Advanced Quantum Information Long Project - S2-25

Investigating barren plateaus in classically trained IQP generative models

Koshik Seeburrun · Mohamed Yaqub Alameen · Alejandro Jackson

Sorbonne Université · Master of Quantum Information | May 2026

IQP circuits and barren plateaus

IQP circuits

Hadamards → diagonal commuting layer D_θ → Hadamards → measure

- Commuting circuit $|\psi(\theta)\rangle = H^{\otimes n} D_\theta H^{\otimes n}|0\rangle$; D_θ is set by a binary generator matrix G — a hypergraph on the qubits.
- Sampling is classically hard, yet expectation values are closed-form — so the model trains classically.

Barren plateaus

gradient variance vs. qubit count n (log scale)

- $\mathrm{Var}[\partial_\theta L] \sim 2^{(-\alpha n)}$: gradients vanish exponentially as the qubit count n grows.
- Mean gradient ≈ 0 by symmetry, so trainability rides on the variance. A flat landscape blinds the optimizer, even when training is classical.

THE PROBLEM

IQP is the sampling-hard model class worth deploying on quantum hardware. With that being said, variational circuits can fall into barren plateaus, so trainability must be tested across circuit family, initialization, and MMD bandwidth.

Why Can We Train IQP Circuits Classically?

The Core Insight

IQP output probabilities decompose via a **Fourier-type expansion** over parity functions:

- ➔ Expectation values of **parity observables** have closed-form solutions
- ➔ No exponential state simulation needed
- ➔ Polynomial cost on classical hardware

Classical Training Loop

Parameters θ
trainable circuit angles

Parity Expectations
⟨f□⟩ computed analytically

MMD Loss ℒ(θ)
compare model ↔ data stats

Computing Gradient

Update parameters (θ)

↻ *iterate*

Training Datasets

Four binary-bitstring distributions spanning random → highly structured

MNIST

Handwritten digit images flattened to binary vectors

Pixel → 1 if > 0.5 else 0
High-dimensional, structured

Ising

Magnetic spin configurations from a physical system

{−1,+1} → binary
Correlated, physically structured

BLOBS

Synthetic multi-cluster data with noise

Multi-modal distribution
Noisy variations around cluster centers

Genomic

Real biological DNA variation data (805 qubits)

Binary-encoded data High-dim, complex real-world structure

All datasets encoded as binary bitstrings $x \in \{0,1\}^n$

MMD Loss & Training Convergence

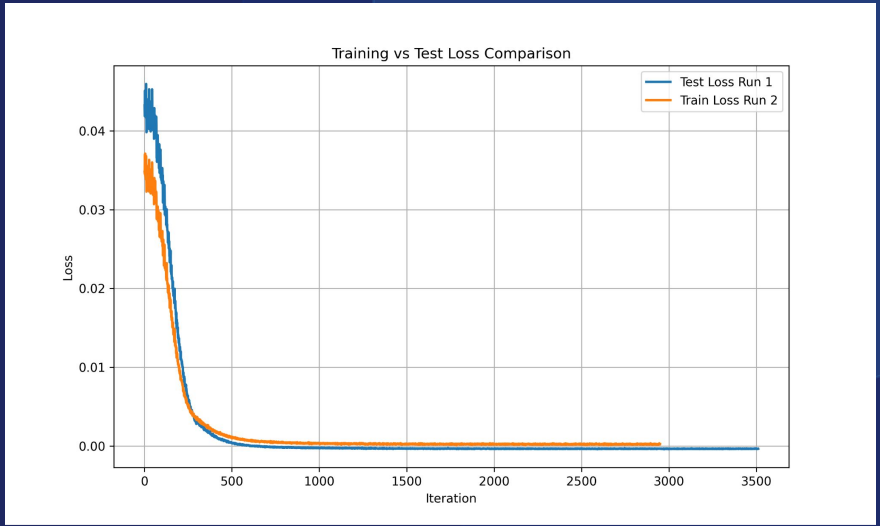
Maximum Mean Discrepancy

Compares distributions through sample statistics — no need for exact likelihoods.

Computed MMD for: Training data
Sigma = 0.6
MMD = 0.001048 ± 0.009328

Computed MMD for: Test data
Sigma = 0.6
MMD = - 0.007612 ± 0.005337

- ➔ Computed for an initially defined 10,000 iterations
- ➔ The loss saturates and stabilizes after certain number of trials, comparing close to zero
- ➔ Bandwidth σ determines which correlation scales the loss prioritizes



Loss drops rapidly, then saturates → model converges to match target distribution

Circuit Connectivity & Initialization

Structural choices that determine trainability (Hyper parameters)

CONNECTIVITY

Product-state

No entanglement — single-qubit only

ZZ Lattice (local gates)

Nearest-neighbor 2-qubit (ZZ) interactions

Sparse random graph

Random sparse ZZ couplings

Complete graph

All-to-all ZZ couplings, Highly entangling and computationally costly

INITIALIZATION

$\theta_i \sim \mathcal{N}(0, \sigma^2)$ where σ = init_scale

Standard: small Gaussian noise near identity

Uniform RISKY

Broad range ⟶ high entanglement ⟶ gradient collapse

Small-angle ($\sigma \approx 10^{-4}$) MODERATE

Near identity start ⟶ stable early, degrades at scale

Data-dependent (init_scale = −1) BETTER

Encodes data covariance ⟶ Often converges faster and more stably

Initialization Strategies

Definition

- Method used to choose the starting vector of parameters, θ of the quantum circuit before training begins.
- Steer the optimizer into a trainable region with high gradient variance

Experimental Setup

- **Dataset:** Ising model (structured data), unless specified.
- **Connectivity:** Lattice (grid-like structure).
- **Hardware Limit:** Up to 25 qubits.

Experimental Metrics

- **Gradient Variance:** Shows if optimizer has a clear slope or if it is stuck in a flat landscape.
- **MMD Loss:** Measures how convincingly the quantum circuit's generated data matches the target data.
- **Patch Radius:** Size of a localized neighborhood of data points. A small **r** captures local relationships, while a large **r** maps long complex patterns.

Note:

Other connectivities and datasets were tested but dropped due to hardware limitations.

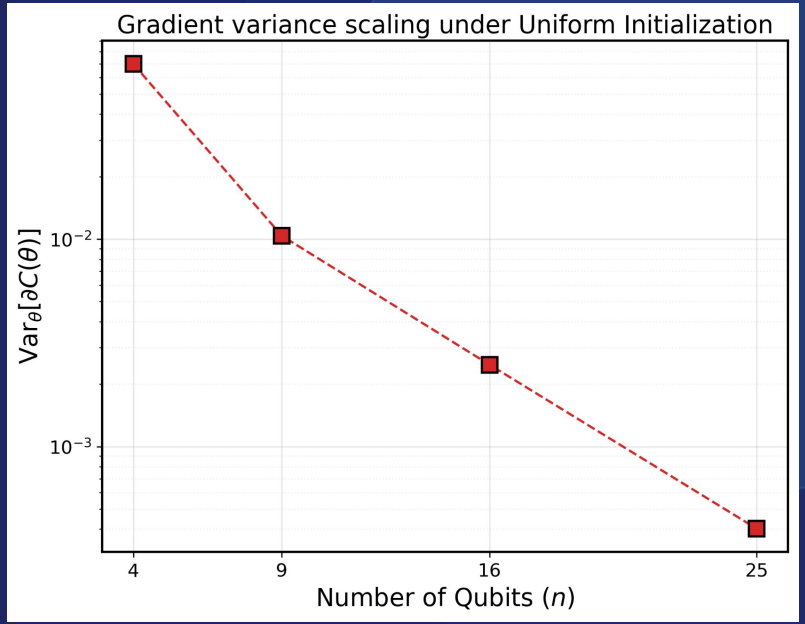
Uniform Initialisation

Definition

Parameters, θ are assigned completely random values across the full range of gate operators. (-π to π)

Takeaway

- **Prior theory:** Random circuits suffer exponential gradient collapse as size of the circuit increases, leading to BP.
- **Experiment:** Empirically reproduced the predicted behavior. Confirms the validity of our code and provides a reliable baseline to test other methods against.



As the number of qubits increases, the variance decays exponentially, this demonstrates barren plateau behavior in IQP-MMD training.

Small angle (Identity)

Definition

Parameters θ are initialized to values at or very close to zero. **($\sigma \approx 10^{-4}$)**

Mechanism

Gates initially act as identity operations, delaying early entanglement growth.

Theory notes

- Easy to implement
- Stable initially, weaker at scale

Data dependent Initialisation

Definition

Encodes classical statistical patterns from the target dataset directly into the initial parameters.

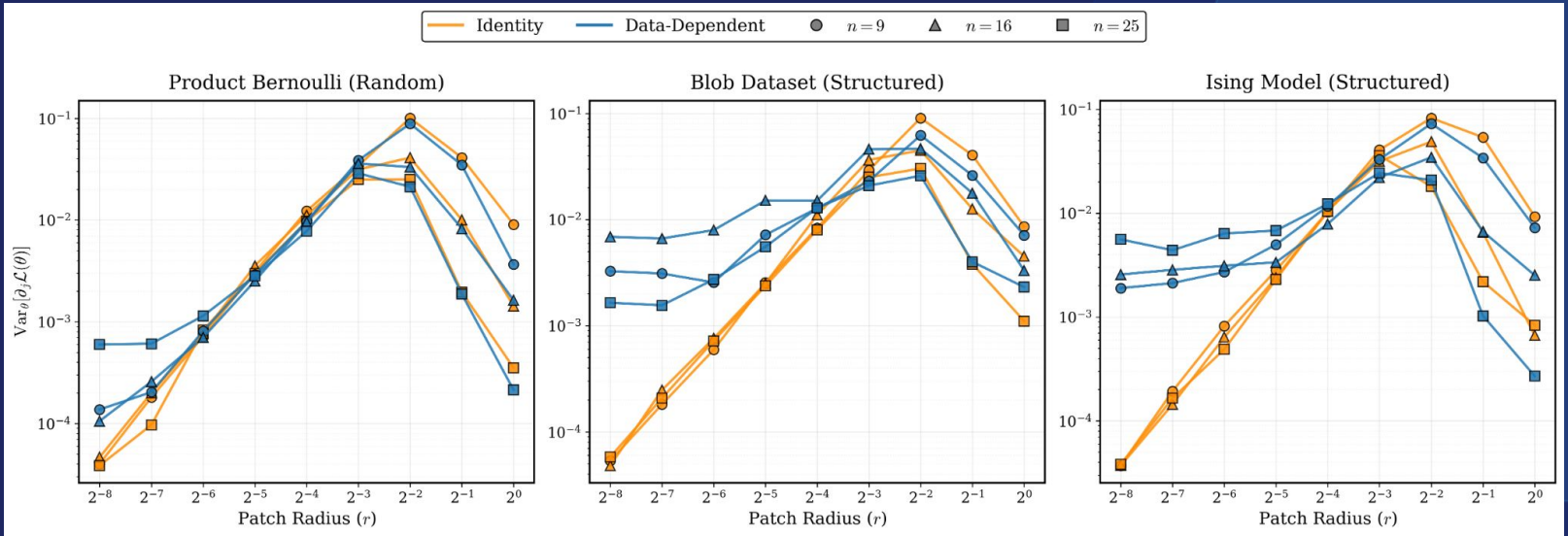
Mechanism

Gives the optimizer a "warm start" by placing the initial state near a favorable local minimum.

Theory notes

- Harder to implement, requires pre-analyzing the data
- Should delay Barren plateau.

Comparison Results



Compares small-angle (identity) initialization and data-dependent initialization across increasing sizes

Observation 1:

Gradient variance remains consistently higher than identity initialization across system sizes.

Observation 2:

For random data, both methods behave similarly because there are no meaningful correlations to encode.

Observation 3:

At 25 qubits, gradient variance decreases for both methods, indicating scaling remains challenging.

Layer-wise Initialisation

Definition

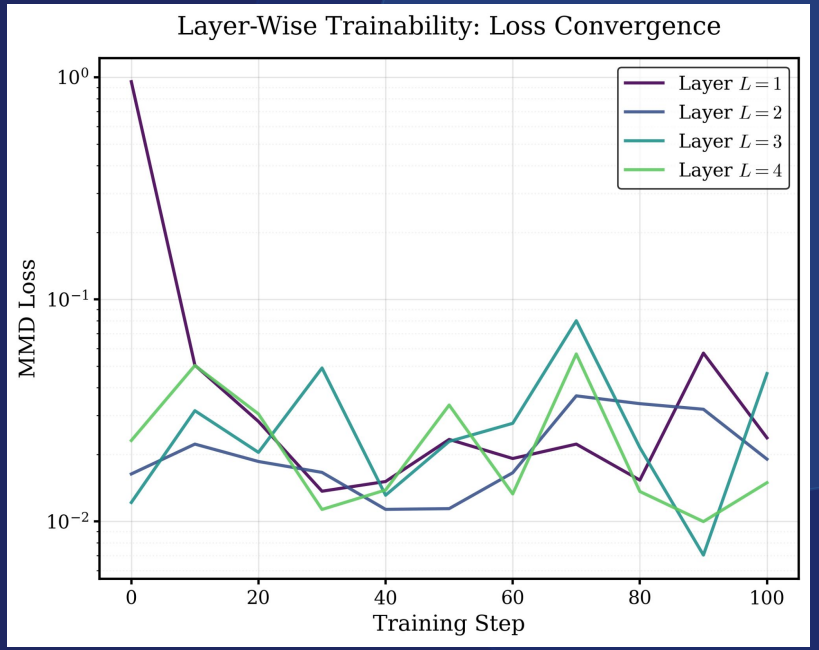
Start with data-dependent initialization, then progressively add new layers after each training round initializing their parameters to exactly zero.

Mechanism

New layers act as identity operators, they preserve the learned state of the previous layers while also increasing the circuit's capacity.

Takeaway

MMD loss remains stable and low across training steps, suggesting this is a promising scaling strategy.



Layer-wise trainability on the Ising dataset using a Lattice IQP architecture. (Size: 25 qubits)

Two diagnostics, not one number

A low MMD loss does not tell you what the model actually learned.

Marginal agreement

relational — compares two distributions

Does the learned q_θ match the target, p, on subsets of variables? Average TV per order $\longrightarrow T_k$.

Anti-concentration

a property of one distribution

How spread is q_θ over the 2^n outcomes?
Scaled second moment: $S(q) = 2^n \Sigma\, q(x)^2$.

Marginals test whether the model learned the target; anti-concentration tests the shape of what it learned.

METHOD

How we compute it: exact for small n, estimator for large n

Exact route · n ≤ 16

- Enumerate all $z \in \{0,1\}^n$
- Phase vector $d(z) = e^{(-i\varphi(z))}$, $\varphi(z) = \Sigma\square\theta\square(-1)^{(z\cdot g\square)}$
- Walsh–Hadamard transform, then square → $q_{\theta(x)}$
- Compute S, T_k, F_k exactly

Estimator route · n = 805

- Exact 2^{805} enumeration is impossible
- Sample random non-identity Pauli strings
- Parseval estimator: $\hat{S} = 1 + (2^n - 1) \cdot \langle Z_a \rangle^2$
- Sampled low-order marginals ($k \leq 8$)

CONCENTRATION SCALAR

$S(q) = 2^n \Sigma_x q(x)^2$
(uniform → 1; larger = more concentrated)

MARGINAL ERROR

T_k = mean over $|S|=k$ of $TV(p_S, q_S)$;

F_k = squared Z-word error

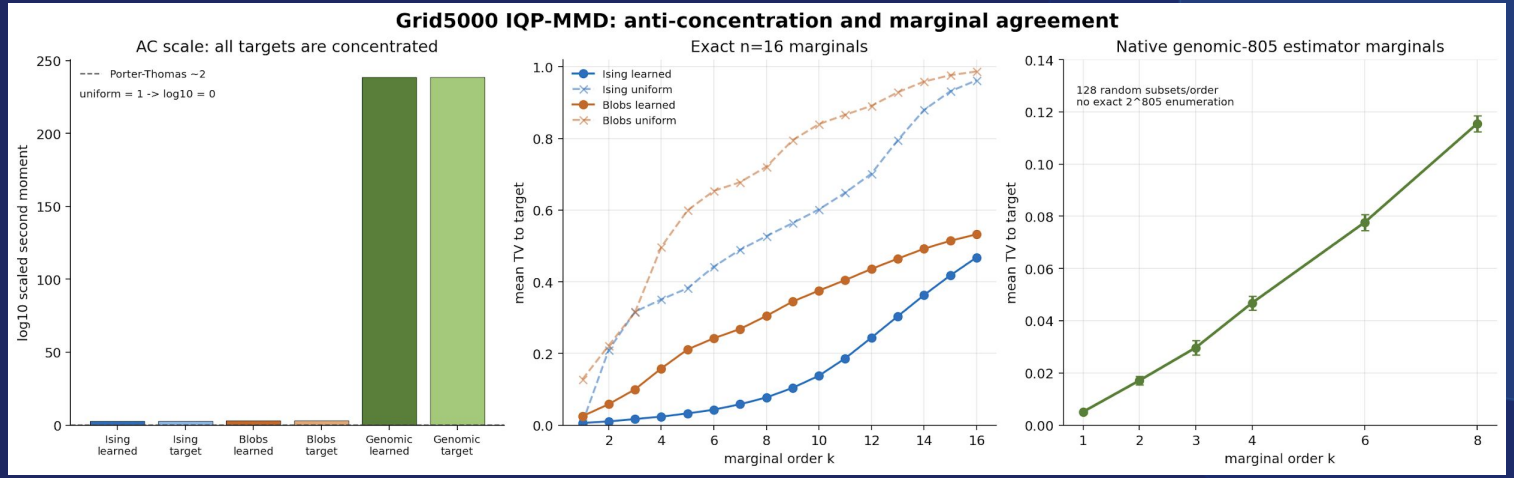
BANDWIDTH FILTER

$MMD^2_\sigma = C_\sigma \Sigma_S \tau^{|S|} (\langle Z_S \rangle_p - \langle Z_S \rangle_q)^2$,

$\tau = \tanh(1/4\sigma^2)$

RESULTS · LEARNED vs TARGET

Concentrated like their targets



Exact n=16 (Ising, blobs) + native genomic-805 estimator. Left: AC scale. Middle: marginal TV vs order k. Right: genomic low-order TV.

CONCENTRATION

Ising 826.9 ≈ target 806 · blobs 939 < 1658 · genomic ~10^{2388}. Not Porter–Thomas since the targets are concentrated too.

ISING MARGINALS

Cleanest case: ~95% lower TV than uniform at k=2; beats uniform at every order.

BLOBS / GENOMIC

Blobs: high-order mode structure only partially captured. Genomic: strong through k=8 (TV 0.005 → 0.12).

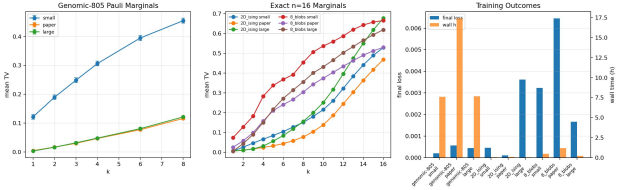
TAKEAWAY

Loss is not learning: the joint picture

The honest readout is the set $\{ S(q_\theta),\ S(p) - S(q_\theta),\ T_{k'},\ F_k \}$ is not a pass/fail flag.

- A low final MMD can hide marginal failure. For example, small-bandwidth genomic reached low loss but k=8 TV ≈ 0.45 vs 0.12 (paper / large).
- The AC “pass” (S ≥ 1) is vacuous — every distribution clears it. Compare to the target, not to uniform.
- The small-bandwidth failure is a red flag, not a diagnosis: consistent with both barren-plateau collapse and classical overfitting.

WHEN IT BREAKS



Bandwidth sweep — low loss can still mean bad marginals.

IQP-MMD trains classically in useful regimes, but success must be checked with gradient scaling, marginal agreement, and target-relative anti-concentration, not assumed from one loss value.

SIGNIFICANCE & CONCLUSION

Establishing benchmarks for large-scale GQML

Our investigation reveals that MMD-based training of IQP circuits scales classically to hundreds of qubits, but success is nuanced.

- **Mitigation strategies:** Data-dependent and layer-wise initialization preserved significantly larger gradient variance and improved scalability.
- **Scalability:** Confirmed training feasibility for n=805 using classical estimators without barren plateau collapse in target-relative regimes.
- **Validation:** Proved that low loss is insufficient; marginal agreement (T_k, F_k) is the primary metric for true learning.
- **Heuristics:** Concentration scalars (**S**) must match targets to avoid vacuous "success" flags.

Future work will focus on deployment on near-term quantum hardware and exploring non-commuting circuit architectures.

FINAL PROJECT IMPACT

Provides a robust framework for validating quantum generative models against classical shadows at scale.

REFERENCES

Key references

2023 **Trainability barriers and opportunities in quantum generative modeling**

M.~S. Rudolph et al.

arXiv:2305.02881, 2023.

2025 **scaling-gqml.**

XanaduAI.

GitHub repository, github.com/XanaduAI/scaling-gqml, accessed 2026-05-26.

2015 **Average-case complexity versus approximate simulation of commuting quantum computations**

M.~J. Bremner, A.~Montanaro, and D.~J. Shepherd.

arXiv:1504.07999, 2015.

2016 **Achieving quantum supremacy with sparse and noisy commuting quantum computations**

M.~J. Bremner, A.~Montanaro, and D.~J. Shepherd.

arXiv:1610.01808, 2016.

2025 **Train on classical, deploy on quantum: scaling generative quantum machine learning to a thousand qubits**

E.~Recio-Armengol, S.~Ahmed, and J.~Bowles.

arXiv:2503.02934, 2025.

2026 **iqp-mmd-barren-plateau**

Alejandro Jackson, Koshik Seeburrun, Mohamed Yaqub Alameen

GitHub repository, github.com/alejack312/iqp-mmd-barren-plateau, accessed 2026-05-26.

2024 **Barren plateaus in variational quantum computing**

M.~Larocca et al.

Nature Reviews Physics, 2024.

2025 **Limits of quantum generative models with classical sampling hardness**

S.~Herbst, I.~Brandic, and A.~Perez-Salinas.

arXiv:2512.24801, 2025.