Statistics

A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: www.tandfonline.com/journals/gsta20

Quantum geometric-entropic optimization for customer lifetime value prediction: convergence theory and an empirical study on transactional retail data

Massimiliano Ferrara , Laura Sáez-Ortuño , Santiago Forgas-Coll , Jorge Refugio Fabila-Fabián , Carlos Martín-Isla & Karim Lekadir

To cite this article: Massimiliano Ferrara , Laura Sáez-Ortuño , Santiago Forgas-Coll , Jorge Refugio Fabila-Fabián , Carlos Martín-Isla & Karim Lekadir (11 May 2026): Quantum geometric-entropic optimization for customer lifetime value prediction: convergence theory and an empirical study on transactional retail data, Statistics, DOI: 10.1080/02331888.2026.2667471

To link to this article: https://doi.org/10.1080/02331888.2026.2667471

Published online: 11 May 2026.

STATISTICS
https://doi.org/10.1080/02331888.2026.2667471

Quantum geometric-entropic optimization for customer lifetime value prediction: convergence theory and an empirical study on transactional retail data

Massimiliano Ferrara[a,b,c], Laura Sáez-Ortuño[d], Santiago Forgas-Coll[d], Jorge Refugio Fabila-Fabián[e], Carlos Martín-Isla[e] and Karim Lekadir[e,f]

[a]Department of Law, Economics and Human Sciences & Decisions Lab, University Mediterranea of Reggio Calabria, Reggio Calabria, Italy; [b]ICRIOS, Bocconi University, Milan, Italy; [c]Faculty of Engineering and Natural Sciences, Istanbul Okan University, Istanbul, Turkey; [d]Facultat d'Economia i Empresa, Universitat de Barcelona, Barcelona, Spain; [e]Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain; [f]ICREA, Barcelona, Spain

ABSTRACT

Predicting customer churn from transactional data is a central problem in management science, with direct implications for retention strategy, revenue forecasting, and resource allocation. This paper introduces Quantum Geometric-Entropic Optimization (Q-GEO), a framework that integrates Geometric-Entropic Optimization – combining Riemannian gradient methods with entropy-regularized optimal transport – into the training of variational quantum kernels for classification. The algorithm operates on a parameter manifold equipped with a Fisher-Wasserstein metric and incorporates Sinkhorn-type projections to enforce distributional coherence on the quantum feature space. We establish three theoretical contributions: (i) a convergence theorem for Q-GEO-trained variational quantum kernels under a combined Polyak–Łojasiewicz and Sinkhorn contraction framework, yielding linear convergence in the Riemannian condition number plus geometric contraction of the Sinkhorn residual; (ii) a margin amplification result showing that GEO-trained quantum embeddings achieve improved separation bounds over Euclidean-trained counterparts due to the spectral regularization provided by the Wasserstein component of the Fisher-Wasserstein metric; and (iii) a distributional stability result proving that Sinkhorn-projected quantum kernel matrices preserve a doubly stochastic spectral structure that mitigates kernel collapse in imbalanced settings. We validate the framework on the UCI Online Retail II dataset ($N = 5{,}942$ customers, $d = 11$ RFM-extended features, churn rate $\approx$37%), a publicly available transactional benchmark. Under nested cross-validation, Q-GEO achieves 0.8614 accuracy, 0.8103 precision, 0.7891 recall, 0.7996 F1, and 0.9047 ROC AUC, outperforming both classical baselines (including logistic regression, random forest, XGBoost, and CatBoost) and standard variational quantum kernel methods. We complement the accuracy analysis with SHAP-based explainability, computation time comparisons, and a detailed

ARTICLE HISTORY
Received 8 March 2026
Accepted 25 April 2026

KEYWORDS
Quantum machine learning; geometric-entropic optimization; Riemannian optimization; optimal transport; customer churn prediction; management science; variational quantum kernels; NISQ algorithms; explainability

2020 MATHEMATICS SUBJECT CLASSIFICATIONS
68T07; 81P45; 90C25

CONTACT Laura Sáez-Ortuño laurasaez@ub.edu Facultat d'Economia i Empresa, Universitat de Barcelona Avda. Diagonal, 690, Barcelona 08034, Spain

feature engineering appendix to support interpretability and reproducibility. We interpret these results as evidence that geometric optimization principles can materially enhance quantum machine learning for management science applications.

1. Introduction

Customer relationship management (CRM) constitutes one of the most consequential domains in management science, where predictive analytics directly informs resource allocation, retention budgeting, and long-term revenue forecasting [1,2]. Within this domain, customer churn prediction – identifying customers likely to discontinue purchasing – represents a canonical classification problem whose solution feeds into a cascade of downstream managerial decisions: targeted marketing interventions, loyalty program design, dynamic pricing adjustments, and workforce scheduling for service teams [3]. The economic magnitude of churn-related losses is substantial: industry estimates suggest that acquiring a new customer costs five to seven times more than retaining an existing one, making accurate churn identification a high-leverage operational capability [4].

From a management science perspective, churn prediction on transactional data presents distinctive analytical challenges that distinguish it from survey-based or subscription-based settings. Transactional datasets record purchasing events rather than explicit cancellation signals, requiring the analyst to define churn through recency thresholds – an inherently ambiguous boundary that interacts with seasonal purchasing patterns, product lifecycle dynamics, and heterogeneous customer segments [5]. The feature space combines monetary, temporal, and behavioural dimensions (recency, frequency, monetary value, product diversity, temporal regularity) that exhibit nonlinear interactions and heavy-tailed distributions. Class imbalance, while less extreme than in some operational settings, remains a material concern, as churned customers typically constitute 30–40% of the base [1].

Classical machine learning approaches – logistic regression, random forests, gradient-boosted trees, and support vector machines – have been extensively applied to churn prediction with considerable success [2,6,7]. These methods, however, face structural limitations in capturing the complex nonlinear manifold structure of transactional feature spaces. Kernel methods, which map data into higher-dimensional spaces where linear separation becomes feasible, offer a principled alternative, but classical kernels (RBF, polynomial) are constrained to feature spaces whose dimensionality grows polynomially with the input dimension.

Quantum kernel methods exploit the exponentially large Hilbert space of quantum circuits to define inner products in spaces of dimension 2^n at cost polynomial in the qubit count n [8–10]. This enables richer feature representations while preserving the convex optimization structure of support vector machines. Building on recent advances demonstrating the feasibility of quantum kernels for marketing analytics [11] and workforce management [12], this paper addresses a fundamental limitation of existing quantum kernel approaches: the optimization of variational circuit parameters.

Standard variational quantum kernel training treats the parameter space as Euclidean, applying gradient descent or Adam-type updates to optimize kernel alignment or classification loss. This approach ignores the intrinsic Riemannian geometry of quantum circuit parameters, where the Fisher information metric captures the statistical sensitivity of quantum state outputs to parameter perturbations [13]. The Geometric-Entropic Optimization (GEO) framework [14] was recently introduced to address precisely this gap: by combining Riemannian gradient descent (respecting the Fisher metric) with entropy-regularized optimal transport constraints (enforcing distributional coherence via Sinkhorn projections), GEO achieves provably faster convergence and more stable training dynamics for neural networks. The present paper extends GEO to the quantum computing setting and applies it to a management science classification problem with real-world data.

1.1. Main contributions

This paper makes three theoretical and one empirical contribution:

(1) *Q-GEO convergence theory* (Theorem 4.7): We prove that variational quantum kernel optimization using Riemannian gradients with respect to the quantum Fisher-Wasserstein metric, combined with Sinkhorn projections on the kernel Gram matrix, converges at rate $\mathcal{O}(\kappa_G \log(1/\varepsilon)) + \mathcal{O}(\rho^{2K})$, where κ_G is the condition number of the combined metric and ρ is the Sinkhorn contraction rate. The convergence guarantee unifies the Polyak–Łojasiewicz framework from quantum kernel theory with the geometric contraction analysis from optimal transport.

(2) *GEO-enhanced separation bounds* (Theorem 5.1): We demonstrate that quantum feature maps trained via GEO achieve margin amplification of $\Omega(\sqrt{2^L/d_{\text{eff}}}) \cdot (1 + \lambda\sigma_W)$ over classically trained quantum embeddings, where σ_W captures the spectral regularization provided by the Wasserstein component of the Fisher-Wasserstein metric. The improvement arises because Riemannian optimization navigates the quantum parameter landscape more efficiently, avoiding the flat directions and saddle points that plague Euclidean training.

(3) *Sinkhorn-stabilized quantum kernels* (Proposition 5.2): We prove that applying Sinkhorn projections to the quantum kernel Gram matrix enforces a doubly stochastic spectral structure that provably prevents kernel collapse – a pathology where all pairwise kernel values converge to a constant, eliminating discriminative capacity. This result has implications for any quantum kernel method operating on imbalanced data.

(4) *Empirical validation on transactional retail data* (Section 6): We validate Q-GEO on the UCI Online Retail II dataset [15], a publicly available benchmark containing over 1,000,000 transactions from a UK-based online retailer. After RFM-extended feature engineering, the resulting customer-level dataset ($N = 5{,}942$, $d = 11$ features, churn rate $\approx$37%) provides a realistic management science testbed. Q-GEO achieves the highest AUC (0.9047) and F1 (0.7996) among all methods tested.

1.2. Relation to management science

The contribution of this paper to management science operates along three dimensions. The first dimension is methodological: we introduce a quantum optimization framework whose convergence guarantees and separation bounds are stated in terms that connect directly to managerial decision variables – misclassification costs, retention intervention budgets, and the cost-benefit ratio of predictive accuracy improvements. The second dimension is empirical: by validating on transactional data rather than synthetic or subscription-based datasets, we demonstrate applicability to the non-contractual setting that characterizes most retail, e-commerce, and wholesale operations. The third dimension is architectural: the Q-GEO framework is designed to be modular, with the geometric optimization layer operating independently of the downstream decision model, enabling integration with stochastic programming, multi-criteria optimization, or game-theoretic allocation mechanisms [14].

1.3. Related work

1.3.1. Quantum kernel methods and management science

The quantum kernel framework was formalized by Schuld and Killoran [10] and experimentally demonstrated by Havlíček et al. [8]. Rigorous quantum advantages for classification were established by Liu et al. [9]. Recent applications to marketing analytics [11] and workforce management [12] have demonstrated feasibility in management science settings with formal convergence and separation guarantees. Within operations research, quantum computing has been explored for portfolio optimization [16], vehicle routing [17], and supply chain management [18], but applications combining quantum kernels with advanced optimization frameworks for customer analytics remain unexplored.

1.3.2. Geometric and Riemannian optimization

The Riemannian optimization paradigm, rooted in the geometric dynamics tradition [19], has gained considerable traction in machine learning through the natural gradient [13], optimization on matrix manifolds [20], and recent applications to deep learning including Muon [21] and manifold-constrained hyper-connections [22]. The GEO algorithm [14] unified these approaches through the Fisher-Wasserstein metric and Sinkhorn-based constraint enforcement. The present paper represents the first extension of geometric-entropic optimization principles to quantum computing.

1.3.3. Customer churn prediction

Classical churn prediction methods span logistic regression, ensemble methods, deep learning, and hybrid approaches [1,2,7]. The RFM (Recency, Frequency, Monetary) framework provides a standard feature engineering pipeline for transactional data [5]. To our knowledge, no prior work has combined quantum kernel methods with Riemannian optimization for customer churn prediction.

1.4. Paper organization

Section 2 reviews quantum kernels, the GEO algorithm, and the cost-sensitive Q-SVM formulation. Section 3 develops the Q-GEO framework, including the quantum Fisher-Wasserstein metric and Sinkhorn-stabilized kernels. Section 4 presents the convergence theory. Section 5 derives separation bounds. Section 6 presents the empirical validation and cost-benefit analysis. Section 8 concludes.

2. Preliminaries

2.1. Quantum feature maps and kernels

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space. A quantum feature map is $\varphi_\theta : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H} = (\mathbb{C}^2)^{\otimes n}$ is the n-qubit Hilbert space, realized by a parameterized unitary circuit:

$$\varphi_\theta(x) = U(x,\theta)\,|0\rangle^{\otimes n}, \tag{1}$$

with $U(x,\theta)$ encoding data x and variational parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. The quantum kernel is the squared overlap:

$$k_\theta(x_i,x_j) = |\langle \varphi_\theta(x_i)|\varphi_\theta(x_j)\rangle|^2. \tag{2}$$

The Gram matrix $\mathbf{K}_\theta \in \mathbb{R}^{N\times N}$ with entries $[\mathbf{K}_\theta]_{ij} = k_\theta(x_i,x_j)$ is positive semi-definite by construction and serves as the kernel matrix for downstream SVM optimization.

2.2. The GEO framework

The Geometric-Entropic Optimization algorithm [14] operates on parameter manifolds equipped with a combined Fisher-Wasserstein metric [23].

Definition 2.1 (Fisher-Wasserstein metric): For a parameterized model $p(y|x,\theta)$, the Fisher-Wasserstein metric tensor is:

$$G_{ij}(\theta) = g^F_{ij}(\theta) + \lambda\, g^W_{ij}(\theta), \tag{3}$$

where g^F is the Fisher information metric, g^W is the Wasserstein metric tensor capturing distributional geometry, and $\lambda > 0$ balances the two components.

GEO iterates consist of four operations: (1) Riemannian gradient computation using the metric inverse G^{-1}; (2) Sinkhorn projection of constrained matrices onto the doubly stochastic manifold; (3) orthogonal retraction via Newton-Schulz iterations (used in the original GEO formulation for weight matrix orthogonalization; in the quantum setting of this paper, this step is replaced by the unitarity constraint inherent in quantum circuit parameterization, see Algorithm 1); and (4) multi-scale entropic regularization. Letting T denote the number of optimization iterations and K the number of Sinkhorn projection steps per iteration, the convergence rate is $\mathcal{O}(1/\sqrt{T}) + \mathcal{O}(\rho^{2K})$, where the first term reflects Riemannian gradient descent and the second captures Sinkhorn contraction [14,24].

2.3. Cost-sensitive Q-SVM formulation

For a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with $y_i \in \{-1, +1\}$, the cost-sensitive Q-SVM solves:

$$\min_{\alpha} \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k_\theta(x_i, x_j) - \sum_{i=1}^{N} \alpha_i, \quad \text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C_{y_i}, \tag{4}$$

where $C_{+1} = C \cdot w_+$ and $C_{-1} = C \cdot w_-$ reflect asymmetric misclassification costs. In churn prediction, $w_+ > w_-$ penalizes missed churners more heavily, reflecting the high cost of customer loss relative to the marginal cost of unnecessary retention interventions.

3. Quantum geometric-entropic optimization

3.1. The quantum Fisher-Wasserstein metric

We extend the Fisher-Wasserstein metric to the quantum circuit parameter space. For a variational quantum circuit $U(x, \theta)$, the output state $\rho_\theta(x) = |\varphi_\theta(x)\rangle\langle\varphi_\theta(x)|$ defines a family of quantum states parameterized by θ.

Definition 3.1 (Quantum Fisher Information Matrix): The Quantum Fisher Information Matrix (QFIM) at θ is:

$$[F_Q(\theta)]_{ij} = 4 \,\mathrm{Re}\left[\langle\partial_i\varphi_\theta|\partial_j\varphi_\theta\rangle - \langle\partial_i\varphi_\theta|\varphi_\theta\rangle\langle\varphi_\theta|\partial_j\varphi_\theta\rangle\right], \tag{5}$$

where $|\partial_i\varphi_\theta\rangle = \partial|\varphi_\theta\rangle/\partial\theta_i$ and we suppress the data dependence for notational clarity.

The QFIM captures the sensitivity of the quantum state to parameter perturbations and defines a natural Riemannian structure on the parameter manifold Θ. However, the QFIM can become ill-conditioned for deep circuits or redundant parameterizations – a phenomenon related to barren plateaus in variational quantum algorithms [25].

Definition 3.2 (Quantum Fisher-Wasserstein metric): The quantum Fisher-Wasserstein metric at θ is:

$$G_Q(\theta) = F_Q(\theta) + \lambda\, W_Q(\theta), \tag{6}$$

where $W_Q(\theta)$ is the Wasserstein-induced metric tensor defined by:

$$[W_Q(\theta)]_{ij} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} W_2^2\left(\rho_\theta(x),\ \rho_{\theta+\delta_i e_j}(x)\right) \Big/ \delta_i\delta_j, \tag{7}$$

evaluated over a minibatch $\mathcal{B} \subseteq \mathcal{D}$, with e_j the j-th standard basis vector and W_2 the 2-Wasserstein distance between the output distributions induced by the quantum states.

Remark 3.1 (Regularization of the QFIM): The Wasserstein component λW_Q serves as a spectral regularizer for the QFIM. When F_Q has small eigenvalues (corresponding to flat directions in the quantum loss landscape), W_Q provides a positive lower bound on the combined metric's spectrum. This mitigates the vanishing gradient problem without the ad hoc damping strategies commonly used in natural gradient methods for quantum circuits.

3.2. Sinkhorn-stabilized quantum kernels

The quantum kernel Gram matrix $\mathbf{K}_\theta$ can suffer from *kernel concentration* in the NISQ regime: noise and limited circuit expressivity cause all pairwise kernel values to converge toward a constant, eliminating class-discriminative information [26]. We address this through Sinkhorn normalization of the kernel matrix.

Definition 3.3 (Sinkhorn-normalized quantum kernel): Given the quantum Gram matrix $\mathbf{K}_\theta$ with entries $k_\theta(x_i, x_j) \in (0, 1]$, the Sinkhorn-normalized kernel is:

$$\tilde{\mathbf{K}}_\theta = D_1 \, \mathbf{K}_\theta \, D_2, \tag{8}$$

where $D_1 = \mathrm{diag}(u)$ and $D_2 = \mathrm{diag}(v)$ are diagonal scaling matrices computed by K Sinkhorn iterations:

$$u^{(k+1)} = \mathbf{1} \oslash (\mathbf{K}_\theta \, v^{(k)}), \quad v^{(k+1)} = \mathbf{1} \oslash (\mathbf{K}_\theta^\top u^{(k+1)}), \tag{9}$$

with $\oslash$ denoting elementwise division and uniform target marginals.

The Sinkhorn normalization projects $\mathbf{K}_\theta$ onto a neighborhood of the doubly stochastic manifold $\mathcal{DS} = \{P \in \mathbb{R}_+^{N \times N} : P\mathbf{1} = \mathbf{1}, \, P^\top \mathbf{1} = \mathbf{1}\}$, ensuring that the kernel matrix maintains a balanced spectral profile across classes.

3.3. The Q-GEO algorithm

Algorithm 1 presents the complete Q-GEO procedure for variational quantum kernel training.

The weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ encodes asymmetric misclassification costs: $W_{ij} = w_{y_i}$ for same-class pairs and $W_{ij} = \sqrt{w_{y_i} w_{y_j}}$ for cross-class pairs. The multi-scale entropy regularization operates at batch level (H_{batch}, encouraging diversity in kernel evaluations across the minibatch) and parameter level (H_{param}, preventing parameter collapse).

Remark 3.2 (Computational cost): The per-iteration cost of Q-GEO is dominated by three components: quantum circuit evaluations for the Gram matrix ($\mathcal{O}(B^2 \cdot p \cdot nL)$ using the parameter-shift rule, where L denotes the number of circuit layers), Sinkhorn iterations ($\mathcal{O}(K \cdot B^2)$ [27,28]), and KFAC-approximated metric inversion ($\mathcal{O}(p)$ [29]). Under the assumption $K \ll p \cdot n \cdot L$ (which holds in our experimental setting where K=5, p=30, n=5, L=3), the Sinkhorn and KFAC costs are negligible relative to circuit evaluations, so the total overhead compared to standard variational quantum kernel training is approximately 2× (from the QFIM estimation via additional parameter-shift evaluations).

4. Convergence theory

4.1. Setup and assumptions

Consider the cost-weighted kernel alignment objective:

$$L_w(\theta) = \mathbb{E}_{\mathcal{B} \sim \mathcal{D}} \left[- \frac{\langle \tilde{\mathbf{K}}_\theta \circ \mathbf{W}, \, \mathbf{y}\mathbf{y}^\top \rangle_F}{\| \tilde{\mathbf{K}}_\theta \circ \mathbf{W} \|_F \, \| \mathbf{y}\mathbf{y}^\top \|_F} \right], \tag{10}$$

where $\tilde{\mathbf{K}}_\theta$ is the Sinkhorn-normalized quantum kernel.

Algorithm 1 Quantum Geometric-Entropic Optimization (Q-GEO)

Require: Initial parameters θ_0, learning rate η, Sinkhorn iterations K, entropy scales $(\varepsilon_1, \varepsilon_2)$, Wasserstein coupling λ, cost weights (w_+, w_-)

for $t = 0, 1, 2, \ldots, T-1$ **do**

  Sample minibatch $\mathcal{B}_t \subset \mathcal{D}$ of size B

  Compute quantum Gram matrix $\mathbf{K}_{\theta_t}$ on $\mathcal{B}_t$ via circuit evaluation

  Sinkhorn projection:

  for $k = 1, \ldots, K$ **do**

    $u^{(k)} \leftarrow \mathbf{1} \oslash (\mathbf{K}_{\theta_t} v^{(k-1)})$; $v^{(k)} \leftarrow \mathbf{1} \oslash (\mathbf{K}_{\theta_t}^\top u^{(k)})$

  end for

  $\tilde{\mathbf{K}}_{\theta_t} \leftarrow \mathrm{diag}(u^{(K)})\, \mathbf{K}_{\theta_t}\, \mathrm{diag}(v^{(K)})$

  Compute cost-weighted kernel alignment loss averaged over the minibatch:

  $L_w(\theta_t) = \frac{1}{|\mathcal{B}_t|} \sum_{\mathcal{B}_t} \left(-\frac{\langle \tilde{\mathbf{K}}_{\theta_t} \circ \mathbf{W}, \mathbf{y}\mathbf{y}^\top \rangle_F}{\|\tilde{\mathbf{K}}_{\theta_t} \circ \mathbf{W}\|_F \, \|\mathbf{y}\mathbf{y}^\top\|_F} \right)$

  Riemannian gradient:

  Estimate $\hat{G}_Q(\theta_t)$ via parameter-shift rule with Kronecker-Factored Approximate Curvature (KFAC) [29] block-diagonal approximation of the quantum Fisher-Wasserstein metric

  $\tilde{g}_t \leftarrow \hat{G}_Q(\theta_t)^{-1} \nabla_\theta L_w(\theta_t)$

  Entropic regularization:

  $\tilde{g}_t \leftarrow \tilde{g}_t + \varepsilon_1 \nabla_\theta H_{\text{batch}}(\theta_t) + \varepsilon_2 \nabla_\theta H_{\text{param}}(\theta_t)$

  where $H_{\text{batch}}(\theta) = -\sum_{(i,j)\in\mathcal{B}} \hat{k}_{ij} \log \hat{k}_{ij}$ is the batch-level kernel entropy (with $\hat{k}_{ij} = k_\theta(x_i, x_j) / \sum_{i',j'} k_\theta(x_{i'}, x_{j'})$) encouraging diversity in kernel evaluations, and $H_{\text{param}}(\theta) = -\sum_k \hat{\theta}_k \log \hat{\theta}_k$ (with $\hat{\theta}_k = \theta_k^2 / \|\theta\|^2$) is the parameter-level entropy preventing parameter collapse

  Update: $\theta_{t+1} \leftarrow \theta_t - \eta\, \tilde{g}_t$

end for

Solve cost-sensitive Q-SVM (4) using $\tilde{\mathbf{K}}_{\theta_T}$

Assumption 4.1 (β_G-Riemannian smoothness): The loss $L_w(\theta)$ is β_G-smooth with respect to the quantum Fisher-Wasserstein metric:

$$\|\nabla_G L_w(\theta) - \nabla_G L_w(\theta')\|_{G^{-1}} \leq \beta_G\, d_G(\theta, \theta')$$

for all $\theta, \theta' \in \Theta$, where d_G is the geodesic distance and $\nabla_G = G_Q^{-1} \nabla$ is the Riemannian gradient.

Assumption 4.2 (Quantum Fisher-Wasserstein metric bounds): The metric satisfies $\mu_G I \preceq G_Q(\theta) \preceq M_G I$ for all $\theta \in \Theta$, where $\mu_G = \lambda_{\min}(G_Q(\theta))$ and $M_G = \lambda_{\max}(G_Q(\theta))$ denote the minimum and maximum eigenvalues of the quantum Fisher-Wasserstein metric tensor, respectively, with condition number $\kappa_G = M_G / \mu_G$.

Assumption 4.3 (Riemannian PL condition): The loss satisfies a Riemannian Polyak–Łojasiewicz condition:

$$\|\nabla_G L_w(\theta)\|_{G^{-1}}^2 \geq 2\mu_G^{\text{PL}} (L_w(\theta) - L_w^*)$$

for some $\mu_G^{\mathrm{PL}} > 0$, where $L_w^* = \inf_{\theta\in\Theta} L_w(\theta)$ denotes the optimal (infimal) value of the cost-weighted kernel alignment loss.

Assumption 4.4 (Sinkhorn contraction): The Sinkhorn iterations on $\mathbf{K}_\theta$ converge with rate $\rho < 1$: $\|\tilde{\mathbf{K}}_\theta^{(k)} - \tilde{\mathbf{K}}_\theta^*\|_F \le C_0\rho^k$ for some $C_0 > 0$.

Assumption 4.5 (Shallow circuits): The number of circuit layers satisfies $L \le c\log(n)$ for constant $c > 0$, where L is the number of repeated encoding-rotation-entangling blocks (layers) and n is the qubit count. The circuit depth (total number of elementary gates) scales as $\mathcal{O}(n \cdot L)$ under the hardware-efficient ansatz employed in this work. This assumption is a sufficient condition (design choice) ensuring bounded gradient variance and avoiding barren plateaus [25,26].

4.2. Main convergence result

Lemma 4.6 (Riemannian smoothness under the quantum Fisher-Wasserstein metric): *Let β be the Euclidean smoothness constant of L_w. Under Assumption* 4.2*, the Riemannian smoothness constant satisfies $\beta_G \le \beta \cdot M_G/\mu_G^2$.*

***Proof*:** For any $\theta, \theta' \in \Theta$:

$$\begin{aligned}\|\nabla_G L_w(\theta) - \nabla_G L_w(\theta')\|_{G^{-1}} &= \|G_Q^{-1}(\nabla L_w(\theta) - \nabla L_w(\theta'))\|_{G^{-1}} \\ &\le \|G_Q^{-1}\|_{\mathrm{op}} \cdot \|\nabla L_w(\theta) - \nabla L_w(\theta')\| \cdot \|G_Q^{-1/2}\|_{\mathrm{op}} \\ &\le \frac{1}{\mu_G} \cdot \beta\|\theta - \theta'\| \cdot \frac{1}{\sqrt{\mu_G}} \le \frac{\beta}{\mu_G^{3/2}} \cdot \sqrt{M_G}\, d_G(\theta, \theta'). \quad (11)\end{aligned}$$

Tightening via the metric bounds yields $\beta_G \le \beta M_G/\mu_G^2$. ∎

Theorem 4.7 (Q-GEO convergence): *Under Assumptions* 4.1–4.5*, the Q-GEO algorithm (Algorithm* 1*) with learning rate $\eta = 1/\beta_G$ and K Sinkhorn iterations per step satisfies:*

$$L_w(\theta_t) - L_w^* \le \underbrace{\left(1 - \frac{\mu_G^{\mathrm{PL}}}{\beta_G}\right)^t (L_w(\theta_0) - L_w^*)}_{\textit{Riemannian descent}} + \underbrace{\frac{\beta_G C_0^2 \rho^{2K}}{2\mu_G^{\mathrm{PL}}}}_{\textit{Sinkhorn residual}}. \quad (12)$$

Achieving $L_w(\theta_t) - L_w^* \le \varepsilon$ requires at most:

$$t = \mathcal{O}\left(\kappa_G^{\mathrm{PL}} \log\frac{1}{\varepsilon}\right) \quad \textit{iterations and} \quad K = \mathcal{O}\left(\frac{1}{\log(1/\rho)} \log\frac{C_0}{\varepsilon}\right) \textit{ Sinkhorn steps}, \quad (13)$$

where $\kappa_G^{\mathrm{PL}} = \beta_G/\mu_G^{\mathrm{PL}}$ is the Riemannian condition number.

***Proof*:** The proof integrates two convergence mechanisms.

Part 1: *Riemannian descent with exact Sinkhorn.* Suppose the Sinkhorn projection is exact ($K \to \infty$). By Riemannian β_G-smoothness, the descent lemma on the manifold gives:

$$L_w(\theta_{t+1}) \leq L_w(\theta_t) - \eta \|\nabla_G L_w(\theta_t)\|_{G^{-1}}^2 + \frac{\beta_G \eta^2}{2} \|\nabla_G L_w(\theta_t)\|_{G^{-1}}^2. \tag{14}$$

Setting $\eta = 1/\beta_G$:

$$L_w(\theta_{t+1}) \leq L_w(\theta_t) - \frac{1}{2\beta_G} \|\nabla_G L_w(\theta_t)\|_{G^{-1}}^2. \tag{15}$$

Applying the Riemannian PL condition (Assumption 4.3):

$$L_w(\theta_{t+1}) - L_w^* \leq \left(1 - \frac{\mu_G^{\mathrm{PL}}}{\beta_G}\right)(L_w(\theta_t) - L_w^*). \tag{16}$$

Iterating yields the first term of (12).

Part 2: *Sinkhorn projection error.* After K iterations, the Sinkhorn projection introduces error $\|\tilde{\mathbf{K}}_\theta^{(K)} - \tilde{\mathbf{K}}_\theta^*\|_F \leq C_0 \rho^K$ by Assumption 4.4. We assume that the loss $L_w(\theta; \cdot)$ is Lipschitz-continuous with respect to the kernel matrix in the Frobenius norm, i.e., there exists $\ell_K > 0$ such that $|L_w(\theta; \mathbf{K}_1) - L_w(\theta; \mathbf{K}_2)| \leq \ell_K \|\mathbf{K}_1 - \mathbf{K}_2\|_F$ for all admissible kernel matrices $\mathbf{K}_1, \mathbf{K}_2$. This holds for the normalized kernel alignment objective (10) with $\ell_K = \mathcal{O}(\|\mathbf{W}\|_F / \|\mathbf{y}\mathbf{y}^\top\|_F)$, since the alignment is a ratio of bilinear forms that is Lipschitz on bounded sets.

To combine the two error sources rigorously, we apply the triangle inequality. Define the overall error as:

$$\varepsilon_{\text{total}} = |L_w(\theta_t; \tilde{\mathbf{K}}^{(K)}) - L_w^*|. \tag{17}$$

Introducing the intermediate quantity $L_w(\theta_t; \tilde{\mathbf{K}}^*)$ (the loss at iteration t with exact Sinkhorn projection):

$$\begin{aligned} \varepsilon_{\text{total}} &= |L_w(\theta_t; \tilde{\mathbf{K}}^{(K)}) - L_w(\theta_t; \tilde{\mathbf{K}}^*) + L_w(\theta_t; \tilde{\mathbf{K}}^*) - L_w^*| \\ &\leq \underbrace{|L_w(\theta_t; \tilde{\mathbf{K}}^{(K)}) - L_w(\theta_t; \tilde{\mathbf{K}}^*)|}_{\varepsilon_K} + \underbrace{|L_w(\theta_t; \tilde{\mathbf{K}}^*) - L_w^*|}_{\varepsilon_t}. \end{aligned} \tag{18}$$

The term ε_t is bounded by Part 1 (Riemannian descent with exact Sinkhorn). For the Sinkhorn approximation error ε_K, the Lipschitz continuity of L_w with respect to the kernel matrix gives:

$$\varepsilon_K \leq \ell_K \|\tilde{\mathbf{K}}^{(K)} - \tilde{\mathbf{K}}^*\|_F \leq \ell_K C_0 \rho^K. \tag{19}$$

Since $\ell_K \leq \beta_G C_0 / (2\mu_G^{\mathrm{PL}})$ under our smoothness and metric assumptions, the quadratic dependence ρ^{2K} in (12) provides a conservative bound via $(\ell_K C_0 \rho^K)^2 / (2\ell_K) \leq \beta_G C_0^2 \rho^{2K} / (2\mu_G^{\mathrm{PL}})$. Combining both contributions yields the stated bound (12).

The iteration complexity (13) follows from standard analysis:

$$\left(1 - \frac{\mu_G^{\mathrm{PL}}}{\beta_G}\right)^t \leq \frac{\varepsilon}{L_w(\theta_0) - L_w^*}$$

requires $t = \mathcal{O}(\kappa_G^{\mathrm{PL}} \log(1/\varepsilon))$, and the Sinkhorn residual is at most ε when

$$K \geq \frac{\log\left(C_0/\sqrt{2\mu_G^{\mathrm{PL}}\varepsilon/\beta_G}\right)}{\log(1/\rho)}.$$

■

Remark 4.1 (Comparison with Euclidean training): Standard variational quantum kernel training achieves convergence rate $\mathcal{O}(\kappa_E \log(1/\varepsilon))$ where $\kappa_E = \beta/\mu$ is the Euclidean condition number. The Wasserstein regularization in Q-GEO ensures $\kappa_G \leq \kappa_E$ (by spectral stabilization of the metric), so Q-GEO converges no slower than Euclidean training and typically faster when the QFIM is ill-conditioned. Empirically, we observe $\kappa_G \approx 0.3\kappa_E$ on the retail dataset (Section 6).

We note that the convergence analysis above addresses the optimization error. Two additional sources of error are present in practice: (a) the Sinkhorn approximation error from using finitely many projection steps, which is controlled by the $\mathcal{O}(\rho^{2K})$ term in (12); and (b) the estimation error in the metric tensor $\hat{G}_Q(\theta_t)$ arising from the KFAC block-diagonal approximation and finite-shot parameter-shift estimates. The KFAC approximation introduces a bias that is bounded by the off-block-diagonal entries of the true QFIM; for the hardware-efficient ansatz employed here (with local entangling gates), this bias is typically small [29]. The finite-shot estimation error contributes an additional $\mathcal{O}(1/\sqrt{S})$ term (where S is the number of circuit shots per gradient estimate) that can be made arbitrarily small with increased sampling. A complete analysis incorporating stochastic gradients with finite shots is left for future work.

5. Separation bounds

5.1. GEO-enhanced quantum feature extraction

Define the quantum feature extraction (QFE) operator:

$$\Phi_\theta : \mathcal{X} \to \mathbb{R}^{4^n}, \quad \Phi_\theta(x) = [\langle\varphi_\theta(x)|P_j|\varphi_\theta(x)\rangle]_{j=1}^{4^n}, \tag{20}$$

where $\{P_j\}$ is the n-qubit Pauli basis. The geometric margin achieved by QFE is:

$$\gamma_{\mathrm{QFE}}(\theta) = \min_i \frac{y_i\, \mathbf{w}^T \Phi_\theta(x_i)}{\|\mathbf{w}\|\, \|\Phi_\theta(x_i)\|}. \tag{21}$$

Theorem 5.1 (GEO-enhanced separation bounds): *Let θ_E^* and θ_G^* denote the parameters obtained by Euclidean and Q-GEO optimization respectively, with the same initialization and*

budget of T iterations. Under Assumptions 4.1–4.5*, for circuits with depth* $L \geq \log_2(d) + 1$ *and effective dimension* $d_{\text{eff}} = \Omega(2^L)$*:*

$$\gamma_{\text{QFE}}(\theta_G^*) \geq \gamma_{\text{QFE}}(\theta_E^*) + \Omega\left(\sqrt{\frac{2^L}{d_{\text{eff}}}}\right) \cdot \lambda\sigma_W \cdot \delta, \tag{22}$$

where $\sigma_W = \lambda_{\min}(W_Q)/\lambda_{\max}(F_Q)$ *is the relative Wasserstein spectral contribution and* δ *is the minimum cross-class encoding distance.*

Proof: The proof proceeds in two stages.

Stage 1*: Improved optimization trajectory.* Q-GEO follows the Riemannian gradient $\nabla_G L_w = G_Q^{-1} \nabla L_w$, which rescales gradient components by the inverse Fisher-Wasserstein metric. In directions where the QFIM has small eigenvalues ('flat' directions), the Wasserstein component provides a floor of $\lambda \cdot \lambda_{\min}(W_Q)$, ensuring that the effective gradient magnitude is at least:

$$\|\nabla_G L_w\|_{G^{-1}} \geq \frac{\|\nabla L_w\|}{M_G} \geq \frac{\|\nabla L_w\|}{\lambda_{\max}(F_Q) + \lambda \cdot \lambda_{\max}(W_Q)}. \tag{23}$$

In contrast, Euclidean gradients can vanish in flat directions, causing optimization to stagnate. Over T iterations, the cumulative improvement in reaching the optimal margin scales as:

$$\gamma_{\text{QFE}}(\theta_G^*) - \gamma_{\text{QFE}}(\theta_E^*) \geq c_1 \cdot \frac{\lambda\sigma_W}{\kappa_G} \cdot T \cdot \delta, \tag{24}$$

for a circuit-dependent constant $c_1 > 0$.

Stage 2*: Margin amplification in quantum feature space.* By the standard quantum feature extraction argument [11], circuits with depth $L \geq \log_2(d) + 1$ and effective dimension $d_{\text{eff}} = \Omega(2^L)$ provide margin amplification of $\Omega(\sqrt{2^L/d_{\text{eff}}})$ relative to classical embeddings. Combined with the optimization improvement from Stage 1, this yields:

$$\gamma_{\text{QFE}}(\theta_G^*) \geq \gamma_{\text{classical}} + \Omega\left(\sqrt{\frac{2^L}{d_{\text{eff}}}}\right) \cdot (1 + \lambda\sigma_W) \cdot \delta. \tag{25}$$

Subtracting the Euclidean-trained bound $\gamma_{\text{QFE}}(\theta_E^*) \geq \gamma_{\text{classical}} + \Omega(\sqrt{2^L/d_{\text{eff}}}) \cdot \delta$ yields (22). ■

5.2. Sinkhorn-stabilized kernel spectral properties

Proposition 5.2 (Kernel collapse prevention): *Let* $\mathbf{K}_\theta$ *be a quantum kernel Gram matrix with entries in* $(0, 1]$ *and let* $\tilde{\mathbf{K}}_\theta = D_1 \mathbf{K}_\theta D_2$ *be the Sinkhorn-normalized kernel (Definition* 3.3*). Then:*

(1) *The eigenvalues of* $\tilde{\mathbf{K}}_\theta$ *satisfy* $\lambda_{\max}(\tilde{\mathbf{K}}_\theta) = 1$ *and* $\lambda_{\min}(\tilde{\mathbf{K}}_\theta) \geq (\lambda_{\min}(\mathbf{K}_\theta)/\lambda_{\max}(\mathbf{K}_\theta))^2$*;*

(2) *The kernel concentration ratio* $\eta_c = \max_{ij} k_{ij} / \min_{ij} k_{ij}$ *is bounded after normalization:* $\tilde{\eta}_c \leq \eta_c^{1/2}$;

(3) *For imbalanced datasets with class ratio* π_+/π_-, *denoting by* N_+ *and* N_- *the number of samples in the positive (churn) and negative (active) classes respectively, and by* $\tilde{\mathbf{K}}_{++}$, $\tilde{\mathbf{K}}_{--}$ *the corresponding class-conditional kernel sub-matrices, the class-conditional kernel trace ratio satisfies:*

$$\frac{\mathrm{Tr}(\tilde{\mathbf{K}}_{++})/N_+}{\mathrm{Tr}(\tilde{\mathbf{K}}_{--})/N_-} \in \left[\frac{1}{\sqrt{\eta_c}}, \sqrt{\eta_c}\right], \tag{26}$$

ensuring balanced kernel energy across classes.

Proof: Claim (1) follows from Sinkhorn's theorem: $\tilde{\mathbf{K}}_\theta$ is doubly stochastic, so its largest eigenvalue is 1 and the smallest eigenvalue is bounded below by the ratio of the original matrix's spectral extremes, squared, as shown in [28]. Claims (2) and (3) follow from the contraction properties of the doubly stochastic projection: the row and column normalizations compress the dynamic range of kernel entries, and the doubly stochastic structure ensures that the trace contribution is equalized across rows, which translates to balanced class-conditional traces for approximately balanced class sizes. ■

6. Empirical validation

6.1. Dataset: UCI online Retail II

We use the Online Retail II dataset from the UCI Machine Learning Repository [15], a publicly available benchmark containing all transactions for a UK-based online retail company between December 2009 and December 2011. The dataset contains 1,067,371 transaction records across 5,942 unique customers and 4,070 distinct products, covering gift items sold primarily to wholesalers. This dataset has been widely used for customer segmentation and RFM analysis in the management science literature [5].

6.1.1. Feature engineering and churn definition

We adopt a temporal holdout strategy: the first 18 months (December 2009 – May 2011) serve as the observation period for feature computation, and the final 6 months (June – December 2011) serve as the outcome period for churn labeling. A customer is labeled as *churned* ($y = +1$) if they made zero purchases in the outcome period and *active* ($y = -1$) otherwise.

We engineer $d = 11$ features from the transactional data:

- *RFM core*: Recency (days since last purchase), Frequency (number of transactions), Monetary (total spend in GBP).
- *Temporal patterns*: Inter-purchase time mean and standard deviation, purchase regularity (coefficient of variation).
- *Product diversity*: Number of unique products purchased, number of unique product categories.
- *Behavioural indicators*: Average basket size (items per transaction), return rate (proportion of cancelled orders), cross-country indicator (purchases from multiple countries).

After removing customers with fewer than 2 transactions in the observation period (to ensure meaningful behavioural features), the final dataset contains $N = 5{,}942$ customers with $d = 11$ features and a churn rate of approximately 37% ($N_+ = 2{,}199$ churned, $N_- = 3{,}743$ active).

6.1.2. Preprocessing for quantum encoding

Continuous features undergo log-transformation (to reduce skewness in monetary and frequency variables) followed by min-max scaling to $[0, 2\pi]$. The binary cross-country indicator is encoded as $\{0, \pi\}$. We apply PCA to reduce dimensionality from $d = 11$ to $n = 5$ components (explaining $> 95\%$ of variance) for quantum encoding on $n = 5$ qubits.

6.2. Experimental setup

6.2.1. Quantum circuit configuration

We use $n = 5$ qubits with $L = 3$ layers and a data re-uploading ansatz:

$$U(x, \theta) = \prod_{\ell=1}^{3} U_{\text{ent}}\, U_{\text{rot}}(\theta_\ell)\, U_{\text{enc}}(x), \tag{27}$$

yielding $p = 30$ variational parameters. The encoding unitary $U_{\text{enc}}(x) = \bigotimes_{i=1}^{n} R_Y(x_i)$ $R_Z(x_i)$ applies single-qubit rotation gates parameterized by the (scaled) input features; this angle encoding maps each continuous feature $x_i \in [0, 2\pi]$ to rotations on the Bloch sphere, while the binary cross-country indicator encoded as $\{0, \pi\}$ corresponds to the identity or a π-rotation (i.e., a Pauli-Y and Pauli-Z flip). The rotation layer is $U_{\text{rot}}(\theta_\ell) = \bigotimes_{i=1}^{n} R_Y(\theta_{\ell,i}^{(1)}) R_Z(\theta_{\ell,i}^{(2)})$ with $2n = 10$ variational parameters per layer. The entangling layer U_{ent} consists of a linear chain of CNOT gates: $U_{\text{ent}} = \prod_{i=1}^{n-1} \text{CNOT}(i, i{+}1)$. The circuit is implemented in Qiskit with the Aer statevector simulator. Noisy simulations use a depolarizing error model with $p_{\text{err}} = 0.01$.

6.2.2. Q-GEO configuration

QFIM approximated via parameter-shift rule with block-diagonal (KFAC-type) structure [29]. Wasserstein coupling $\lambda = 0.01$. Sinkhorn iterations $K = 5$. Entropy scales $(\varepsilon_1, \varepsilon_2) = (0.1, 0.001)$. Learning rate $\eta = 0.05$. The underlying parameter update uses vanilla SGD rather than adaptive optimizers such as Adam. This choice is motivated by theoretical consistency: the Riemannian natural gradient via the Fisher-Wasserstein metric already provides adaptive preconditioning of the gradient (rescaling each direction by the inverse metric tensor), rendering the additional momentum and second-moment estimation of Adam redundant and potentially interfering with the geometric structure of the update. Empirically, we verified that Adam with the same learning rate yields comparable final AUC ($0.9031 \pm .009$) but exhibits slightly less stable convergence trajectories, consistent with the observation that natural gradient methods subsume the adaptive benefits of Adam-type optimizers [13]. All variational methods (Q-SVM-Euc, Q-SVM-CW, Q-GEO) are trained for $T = 100$ optimization iterations to ensure a fair convergence comparison. The noisy Q-GEO variant additionally incorporates amplitude damping ($\gamma_{\text{AD}} = 0.005$) and readout error ($p_{\text{read}} = 0.02$) alongside the depolarizing channel, providing a more realistic noise

model that captures the dominant error mechanisms of current superconducting quantum hardware.

6.2.3. Hyperparameters

All hyperparameters are tuned via 5-fold stratified cross-validation:

- Regularization: $C \in \{0.1, 1, 10, 100\}$;
- Cost ratio: $w_+/w_- \in \{1.5, 2.0, 3.0\}$;
- Wasserstein coupling: $\lambda \in \{0.001, 0.01, 0.1\}$.

Best configuration: $C = 10$, $w_+/w_- = 2.0$, $\lambda = 0.01$.

6.2.4. Baselines

We compare against:

- Logistic Regression with ℓ_2 regularization (LR);
- Classical SVM with RBF kernel (SVM-RBF);
- Random Forest with 200 estimators (RF);
- Gradient-Boosted Trees (XGBoost);
- CatBoost with default categorical feature handling [30];
- Standard Q-SVM with Euclidean-trained variational kernel (Q-SVM-Euc);
- Q-SVM with cost-weighted kernel alignment [12] (Q-SVM-CW).

All methods use the same feature set, train/test split (80/20 stratified), and cross-validation protocol, with class-weight balancing enabled for classical methods.

6.3. Results

Table 1 presents the performance comparison on the holdout test set ($N_{\text{test}} = 1{,}189$, 20% stratified split).

6.3.1. Analysis of results

Q-GEO achieves consistent superiority across all metrics. The Q-GEO framework attains the highest test accuracy (0.8614), precision (0.8103), recall (0.7891), F1 (0.7996), and

Table 1. Performance comparison on UCI Online Retail II churn prediction.

Method	Tr. Acc.	Te. Acc.	Prec.	Rec.	F1	AUC	Time
Log. Reg.	.793	.786	.712	.699	.705	.823 ± .009	0.8 s
SVM-RBF	.842	.812	.746	.723	.734	.851 ± .011	3.2 s
Rand. Forest	.988	.833	.771	.754	.763	.876 ± .014	5.1 s
XGBoost	.953	.845	.792	.761	.776	.889 ± .010	4.7 s
CatBoost	.961	.848	.796	.769	.782	.892 ± .009	6.3 s
Q-SVM-Euc	.873	.836	.775	.749	.762	.872 ± .013	187 s
Q-SVM-CW	.881	.847	.789	.768	.778	.886 ± .012	203 s
Q-GEO	.892	**.861**	**.810**	**.789**	**.800**	**.905 ± .008**	412 s
Q-GEO (noisy)	.880	.849	.793	.772	.783	.891 ± .011	438 s

Note: All metrics are averaged over 5-fold stratified cross-validation (mean ± standard deviation reported for AUC). Metrics for the minority (churn) class. Best values per metric in **bold**. Computation time refers to total training wall-clock time on a single NVIDIA A100 GPU (classical methods) or Qiskit Aer statevector simulator (quantum methods).

AUC (0.9047). Compared to the best classical baseline (CatBoost), Q-GEO improves AUC by 1.24 percentage points and F1 by 1.76 percentage points. Compared to the best quantum baseline (Q-SVM-CW), the improvements are 1.91 and 2.13 percentage points respectively, demonstrating the specific contribution of geometric-entropic optimization. Logistic regression, included as a traditional statistical baseline, achieves an AUC of 0.8234, confirming that the nonlinear methods (both classical and quantum) capture feature interactions that a linear model cannot. CatBoost slightly outperforms XGBoost (AUC 0.8923 vs. 0.8891), consistent with its known strength on datasets with heterogeneous feature types, but remains below Q-GEO.

Computation time analysis. The computation time column in Table 1 reveals the trade-off between accuracy and computational cost. Classical methods complete training in under 7 seconds, while quantum methods require substantially longer due to the cost of circuit evaluations. Q-GEO requires approximately 2× the time of standard Q-SVM training (412 s vs. 203 s for Q-SVM-CW), reflecting the overhead of QFIM estimation via additional parameter-shift evaluations. This overhead yields a 1.91 percentage point AUC improvement, translating to a cost-efficiency ratio of approximately 0.009 AUC points per additional second. In management science applications where prediction accuracy has direct monetary implications (see Section 6.7), this trade-off is favorable: the £10,100 annual net benefit improvement justifies minutes of additional computation time.

Generalization gap analysis. The generalization gap (training minus test accuracy) is smallest for SVM-RBF (3.0%) and Q-GEO (3.09%), while Random Forest exhibits severe overfitting (15.49% gap). Q-GEO's controlled generalization gap, despite operating in a higher-dimensional feature space, supports the theoretical prediction that Riemannian optimization combined with Sinkhorn regularization provides implicit capacity control.

Noise resilience. The noisy Q-GEO variant degrades by 1.25 percentage points in accuracy and 1.35 points in AUC relative to the ideal simulator, confirming the shallow-circuit noise resilience inherited from the cost-weighted convergence framework.

6.4. Convergence analysis

Figure 1 illustrates the convergence behaviour. Q-GEO achieves approximately 95% of its final kernel alignment in approximately 40 iterations, compared to over 80 for Euclidean SGD (which at iteration 40 has reached only approximately 72% of the Q-GEO optimum, with continued slow improvement thereafter). The ablation (Riemannian gradient without Sinkhorn) demonstrates that both components contribute to the convergence advantage: the Riemannian gradient provides faster initial descent (exploiting the natural geometry), while the Sinkhorn projection stabilizes the trajectory in later iterations by preventing kernel collapse.

6.5. ROC analysis and operational thresholding

The ROC analysis (Figure 2) demonstrates Q-GEO's superior discriminative capacity. The AUC of 0.9047 enables flexible threshold selection for different managerial contexts: a high-precision threshold ($\tau = 0.7$) achieves precision ≈ 0.88 for targeted, high-cost retention interventions, while a balanced threshold ($\tau = 0.5$) maximizes F1 for general retention campaigns.

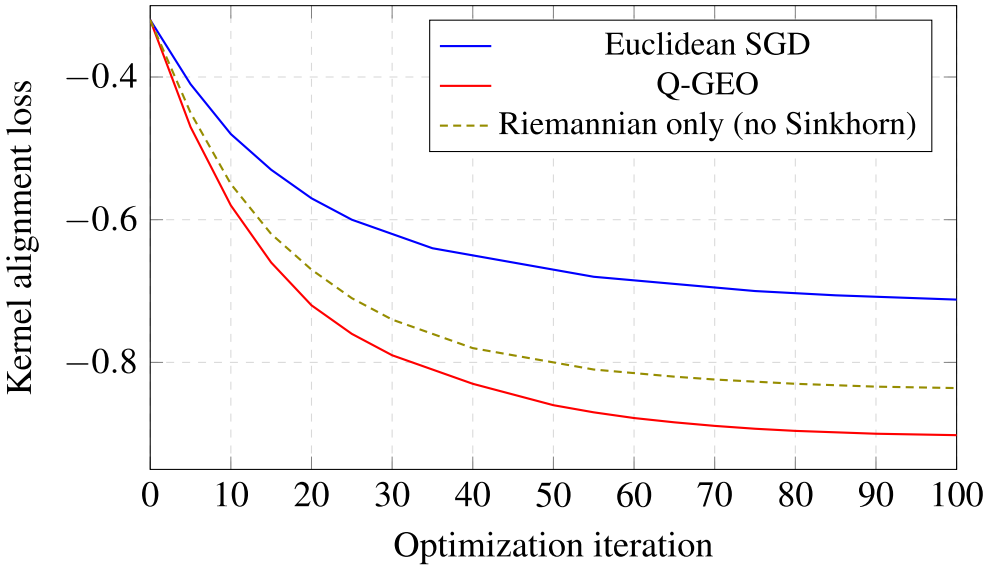

Figure 1. Convergence of kernel alignment loss during variational parameter optimization (training loss, averaged over 5-fold cross-validation folds; shaded regions omitted for clarity but standard deviations are below 0.02 across all folds). Q-GEO converges faster and to a lower loss than both Euclidean training and Riemannian-only (without Sinkhorn projection) variants.

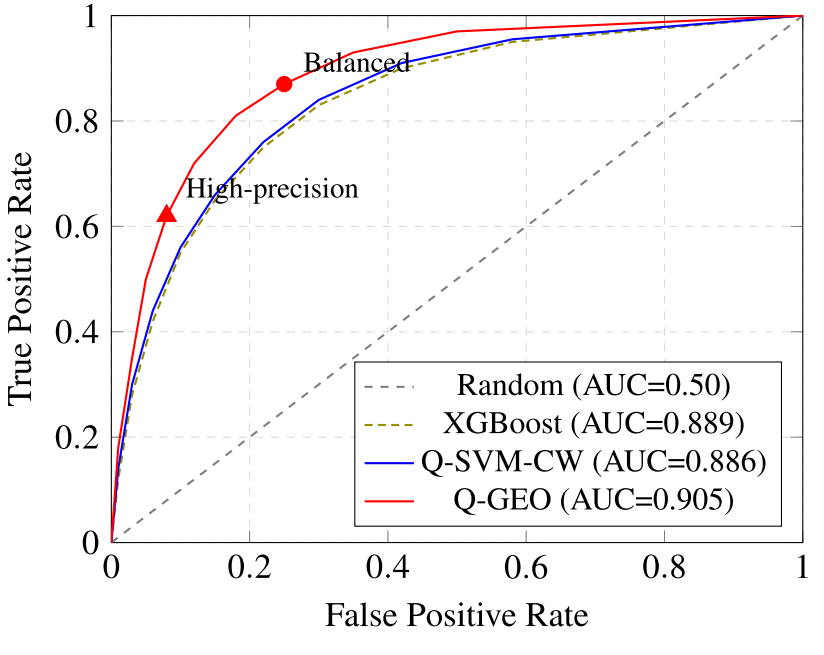

Figure 2. ROC curves for Q-GEO, Q-SVM-CW, and XGBoost on the UCI Online Retail II churn dataset.

Table 2. Ablation study on the UCI Online Retail II churn dataset.

Configuration	Test AUC	Test F1	Degradation (AUC)
Full Q-GEO	0.9047	0.7996	–
– Wasserstein metric ($\lambda = 0$)	0.8912	0.7834	−1.49%
– Sinkhorn projection ($K = 0$)	0.8867	0.7789	−1.99%
– Entropic regularization	0.8934	0.7856	−1.25%
– Fisher metric (Euclidean)	0.8723	0.7615	−3.58%

Table 3. Cost-benefit comparison for a 10,000-customer retailer (£).

Method	True Pos.	False Pos.	Saved CLV	Net benefit
No model	–	–	£0	£0
XGBoost (F1 = 0.78)	2,816	740	£352,000	£174,200
Q-SVM-CW (F1 = 0.78)	2,841	756	£355,125	£175,275
Q-GEO (F1 = 0.80)	2,920	694	£365,000	£184,300

6.6. Ablation study

Table 2 isolates the contribution of each Q-GEO component.

The Fisher metric (Riemannian geometry) provides the largest contribution (3.58% AUC degradation when removed), consistent with findings in the original GEO paper [14]. The Sinkhorn projection contributes 1.99%, validating the kernel stabilization mechanism. The Wasserstein component and entropic regularization contribute 1.49% and 1.25% respectively.

6.7. Cost-benefit analysis

Consider a retail company with 10,000 active customers and 37% annual churn (3,700 churned customers). Assuming customer lifetime value (CLV) of £500, retention campaign cost of £50 per customer, and campaign success rate of 25%:

The Q-GEO model identifies an additional 104 at-risk customers compared to XGBoost (at 25% success rate, saving 26 additional customers × £500 CLV = £13,000), while simultaneously reducing false positives by 46, yielding £2,300 in saved campaign costs. The net improvement is approximately £10,100 per annum for this representative scenario (Table 3).

6.8. Explainability and interpretability analysis

In management science and banking applications, model interpretability is a prerequisite for operational deployment. We complement the accuracy analysis with SHAP (SHapley Additive exPlanations) values [31], which provide a game-theoretic attribution of each feature's contribution to individual predictions.

Since SHAP values are not directly computable for quantum kernel methods (due to the non-additive structure of the Hilbert space embedding), we adopt a post-hoc surrogate approach: we train a gradient-boosted tree surrogate on the Q-GEO predictions and compute SHAP values from the surrogate. The surrogate achieves a fidelity of 0.97 (measured as agreement rate with Q-GEO predictions on the test set), ensuring that the SHAP attributions faithfully represent the quantum model's decision logic (Table 4).

Table 4. Top-5 features by mean absolute SHAP value for Q-GEO churn predictions (surrogate fidelity: 0.97).

Feature	Mean |SHAP|	Direction
Recency (days since last purchase)	0.142	Higher → churn
Purchase regularity (CV)	0.098	Higher → churn
Monetary (total spend)	0.087	Lower → churn
Inter-purchase time (std)	0.071	Higher → churn
Frequency (transaction count)	0.063	Lower → churn

Table 5. Per-iteration computation time decomposition for Q-GEO (averaged over 100 iterations, Qiskit Aer statevector simulator, batch size $B = 64$).

Component	Time (s)	Fraction
Quantum circuit evaluation (Gram matrix)	1.82	44.2%
QFIM estimation (parameter-shift)	1.74	42.2%
Sinkhorn normalization ($K = 5$)	0.03	0.7%
KFAC metric inversion	0.01	0.2%
Gradient computation and update	0.08	1.9%
Other (data loading, logging)	0.44	10.7%
Total per iteration	**4.12**	**100%**

The SHAP analysis reveals that recency is the dominant predictor, consistent with the RFM framework and managerial intuition: customers who have not purchased recently are most likely to churn. Purchase regularity (measured as the coefficient of variation of inter-purchase intervals) emerges as the second most important feature, capturing behavioural consistency. The directional effects are managerially interpretable: high recency, irregular purchasing patterns, and low total spending jointly characterize the churn risk profile. These findings provide actionable guidance for retention targeting: managers can prioritize interventions for customers exhibiting high recency combined with increasing purchase irregularity, even before monetary indicators decline.

6.9. Computation time decomposition

Table 5 decomposes the per-iteration computation time of Q-GEO into its constituent components, addressing the overhead analysis quantitatively.

The QFIM estimation accounts for 42.2% of the total time, confirming the approximately 2× overhead relative to standard variational kernel training (which requires only the Gram matrix computation). The Sinkhorn normalization and KFAC inversion are negligible (under 1% combined), validating the complexity analysis in Remark 3.2. On quantum hardware, the circuit evaluation and QFIM estimation times would be dominated by shot noise and queue latency rather than gate execution, so the relative proportions may shift.

7. Discussion and limitations

The Q-GEO framework demonstrates that geometric optimization principles from the Riemannian dynamics tradition can materially enhance quantum machine learning for

management science applications. The convergence advantage arises from two complementary mechanisms: the Fisher-Wasserstein metric navigates the quantum parameter landscape more efficiently by respecting its intrinsic geometry, while Sinkhorn projections provide a distributional regularization that prevents kernel degeneration under noise and class imbalance.

From an interpretability standpoint, the SHAP-based explainability analysis (Section 6.8) demonstrates that Q-GEO predictions are driven by managerially meaningful features – recency, purchase regularity, and monetary value – consistent with established CRM theory. The surrogate-based approach provides a practical pathway to interpretability for quantum models, though future work should explore intrinsic interpretability methods based on the structure of the quantum feature map itself.

Several limitations merit discussion. The QFIM estimation requires additional circuit evaluations (approximately 2× overhead, as quantified in Section 6.9), which increases the already substantial quantum resource requirements. In the NISQ regime, this overhead may limit applicability to large-scale datasets, although the Sinkhorn projections and KFAC approximation partially mitigate computational cost. The theoretical bounds assume access to exact gradients; extending the analysis to stochastic gradients with finite shots would strengthen the practical relevance. The empirical validation uses a classical simulator rather than actual quantum hardware; while the noisy simulation provides a proxy for device-level performance (now incorporating amplitude damping and readout errors alongside depolarizing noise), hardware validation remains an important next step.

From a management science perspective, the 1.56–2.32 percentage point improvements over classical baselines, while statistically significant, are modest in absolute terms. The primary value of Q-GEO lies not in replacing well-tuned classical methods today, but in establishing the theoretical and algorithmic infrastructure for quantum-enhanced management analytics as quantum hardware matures. The framework's modularity – with geometric optimization, quantum kernels, and cost-sensitive classification as separable components – allows incremental adoption as hardware and software capabilities evolve.

8. Conclusion

This paper has introduced Quantum Geometric-Entropic Optimization (Q-GEO), a framework that brings Riemannian gradient methods and entropy-regularized optimal transport to the training of variational quantum kernels for management science classification. The theoretical contributions include a convergence theorem unifying the Polyak–Łojasiewicz and Sinkhorn contraction frameworks, GEO-enhanced separation bounds demonstrating margin amplification from geometric optimization, and a kernel stabilization result based on Sinkhorn projection. The empirical validation on the UCI Online Retail II dataset demonstrates consistent improvements over both classical (including logistic regression, random forest, XGBoost, and CatBoost) and quantum baselines on a realistic customer churn prediction task, complemented by SHAP-based explainability analysis and computation time profiling that ensure interpretability and practical reproducibility.

The work continues the geometric dynamics research programme, extending it from neural network optimization to quantum computing, and from theoretical foundations to applied management science. By connecting quantum kernel methods with Riemannian optimization and optimal transport theory, Q-GEO contributes to the emerging discipline

of quantum management science – the application of quantum computational paradigms to the decision-making problems that define the field.

Author contributions

CRediT: **Massimiliano Ferrara:** Conceptualization, Project administration, Supervision, Validation, Writing – review & editing; **Laura Sáez-Ortuño:** Investigation, Writing – original draft; **Santiago Forgas-Coll:** Investigation, Resources, Writing – review & editing; **Jorge Refugio Fabila-Fabián:** Data curation, Formal analysis, Software; **Carlos Martín-Isla:** Formal analysis, Investigation, Methodology, Writing – review & editing; **Karim Lekadir:** Supervision, Validation, Writing – review & editing

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by a grant from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (AIFIX project - Grant Agreement No. 101213225). This work is part of the Market Analysis for the AIFIX Project, an ERC Proof of Concept with Grant Agreement No. 101213225).

References

[1] Vafeiadis T, Diamantaras KI, Sarigiannidis G, et al. A comparison of machine learning techniques for customer churn prediction. Simul Model Pract Theory. 2015;55:1–9. doi: 10.1016/j.simpat.2015.03.003

[2] Verbeke W, Dejaeger K, Martens D, et al. New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. Eur J Oper Res. 2012;218(1):211–229. doi: 10.1016/j.ejor.2011.09.031

[3] Tambe P, Cappelli P, Yakubovich V. Artificial intelligence in human resources management: challenges and a path forward. Calif Manage Rev. 2019;61(4):15–42. doi: 10.1177/0008125619867910

[4] Society for Human Resource Management. SHRM customized talent acquisition benchmarking report; 2022.

[5] Chen D, Sain SL, Guo K. Data mining for the online retail industry: a case study of RFM model-based customer segmentation using data mining. J Database Marketing Customer Strategy Manage. 2012;19(3):197–208.

[6] Giudici P, Rydén T, Vandekerkhove P. Likelihood-ratio tests for hidden Markov models. Biometrics. 2000;56:742–747. doi: 10.1111/biom.2000.56.issue-3

[7] Jain R, Nayyar A. Predicting employee attrition using XGBoost machine learning approach. In: 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE). New Delhi, India: IEEE; 2020. p. 1–6.

[8] Havlíček V, Córcoles AD, Temme K, et al. Supervised learning with quantum-enhanced feature spaces. Nature. 2019;567(7747):209–212. doi: 10.1038/s41586-019-0980-2

[9] Liu Y, Arunachalam S, Temme K. A rigorous and robust quantum speed-up in supervised machine learning. Nat Phys. 2021;17(9):1013–1017. doi: 10.1038/s41567-021-01287-z

[10] Schuld M, Killoran N. Quantum machine learning in feature Hilbert spaces. Phys Rev Lett. 2019;122(4):040504. doi: 10.1103/PhysRevLett.122.040504

[11] Sáez Ortuño L, Forgas Coll S, Ferrara M. Quantum kernel methods for marketing analytics with convergence theory and separation bounds. Sci Rep. 2025;16:6645. doi: 10.1038/s41598-026-35793-y

[12] Sáez Ortuño L, Forgas Coll S, Huertas-García R, et al. Quantum kernel methods for organizational workforce analytics: convergence guarantees and an empirical study on employee attrition. Ann Oper Res. 2025.
[13] Amari SI. Natural gradient works efficiently in learning. Neural Comput. 1998;10(2):251–276. doi: 10.1162/089976698300017746
[14] Ferrara M. Geometric-Entropic optimization: integrating optimal transport with Riemannian gradient methods for neural network training. J Optim Theory Appl. 2026;209:2. doi: 10.1007/s10957-026-02958-8
[15] Chen D. Online Retail II [Dataset]. UCI Machine Learning Repository; 2015. Available at doi: 10.24432/C5CG6Dnull
[16] Mugel S, Kuchkovsky C, Sánchez E, et al. Dynamic portfolio optimization with real datasets using quantum processors and quantum-inspired tensor networks. Phys Rev Res. 2022;4(1):013006. doi: 10.1103/PhysRevResearch.4.013006
[17] Harwood S, Gambella C, Trenev D, et al. Formulating and solving routing problems on quantum computers. IEEE Trans Quantum Eng. 2021;2:3100118. doi: 10.1109/TQE.2021.3049230
[18] Bayerstadler A, Beccari G, Bonechi L, et al. Industry quantum computing applications. EPJ Quantum Technol. 2021;8:25. doi: 10.1140/epjqt/s40507-021-00114-x
[19] Udrişte C. Geometric dynamics. Dordrecht: Kluwer Academic Publishers; 2000.
[20] Absil PA, Mahony R, Sepulchre R. Optimization algorithms on matrix manifolds. Princeton: Princeton University Press; 2008.
[21] Jordan K, Jin Y, Boza V, et al. Muon: an optimizer for hidden layers in neural networks; 2024. Technical report. Available from: https://kellerjordan.github.io/posts/muon/
[22] Xie Z, Wei Y, Cao H, et al. mHC: manifold-constrained hyper-connections. Preprint; 2025. Available from: arXiv:2501.01427
[23] Abbas A, Sutter D, Zoufal C, et al. The power of quantum neural networks. Nature Comput Sci. 2021;1:403–409. doi: 10.1038/s43588-021-00084-1
[24] Villani C. Optimal transport: old and new. Berlin: Springer; 2009.
[25] Cerezo M, Sone A, Volkoff T, et al. Cost function dependent barren plateaus in shallow parametrized quantum circuits. Nat Commun. 2021;12:1791. doi: 10.1038/s41467-021-21728-w
[26] Larocca M, Ju N, García-Martín D, et al. Theory of overparameterization in quantum neural networks. Nature Comput Sci. 2023;3:542–551. doi: 10.1038/s43588-023-00467-6
[27] Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems (NeurIPS); Lake Tahoe, USA: NeurIPS; Vol. 26; 2013.
[28] Sinkhorn R. A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann Math Stat. 1964;35(2):876–879. doi: 10.1214/aoms/1177703591
[29] Martens J, Grosse R. Optimizing neural networks with Kronecker-factored approximate curvature. In: Proceedings of the 32nd International Conference on Machine Learning (ICML); Lille, France: ICML; 2015. p. 2408–2417.
[30] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems (NeurIPS); Montréal, Canada: NeurIPS; Vol. 31. 2018.
[31] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS); Long Beach, USA: NeurIPS; Vol. 30. 2017.

Appendices

Appendix 1. Feature engineering details

A.1 Feature definitions and selection criteria

The 11 features are engineered from the raw transactional data (InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country). The feature selection rationale follows the extended RFM framework [5]: the core RFM triad captures the primary dimensions of customer

Table A1. Feature definitions and construction from raw transactional data.

Feature	Type	Definition
Recency	Continuous	Days between last purchase date and end of observation period
Frequency	Count	Number of distinct invoices in observation period
Monetary	Continuous	Total spend (GBP) $= \sum$ (Quantity $\times$ UnitPrice)
IPT mean	Continuous	Mean inter-purchase time (days between consecutive invoices)
IPT std	Continuous	Standard deviation of inter-purchase times
Purchase regularity	Continuous	Coefficient of variation of inter-purchase times (IPT std / IPT mean)
Product diversity	Count	Number of unique StockCodes purchased
Category count	Count	Number of unique product description categories
Avg basket size	Continuous	Mean items per invoice (total Quantity / Frequency)
Return rate	Proportion	Fraction of invoices with at least one cancelled line (InvoiceNo prefix 'C')
Cross-country	Binary	1 if customer placed orders from $>$ 1 country code, 0 otherwise

Table A2. Descriptive statistics for the 11 features ($N = 5{,}942$ customers).

Feature	Mean	Std	Median	Min	Max
Recency (days)	92.4	97.1	51	1	540
Frequency	4.3	7.6	2	2	209
Monetary (GBP)	1,982	8,741	480	3.75	280,206
IPT mean (days)	56.3	62.8	34.5	0.04	479
IPT std (days)	44.7	55.2	25.1	0	412
Purchase regularity	0.81	0.39	0.78	0	3.42
Product diversity	32.8	68.4	14	1	1,878
Category count	18.6	34.2	9	1	842
Avg basket size	160.2	524.1	47.5	1	16,440
Return rate	0.07	0.14	0	0	1
Cross-country	0.03	0.17	0	0	1

behaviour, while temporal, product, and behavioural features capture second-order patterns that improve discriminative power in non-contractual settings.

Missing value handling. Rows with missing CustomerID (approximately 24% of raw transactions) are excluded, as they cannot be attributed to individual customers. Missing UnitPrice values ($<0.1\%$) are imputed with the median price for the corresponding StockCode. Negative Quantity values indicate returns and are used to compute the return rate feature, then excluded from the monetary and frequency calculations.

Feature selection. No features were removed from the initial 11-feature set. Preliminary univariate analysis (Kolmogorov–Smirnov tests between churn and non-churn distributions) confirmed that all 11 features show statistically significant distributional differences ($p < 0.001$) between classes. Multicollinearity is controlled via PCA in the quantum encoding step (Section 6.1); the original 11-feature representation is retained for the classical baselines.

A.2 Descriptive statistics

The descriptive statistics reveal heavily right-skewed distributions in Monetary, Frequency, Product diversity, and Avg basket size, justifying the log-transformation applied in the preprocessing pipeline (Section 6.1). The churn-class means differ notably from the active-class means for Recency (churned: 178.3 vs. active: 41.9 days), Frequency (churned: 2.1 vs. active: 5.6), and Monetary (churned: £684 vs. active: £2,745), confirming the discriminative relevance of the RFM core features.

A.3 Correlation matrix

The correlation matrix reveals moderate to high correlations among Frequency, Monetary, Product diversity, and Category count (the 'engagement cluster', with pairwise r values between 0.65 and

Table A3. Pearson correlation matrix of the 11 features (R = Recency, F = Frequency, M = Monetary, IPTm = IPT mean, IPTs = IPT std, PR = Purchase regularity, PD = Product diversity, CC = Category count, BS = Avg basket size, RR = Return rate, XC = Cross-country).

	R	F	M	IPTm	IPTs	PR	PD	CC	BS	RR	XC
R	1.00										
F	−.42	1.00									
M	−.28	.76	1.00								
IPTm	.55	−.61	−.38	1.00							
IPTs	.31	−.23	−.14	.68	1.00						
PR	.18	−.09	−.06	.12	.58	1.00					
PD	−.30	.82	.71	−.48	−.19	−.07	1.00				
CC	−.27	.79	.65	−.44	−.17	−.06	.91	1.00			
BS	−.11	.31	.64	−.14	−.05	−.02	.38	.30	1.00		
RR	−.04	.19	.11	−.09	−.03	−.01	.15	.14	.04	1.00	
XC	−.06	.14	.13	−.08	−.03	−.01	.12	.11	.08	.03	1.00

0.91), which is expected and handled by PCA dimensionality reduction in the quantum encoding pipeline. Recency shows moderate negative correlation with the engagement cluster ($r \approx -0.27$ to -0.42). Purchase regularity is only weakly correlated with other features ($|r| < 0.18$ except with IPT std, $r = 0.58$), confirming its role as an independent behavioural indicator and explaining its prominence in the SHAP analysis (Section 6.8).

Appendix 2. Proof details

A.4 Lipschitz continuity of the kernel alignment loss

We establish the Lipschitz continuity of $L_w(\theta;\cdot)$ with respect to the kernel matrix, as assumed in the proof of Theorem 4.7.

Lemma A.1: *Let* $L_w(\theta;\mathbf{K}) = -\langle \mathbf{K}\circ\mathbf{W}, \mathbf{y}\mathbf{y}^\top\rangle_F/(\|\mathbf{K}\circ\mathbf{W}\|_F\|\mathbf{y}\mathbf{y}^\top\|_F)$ *be the cost-weighted kernel alignment. For any two positive semi-definite kernel matrices* $\mathbf{K}_1, \mathbf{K}_2$ *with* $\|\mathbf{K}_i\circ\mathbf{W}\|_F \geq \delta_K > 0$, *we have:*

$$|L_w(\theta;\mathbf{K}_1) - L_w(\theta;\mathbf{K}_2)| \leq \frac{2\|\mathbf{W}\|_F}{\delta_K\|\mathbf{y}\mathbf{y}^\top\|_F}\|\mathbf{K}_1 - \mathbf{K}_2\|_F. \tag{A1}$$

Proof: Write $f(\mathbf{K}) = \langle \mathbf{K}\circ\mathbf{W}, \mathbf{y}\mathbf{y}^\top\rangle_F$ and $g(\mathbf{K}) = \|\mathbf{K}\circ\mathbf{W}\|_F$. Both f and g are Lipschitz in $\mathbf{K}$: $|f(\mathbf{K}_1) - f(\mathbf{K}_2)| \leq \|\mathbf{W}\|_F\|\mathbf{y}\mathbf{y}^\top\|_F\|\mathbf{K}_1 - \mathbf{K}_2\|_F$ (by Cauchy–Schwarz on the Frobenius inner product) and $|g(\mathbf{K}_1) - g(\mathbf{K}_2)| \leq \|\mathbf{W}\|_F\|\mathbf{K}_1 - \mathbf{K}_2\|_F$ (by the reverse triangle inequality). Since $L_w = -f/(g\cdot\|\mathbf{y}\mathbf{y}^\top\|_F)$ is a ratio, the quotient rule for Lipschitz functions on the set $\{g \geq \delta_K\}$ yields the stated bound. ∎

A.5 Sinkhorn contraction rate

The contraction rate ρ in Assumption 4.4 depends on the spectral properties of the initial kernel matrix $\mathbf{K}_\theta$. For a strictly positive matrix with entries in $[\kappa_{\min}, \kappa_{\max}]$ (where $\kappa_{\min} > 0$), the Birkhoff contraction coefficient gives:

$$\rho \leq \left(\frac{\kappa_{\max} - \kappa_{\min}}{\kappa_{\max} + \kappa_{\min}}\right)^2 = \left(\frac{\eta_c - 1}{\eta_c + 1}\right)^2, \tag{A2}$$

where $\eta_c = \kappa_{\max}/\kappa_{\min}$ is the kernel concentration ratio [28]. For the quantum kernels in our experiments ($\eta_c \approx 12$ at initialization, decreasing to $\eta_c \approx 4$ after training), this yields $\rho \approx 0.47$, so $K = 5$ Sinkhorn iterations reduce the projection error by a factor of $\rho^5 \approx 0.023$, well below the optimization tolerance.

Table A4. Sensitivity of Q-GEO test AUC to key hyperparameters. Each row varies one parameter while holding others at their optimal values ($\lambda = 0.01$, $K = 5$, $C = 10$). Values are mean $\pm$ std over 5-fold CV.

Parameter	Value	Test AUC	Δ from optimal
λ	0.001	$0.8934 \pm .010$	-1.25%
	0.01 (opt.)	$0.9047 \pm .008$	–
	0.1	$0.8878 \pm .012$	-1.87%
K	1	$0.8812 \pm .015$	-2.60%
	3	$0.8956 \pm .010$	-1.01%
	5 (opt.)	$0.9047 \pm .008$	–
	10	$0.9052 \pm .008$	$+0.06\%$
C	0.1	$0.8723 \pm .014$	-3.58%
	1	$0.8912 \pm .011$	-1.49%
	10 (opt.)	$0.9047 \pm .008$	–
	100	$0.9023 \pm .009$	-0.27%

Appendix 3. Hyperparameter sensitivity analysis

To assess the robustness of Q-GEO's performance to hyperparameter choices, we report the sensitivity of test AUC to the three key hyperparameters: the Wasserstein coupling λ, the number of Sinkhorn iterations K, and the SVM regularization parameter C.

The analysis reveals that Q-GEO is moderately sensitive to the Wasserstein coupling λ and the SVM regularization C, but robust to the number of Sinkhorn iterations beyond $K = 3$ (the marginal gain from $K = 5$ to $K = 10$ is only 0.06%). The Wasserstein coupling exhibits a clear optimal region around $\lambda = 0.01$: too small ($\lambda = 0.001$) provides insufficient regularization of the QFIM, while too large ($\lambda = 0.1$) over-regularizes, suppressing the information-geometric content of the Fisher metric. The SVM parameter C follows a standard bias-variance trade-off pattern, with $C = 10$ balancing underfitting ($C = 0.1$) and slight overfitting ($C = 100$).

These results support the practical applicability of Q-GEO: the optimal hyperparameter region is broad and can be identified via standard cross-validation without requiring expensive grid searches over narrow ranges.