

Generative Quantum Data Embeddings for Supervised Learning

Jaewoong Heo¹ and Daniel K. Park^{1,2,3}

¹Department of Statistics and Data Science, Yonsei University, Seoul 03722, Republic of Korea

²Department of Applied Statistics, Yonsei University, Seoul 03722, Republic of Korea

³Department of Quantum Information, Yonsei University, Seoul 03722, Republic of Korea

Abstract

The classification performance of quantum supervised learning on classical data is fundamentally controlled by the feature map that embeds input into quantum states. To promote quantum state distinguishability, we propose an energy-based generative learning framework that synthesizes gate sequences to optimize embedding structures and refine data-tailored parameters using a fidelity-based surrogate objective. Moreover, we explain the saturation of classification performance by extending quantum risk bounds into the classical input space, demonstrating that the achievable empirical risk across an embedding family is bounded by the 1-Wasserstein distance. Our research establishes a data-driven quantum embedding framework while showing that classical data geometry provides an *a priori* diagnostic for optimization gains.

1. Introduction

Supervised quantum machine learning (QML) on classical data begins with data embedding, which specifies how classical inputs are encoded into quantum states [1, 2]. Despite its central role, most QML studies adopt a fixed mapping scheme chosen *a priori* [3]. Recent methods like neural quantum embedding (NQE) [4] introduce data dependence by transforming inputs via classical neural networks before passing them to a fixed quantum circuit. Although NQE demonstrates the efficiency of data-driven preprocessing, the embedding remains confined to the transformations expressible by the pre-selected structure. This fixed structure imposes an inductive bias that may misalign with the geometry of a given dataset, fundamentally limiting achievable class separation [5, 6]. In this sense, maximizing separability requires a fully data-driven approach where the embedding circuit structure itself is treated as the optimization variable.

2. Separability Objective and Geometric Limit

Binary classification in QML corresponds to a quantum state discrimination problem [7, 8]. The embedding-induced geometry sets a fundamental limit on achievable class separation. For an embedding structure Φ , the quantum state representation of an input x is $\rho_\Phi(x) = |\psi_\Phi(x)\rangle\langle\psi_\Phi(x)|$. Given a labeled dataset $\{(x_i, y_i)\}_{i=1}^N$ with $y_i \in \{-1, +1\}$, the empirical class priors and class-conditional mixture states are $p^\pm = N_\pm/N$ and $\rho^\pm(\Phi) = \sum_{i:y_i=\pm 1} \rho_\Phi(x_i)/N_\pm$. The distinguishability between prior-weighted class states is determined by the trace distance, $D_{\text{tr}}(p^+ \rho^+(\Phi), p^- \rho^-(\Phi))$. Regardless of the downstream classifier, the empirical risk $\hat{R}(\Phi)$ is lower-bounded by this distinguishability as Eq. (1) [9],

$$\hat{R}(\Phi) \geq \frac{1}{2} - D_{\text{tr}}(p^- \rho^-(\Phi), p^+ \rho^+(\Phi)). \quad (1)$$

Since computing the trace distance is expensive on near-term quantum devices [10], a pairwise fidelity loss $\ell_{ij}(\Phi) = |\delta_{y_i y_j} - F_\Phi(x_i, x_j)|$ serves as a tractable surrogate objective, where $F_\Phi(x_i, x_j) = |\langle\psi_\Phi(x_i)|\psi_\Phi(x_j)\rangle|^2$ [1, 11]. Minimizing this loss achieves two objectives: it maximizes same-class fidelities to increase the purity and concentration of the class-conditional states, and it minimizes cross-class fidelities to increase the upper bound on the attainable trace distance via the convexity of the trace norm [12, 13].

To determine the ultimate limits of structural optimization, we extend the risk bounds of quantum state discrimination back into the classical input space. Denoting $\kappa_{\mathcal{F}}$ as the Lipschitz constant of embedding family $\mathcal{F}$ governed by the maximum sequence length and available gate pool, the relaxed uniform trace distance

continuity bound per sample is $D_{\text{tr}}(\rho_{\Phi}(x_i), \rho_{\Phi}(x_j)) \leq \kappa_{\mathcal{F}} \|x_i - x_j\|_1$. This per-sample relation extends to the ensemble level by considering a coupling π whose marginals are the empirical class-conditional input distributions $\hat{P}_+(x)$ and $\hat{P}_-(x)$. Applying the convexity of trace norm and Jensen’s inequality [14] yields $D_{\text{tr}}(\rho^+(\Phi), \rho^-(\Phi)) \leq \mathbb{E}_{(x_i, x_j) \sim \pi} [D_{\text{tr}}(\rho_{\Phi}(x_i), \rho_{\Phi}(x_j))] \leq \kappa_{\mathcal{F}} \mathbb{E}_{(x_i, x_j) \sim \pi} [\|x_i - x_j\|_1]$. Minimizing the right-hand side over all valid joint distributions corresponds to the 1-Wasserstein distance $W_1(\hat{P}_+(x), \hat{P}_-(x))$ [15]. Assuming balanced class priors, $p^+ = p^- = 1/2$, combining this ensemble trace distance bound with Eq. (1) yields the final risk lower bound for any embedding $\Phi \in \mathcal{F}$,

$$\hat{R}^*(X, y; \mathcal{F}) \geq \frac{1}{2} - \frac{1}{2} \kappa_{\mathcal{F}} W_1(\hat{P}_+(x), \hat{P}_-(x)), \quad (2)$$

where $\hat{R}^*(X, y; \mathcal{F})$ denotes the minimum empirical risk achievable over the embedding family when the Helstrom measurement is used. Equation (2) demonstrates that when empirical class distributions are close in Wasserstein distance, the achievable class separation is fundamentally limited, rendering substantial architecture-dependent gains unlikely.

3. Energy-based Generative Architecture Search and Refinement

To navigate the combinatorial search space of discrete circuits, we propose the energy-based generative architecture search (EGAS) framework. In EGAS, an embedding circuit is a sequence $s = (c_1, \dots, c_D)$ of depth-one subcircuits drawn from a tokenized pool $\mathcal{C}$. A generative pre-trained transformer (GPT) [16] parameterized by θ autoregressively samples candidate sequences under a learned distribution $p_{\theta}(s)$. Each generated circuit Φ_s is evaluated via a quantum simulator to compute its energy $E(s)$, defined as the batch-averaged pairwise fidelity loss over labeled pairs (Fig. 1).

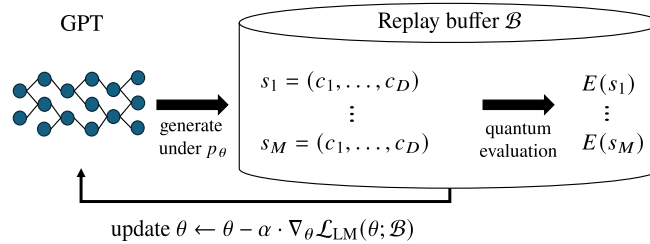


Figure 1: Overview of the Energy-based Generative Architecture Search (EGAS) training. In each iteration, a generative pre-trained transformer (GPT) parameterized by θ autoregressively samples M candidate sequences, which are translated into embedding circuits and evaluated using a quantum simulator to assign an energy value $E(s_m)$ based on the pairwise fidelity surrogate loss. The sequence-energy pairs are stored in a replay buffer $\mathcal{B}$ and used to update the generator parameter via a gradient step on the logit-matching loss $\mathcal{L}_{\text{LM}}$ with learning rate α .

At each training iteration, the GPT samples M candidate sequences. To align the cumulative logit score $w_{\text{sum}}(s_m; \theta)$ of the m -th sequence with the true landscape, the GPT is optimized via a gradient step on the logit-matching loss

$$\mathcal{L}_{\text{LM}}(\theta) = \frac{1}{M} \sum_{m=1}^M \left(e^{-\gamma w_{\text{sum}}(s_m; \theta)} - e^{-\gamma E(s_m)} \right)^2 \quad (3)$$

where $\gamma > 0$ is an inverse-temperature hyperparameter. By matching the logit score to the evaluated energy, the generator approximates a Boltzmann distribution where γ controls the sampling preference toward lower-energy circuits [17].

To exploit continuous degrees of freedom within optimized structures, we apply a post-search continuous parameter refinement step. For each parameterized gate, the data-dependent angle $\phi(x) = rx_i$ is replaced with $\tilde{\phi}(x) = rx_i + b_{\omega}(x)$, where $b_{\omega}(x)$ is a parameterized bias function providing a learnable additive offset. Initializing $b_{\omega}(x) = 0$ ensures the optimization does not prematurely leave the local optimum discovered during

discrete search. During refinement, the circuit structure is held fixed, and only the bias parameters ω are optimized via binary cross-entropy loss, ensuring local continuous adjustments without dismantling the optimized gate structure.

4. Results

We evaluate the searched embeddings on binary classification using a quantum kernel support vector machine (QKSVM) [1, 18], which utilizes the precomputed fidelity matrix $K_{ij} = F_{\Phi}(x_i, x_j)$ for the decision boundary. Across ten independent train-test splits, the selected embeddings from the low-energy group with continuous refinement, denoted as $G^*(\text{Bias})$, consistently outpaced the classical linear SVM, the data-agnostic ZZ feature map [1], and the NQE baseline across various datasets (Fig. 2) [19–26].

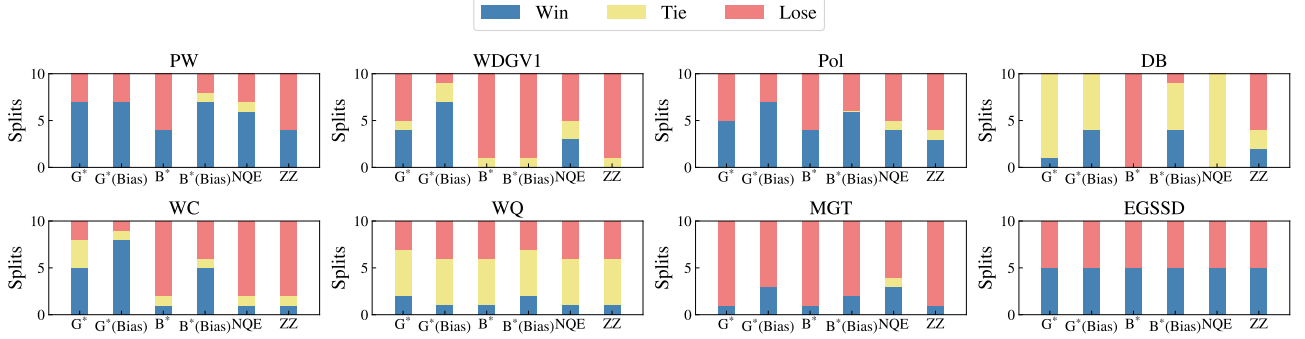


Figure 2: Split-wise comparison against the classical baseline across datasets. For each dataset, stacked bars reports the numbers of wins (blue), ties (yellow), and losses (red) over ten test split relative to the linear SVM. The symbol * denotes the strongest representative selected from each EGAS-derived group [G , $G(\text{Bias})$, B , and $B(\text{Bias})$], alongside the fixed-structure ZZ feature map and NQE baselines.

The classification performance trends validate our 1-Wasserstein diagnostic. To quantify sensitivity to the embedding choice, the interquartile range (IQR) of mean test accuracies across all evaluated models is analyzed. As Table 1 shows, datasets with small input-space class separation (WQ, MGT, EGSSD) display small IQRs and tightly clustered accuracies. This confirms performance is insensitive to embedding choice in weakly separated regimes (Fig. 3).

Table 1: Empirical 1-Wasserstein distances between the class-conditional input distributions.

Dataset	PW	WDGVI	Pol	DB	WC	WQ	MGT	EGSSD
W_1 dist.	5.24	5.16	4.35	13.91	10.86	3.01	3.30	3.56

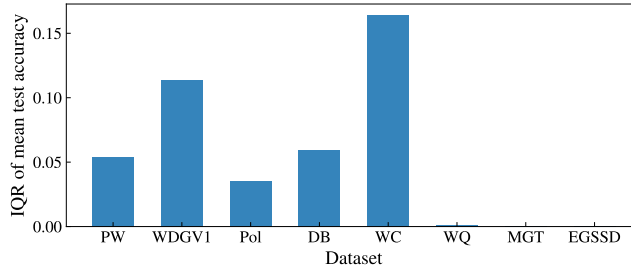


Figure 3: Dataset-level dispersion of QKSVM test accuracy across the evaluated quantum embeddings. For each dataset, the bar height represents the interquartile range (IQR) of mean test accuracies over the fixed quantum baselines (ZZ and NQE) and the EGAS-derived candidate embeddings from the G , $G(\text{Bias})$, B , and $B(\text{Bias})$ groups.

References

- [1] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747): 209–212, 2019.
- [2] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020.
- [3] Maria Schuld and Francesco Petruccione. *Machine learning with quantum computers*, volume 676. Springer, 2021.
- [4] Tak Hur, Israel F Araujo, and Daniel K Park. Neural quantum embedding: Pushing the limits of quantum supervised learning. *Physical Review A*, 110(2):022411, 2024.
- [5] Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf. The inductive bias of quantum kernels. *Advances in Neural Information Processing Systems*, 34:12661–12673, 2021.
- [6] 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, 2021.
- [7] Joonwoo Bae and Leong-Chuan Kwek. Quantum state discrimination and its applications. *Journal of Physics A: Mathematical and Theoretical*, 48(8):083001, 2015.
- [8] Roberto Giuntini, Hector Freytes, Daniel K Park, Carsten Blank, Federico Holik, Keng Loon Chow, and Giuseppe Sergioli. Quantum state discrimination for supervised classification. *arXiv preprint arXiv:2104.00971*, 2021.
- [9] Carl W Helstrom. Quantum detection and estimation theory. *Journal of Statistical Physics*, 1(2):231–252, 1969.
- [10] Qisheng Wang and Zhicheng Zhang. Fast quantum algorithms for trace distance estimation. *IEEE Transactions on Information Theory*, 70(4):2720–2733, 2023.
- [11] Harry Buhrman, Richard Cleve, John Watrous, and Ronald De Wolf. Quantum fingerprinting. *Physical review letters*, 87(16):167902, 2001.
- [12] Mark M Wilde. *Quantum information theory*. Cambridge university press, 2013.
- [13] Christopher A Fuchs and Jeroen Van De Graaf. Cryptographic distinguishability measures for quantum-mechanical states. *IEEE Transactions on Information Theory*, 45(4):1216–1227, 2002.
- [14] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.
- [15] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.
- [16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [17] Kouhei Nakaji, Lasse Bjørn Kristensen, Jorge A Campos-Gonzalez-Angulo, Mohammad Ghazi Vakili, Haozhe Huang, Mohsen Bagherimehrab, Christoph Gorgulla, FuTe Wong, Alex McCaskey, Jin-Sung Kim, et al. The generative quantum eigensolver (gqe) and its application for ground state search. *arXiv preprint arXiv:2401.09253*, 2024.

- [18] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nature communications*, 12(1):2631, 2021.
- [19] Rami Mohammad and Lee McCluskey. Phishing Websites. UCI Machine Learning Repository, 2012. DOI: <https://doi.org/10.24432/C51W2X>.
- [20] L. Breiman and C.J. Stone. Waveform Database Generator (Version 1). UCI Machine Learning Repository, 1984. DOI: <https://doi.org/10.24432/C5CS3C>.
- [21] Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.
- [22] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.
- [23] Dry Bean. UCI Machine Learning Repository, 2020. DOI: <https://doi.org/10.24432/C50S4B>.
- [24] Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: <https://doi.org/10.24432/C56S3T>.
- [25] R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. DOI: <https://doi.org/10.24432/C52C8B>.
- [26] Vadim Arzamasov. Electrical Grid Stability Simulated Data . UCI Machine Learning Repository, 2018. DOI: <https://doi.org/10.24432/C5PG66>.