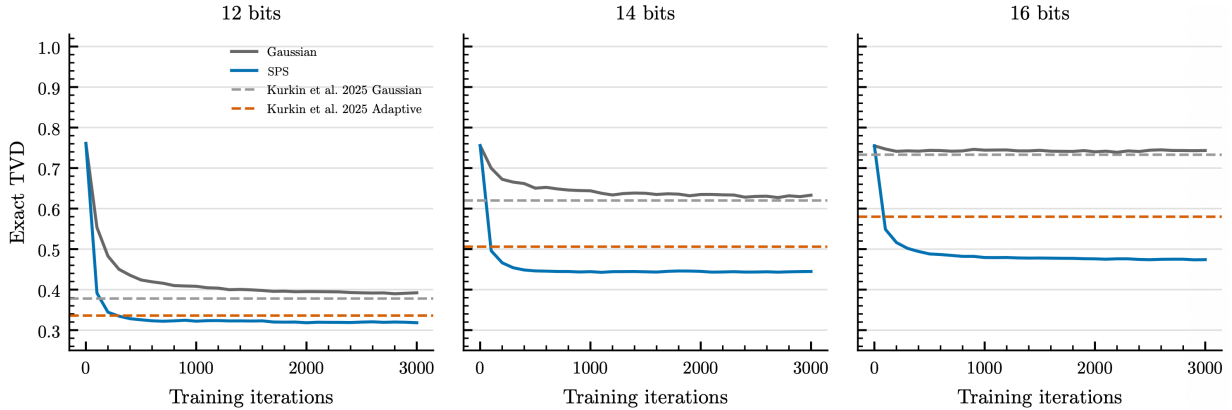


Figure 1: Exact TVD on the synthetic parity-check benchmark. Gaussian and SPS curves show mean TVD across five fixed-protocol repeats under the training protocol from Kurkin et al. [17]. The MMD operator budget is fixed 1000 in all cases. Horizontal lines denote the values reported by Kurkin et al. [17] for the Gaussian kernel and adaptive-FVSN, their proposed solution.

where π_σ is the normalized isotropic Gaussian spectrum. The learned term concentrates MMD mass on target-specific parities, while the isotropic floor keeps $a_S > 0$ for every mask. Thus the population kernel remains characteristic. The sampled training objective then uses this spectrum to spend a finite operator budget on discovered directions rather than treating all same-degree parities identically.

We evaluate SPS on the synthetic parity-check benchmark of Kurkin et al. [17], where target samples are drawn uniformly from the nullspace of a binary parity-check matrix $H \in \{0, 1\}^{2 \times n}$. Relative to the uniform distribution on $\{0, 1\}^n$, the nontrivial Fourier support consists of the two parity-check rows and their XOR. This makes the task a clean test of whether a kernel can allocate finite operator budget to a small set of high-magnitude Fourier directions. We match the reference protocol: degree-6 IQP generator, data-based initialization, Adam optimization for 3000 steps, 1000 IQP expectation samples, exact TVD evaluation every 100 steps, and fixed MMD operator budget 1000.

The parity-check results support the intended mechanism. Figure 1 shows that SPS improves exact TVD over the local fixed Gaussian baseline at all three bit widths, from 0.392 to 0.318 for $n = 12$, 0.633 to 0.444 for $n = 14$, and 0.743 to 0.474 for $n = 16$, all at budget 1000. These also fall below the reported adaptive-FVSN references of 0.336, 0.506, and 0.580. Notably, the adaptive-FVSN kernel changes the spectral measure during training via adversarial learning, while SPS performs pre-training target-sample discovery and then trains with a fixed anisotropic characteristic kernel. Upon inspection of the selected operators, we find that SPS recovers the exact three-mask Fourier support during the operator discovery phase, which explains why the method is effective in this setting. However, SPS is not meant to be a universal replacement for isotropic kernels; it is most useful when the target distribution has clear coordinate-specific structure that an isotropic spectrum undersamples.

We further evaluate SPS on the 8-node bipartite graph datasets from Balló-Gimbernát et al. [18], using the same shallow IQP ansatz and training protocol. SPS raises the generated bipartite percentage from 44–77% (isotropic MMD baseline) to 71–81% across all three 28-qubit datasets, confirming that observable allocation improves global structural feature learning beyond isotropic MMD.

In practice, the MMD loss exposes the model to a limited subset of Pauli-Z observables. This makes the kernel design an observable allocation problem. SPS provides an interpretable way to target relevant observables from samples while preserving characteristicness through an isotropic floor. The evidence is strongest for sparse parity constraints; it also highlights that discovering useful constraints is a necessary but not sufficient step. Future work should combine pre-training discovery with training-time spectral adaptation and generator changes matched to discovered observables, connecting naturally to recent work on generative quantum advantage [19] and IQP-based generative modeling [18, 20–22].

References

- [1] Manuel S. Rudolph, Sacha Lerch, Supanut Thanasilp, Oriel Kiss, Oxana Shaya, Sofia Vallecorsa, Michele Grossi, and Zoë Holmes. Trainability barriers and opportunities in quantum generative modeling. *npj Quantum Information*, 10(1):116, November 2024.
- [2] Jin-Guo Liu and Lei Wang. Differentiable Learning of Quantum Circuit Born Machine. *Physical Review A*, 98(6):062324, December 2018.
- [3] 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.
- [4] Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy, January 2021.
- [5] Dan Shepherd and Michael J. Bremner. Instantaneous Quantum Computation. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 465(2105):1413–1439, May 2009.
- [6] 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, August 2016.
- [7] Simon C. Marshall, Scott Aaronson, and Vedran Dunjko. Improved separation between quantum and classical computers for sampling and functional tasks, October 2024.
- [8] M. Van den Nest. Simulating quantum computers with probabilistic methods, April 2010.
- [9] Erik Recio-Armengol and Joseph Bowles. IQPopt: Fast optimization of instantaneous quantum polynomial circuits in JAX, January 2025.
- [10] Erik Recio-Armengol, Shahnawaz Ahmed, and Joseph Bowles. Train on classical, deploy on quantum: Scaling generative quantum machine learning to a thousand qubits, March 2025.
- [11] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.
- [12] Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R. Figueiras-Vidal. Sparse spectrum gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.
- [13] Oded Goldreich and Leonid A. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 25–32, New York, NY, USA, 1989. Association for Computing Machinery.
- [14] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM*, 40(3):607–620, 1993.
- [15] Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993.
- [16] Ryan O’Donnell. *Analysis of Boolean Functions*. Cambridge University Press, Cambridge, 2021.
- [17] Andrii Kurkin, Kevin Shen, Susanne Pielawa, Hao Wang, and Vedran Dunjko. Universality and kernel-adaptive training for classically trained, quantum-deployed generative models, October 2025.

- [18] Oriol Balló-Gimbernat, Marcos Arroyo-Sánchez, Paula García-Molina, Adan Garriga, and Fernando Vilarino. Shallow instantaneous quantum polynomial-time circuits for generative modeling on noisy intermediate-scale quantum hardware. *Physical Review A*, 113:042617, 2026.
- [19] Hsin-Yuan Huang, Michael Broughton, Norhan Eassa, Hartmut Neven, Ryan Babbush, and Jarrod R. McClean. Generative quantum advantage for classical and quantum problems, September 2025.
- [20] Xiaocheng Zou, Shijin Duan, Charles Fleming, Gaowen Liu, Ramana Rao Kompella, Shaolei Ren, and Xiaolin Xu. ConQuER: Modular Architectures for Control and Bias Mitigation in IQP Quantum Generative Models, October 2025.
- [21] Kevin Shen, Susanne Pielawa, Vedran Dunjko, and Hao Wang. Characterizing Trainability of Instantaneous Quantum Polynomial Circuit Born Machines, February 2026.
- [22] Sacha Lerch, Joseph Bowles, Ricard Puig, Erik Armengol, Zoë Holmes, and Supanut Thanasilp. IQP Born Machines under Data-dependent and Agnostic Initialization Strategies, March 2026.