Mathematics

AIMS Mathematics, 11(10): 33635–33672.
DOI: 10.3934/math.20261316
Received: 08 August 2026
Revised: 20 September 2026
Accepted: 22 September 2026
Published: 15 October 2026

https://www.aimspress.com/journal/Math

Research article

Entropy-regularized Riemannian hyperstate transport with adaptive nonlocal tunneling for stochastic deep neural optimization

Mashail M. AL Sobhi[1,*] and Irsa Sajjad[2,*]

[1] Department of Mathematics, Umm Al-Qura University, Makkah 24227, Saudi Arabia
[2] Department of Mathematics, Faculty of Engineering and Computing, National University of Modern Languages, Islamabad, Pakistan

* **Correspondence:** Email: mmsobhi@uqu.edu.sa, irsa.sajjad@numl.edu.pk.

Abstract: This study investigates an improved quantum-inspired stochastic geometric hyperstate optimization (QISGHO) method for stochastic deep neural optimization, where learning is viewed as the evolution and transport of multiple interacting probabilistic hyperstates on a Riemannian statistical manifold. The proposed optimization framework is based on the general stochastic-geometric hyperstate optimization paradigm, with a more explicit entropy-regularized Riemannian probability-transport mechanism, an interaction potential for hyperstates, geometry and uncertainty-adaptive diffusion, and an acceptance-controlled nonlocal tunneling process with loss barriers, manifold distance, local curvature, probability mass, and annealing temperature. The ensemble also includes a collapse-prevention, consolidation, and rejuvenation mechanism to preserve diversity within a hyperstate. The dynamics of these components are then linked to nonlinear Riemannian Fokker–Planck and nonlocal master-equation descriptions, thereby offering a population-level perspective on stochastic optimization, and are unified by a variational free-energy formulation. The framework establishes a principled balance between global exploration and local exploitation, optimization stability, uncertainty awareness, and geometric consistency. The paired-run experiments compare and contrast the algorithm's predictive performance, calibration, robustness to perturbations, optimization behavior, and computational cost with SGD, Adam, AdamW, Lion, and SGLD on one synthetic benchmark and four public ones. The reported estimates and multiplicity-controlled comparisons suggest the advantages of the full method and its components for the following architectures and computational budgets. While these are robust theoretical findings, the conclusions do not imply global convergence for a generic non-convex neural-network objective under the sum of diffusion, jump, consolidation, and rejuvenation perturbations.

Keywords: quantum-inspired optimization; deep neural networks; stochastic optimization; information geometry; hyperstate learning; optimal transport
Mathematics Subject Classification: 60H10, 68T07

1. Introduction

Deep neural networks have revolutionized computer vision, natural-language processing, scientific computing, autonomous systems, and data-driven decision-making by learning highly nonlinear representations directly from complex observations [1,2]. However, in practice, they largely rely on the tail of the optimization procedure to estimate millions or billions of trainable parameters. Neural network loss functions are generally high-dimensional, nonconvex, anisotropic, and stochastic, with sharp minima, flat plateaus, narrow valleys, saddle points, and multiple attraction basins [3]. As a result, training can converge slowly, be sensitive to initialization, produce unstable gradients, converge early, or perform poorly on noisy data.

Stochastic gradient descent remains widely considered the cornerstone of neural network optimization because it is cheap to compute and well-suited to mini-batch learning [3]. Adaptive techniques like AdaGrad and Adam adjust learning rates using information from accumulated parameter gradients [4,5]. Some Adam update sequences can cause convergence failures; to address this, AMSGrad was introduced [6], and AMSGrad shares this goal with AdamW, which further decouples weight decay from adaptive gradient scaling [7]. Lion later showed that the sign-based momentum updates learned using symbolic program search are competitive with popular adaptive optimizers [8]. Although successful in practice, they typically simulate a single parameter vector and are largely based on the local gradient information. Thus, they have limited ability to encode uncertainty, maintain potentially viable alternative solutions, and make structured transitions between distant regions of the loss landscape.

Stochastic gradient sampling offers a more general view of optimization. Stochastic gradient Langevin dynamics (SGLD) [9] adds diffusion noise to the gradient updates, bringing together neural training, Bayesian sampling and stochastic differential equations. Another interpretation of constant-step stochastic-gradient dynamics is approximating a stationary parameter distribution through Bayesian inference [10]. Langevin-based methods can help escape certain local minima, but are typically specified in Euclidean parameter space and may not account for the highly nonuniform geometry of deep predictive models [11].

To address this limitation, a statistical manifold represents model parameters in information geometry. The natural gradient rescales the ordinary gradient by the local sensitivity of the predictive distribution [12]. This yields parameterization-aware updates and connects optimization, probability theory, and differential geometry [13]. At the same time, optimal transport theory provides an elegant, mathematically sound vocabulary for studying how probability distributions move through space. The Jordan–Kinderlehrer–Otto formulation sees diffusion as a Wasserstein gradient flow of free energy [14] and entropy-regularized optimal transport and Sinkhorn iterations render the distributional transport numerically accessible [15,16]. These developments suggest that one can think of optimizing neural networks as evolving a probability distribution over parameter configurations, rather than evolving a single point.

Another conceptual basis is quantum machine learning. Quantum systems model information in terms of probabilistic amplitudes, superposition, interference, and measurement, and generate computational structures that differ fundamentally from classical, deterministic systems [17]. One major approach for near-term quantum computation is variational quantum algorithms, which use parameterized quantum circuits and classical optimization [18]. In practice, qubit availability, noise, finite circuit depth, and the lack of fully fault-tolerant hardware limit deployment [19]. However, parameterized quantum models can also suffer from barren plateaus, where gradient magnitudes rapidly decrease to near zero [20], and hardware noise can further worsen trainability [21].

Because of these limitations, researchers have encouraged quantum-inspired classical optimization and the power of data in quantum machine learning [22–24]. Quantum-inspired algorithms do not require actual quantum processors; they map computational concepts like probability amplitudes, populations in superpositions, tunneling, and probabilistic state transitions into algorithms that can run on traditional processors [25,26]. These methods are especially applicable to training neural networks, where many candidate states can traverse different portions of a non-convex space and share information during optimization.

Sajjad and Al Sobhi (2026) proposed quantum-inspired adaptive superposition optimization, in which amplitude-based probabilistic adaptation retains multiple neural network configurations [27]. Later, Sajjad et al. proposed adaptive Grover-driven optimization of quantum-inspired deep learning, a gradient-free classical deep learning framework that combines Grover-inspired search, discretization adaptation, and fidelity-based regulation [28]. Both articles are published in AIMS Mathematics and lay the foundations of a direct methodological approach to extending quantum-inspired neural optimization to stochastic geometry and interacting probability distributions.

However, an important research gap remains: while most classical optimizers follow a single local trajectory [29], many stochastic optimizers rely on isotropic diffusion, natural-gradient methods typically exclude nonlocal tunneling, and existing quantum-inspired optimizers rarely offer a unified probability-manifold formulation. Furthermore, researchers often apply explainability after training rather than before. The key point is that the new stochastic perturbation mechanism is not isolated. Instead, QISGHO proposes redefining neural network learning as an interacting probability-flow problem on a curved statistical manifold. The framework is designed to achieve the following in a highly nonlinear learning environment, with respect to exploration, convergence stability, generalization, calibration, and interpretability: connections to information geometry, stochastic differential equations, optimal transport, quantum-inspired search, and uncertainty-aware explanation. In contrast to the previous QSGHO formulation [30], which established the general concept of entropy-regularized stochastic hyperstate optimization, the present QISGHO formulation brings forth an explicit Fisher-Riemannian probability-flow model with interacting hyperstate potentials, geometry-adaptive diffusion, acceptance-controlled nonlocal tunneling, diversity-preserving consolidation, and population-level Riemannian Fokker–Planck dynamics; more differences can be seen in Table 1.

Table 1. Key differences between the previous QSGHO study and the present QISGHO framework.

Aspect	Previous QSGHO study [30]	Present QISGHO study
Core framework	Established QISGHO for deep neural learning.	Extends the framework into an explicit entropy-regularized Riemannian hyperstate transport formulation.
Optimization dynamics	Probabilistic hyperstate evolution with entropy-controlled diffusion and tunneling.	Adds explicit Riemannian probability transport, hyperstate interaction, and geometry- and uncertainty-adaptive diffusion.
Nonlocal exploration	Introduced adaptive, quantum-inspired tunneling to escape local minima.	Develops curvature- and barrier-aware nonlocal tunneling with controlled acceptance probability and manifold distance.
Mathematical formulation	Stochastic manifold optimization with convergence and stability analysis.	Further formulates the dynamics through variational free energy, Riemannian Fokker–Planck and nonlocal master-equation representations.
Validation	Primarily demonstrated on nonlinear stochastic benchmark settings.	Adds broader benchmark evaluation, computational analysis, and systematic component-level ablation to validate the new mechanisms.

This study addresses this gap using quantum-inspired stochastic geometric hyperstate optimization (QISGHO). The framework comprises several probability states (hyperstates) that interact with one another, rather than a single parameter vector. These hyperstates emerge via: entropy-regularized probability transport, Riemannian natural-gradient drift, adaptive diffusion, nonlocal tunneling, free-energy minimization and probabilistic consolidation. The whole thing is classical, and the quantum concepts are used only as optimization analogies and do not imply physical quantum computation. This aligns with the manuscript's goal of merging hyperstate representation, entropy-regulated manifold transport, tunneling diffusion, and explainable probabilistic dynamics in a single learning architecture. In recent years, quantum learning research has focused on model aspects directly relevant to robustness and representation, moving beyond general statements of quantum advantage. Recently, Li et al. [31] proposed a quantum (K)-nearest neighbor formulation in which a quantum circuit simultaneously selects the value of (K) and the (K) nearest neighbors. This is an adaptive observation-neighborhood method, where the state is the training observations and the task is to compute a neighborhood for classification. Li et al. [32] have also proposed a quantum support vector machine with sample-dependent weighting, which adds sample-dependent weighting to the hinge-loss objective and recovers the separating hyperplane using a quantum circuit. Later, Xu and Li [33] introduced nonlinear quantum feature selection via multikernel circuits, incorporating representation selection into the quantum learning procedure.

2. Review of literature

2.1. Gradient-based and adaptive optimization

The adoption of modern deep learning has closely followed advances in stochastic optimization. Large-scale neural training is made possible by approximating the full gradient through random mini-batch sampling [3]. However, its effectiveness hinges heavily on learning-rate schedules, momentum, gradient noise, and landscape conditioning. AdaGrad, proposed by the authors of [4], adapts the step size per coordinate based on the sum of past squared gradients. This mechanism is useful when little learning is required, but in a continuous process it can significantly reduce learning power. Adam combines the first and second moments with exponentially weighted averages, yielding efficient adaptive updates with low computational complexity [5]. Later work showed that in certain stochastic settings, Adam does not converge, leading to AMSGrad [6]. AdamW also decouples weight decay from the gradient update [7]. To show that some relatively simple optimizer structures can achieve competitive results, Lion uses sign-based momentum and stores fewer auxiliary variables than Adam. Nevertheless, adaptive optimizers remain local, single-state optimizers. Their moment estimates characterize gradient history, not a distribution over alternative neural configurations. This can cause them to get caught in sharp basins, or in basins with different widths or heights depending on initialization, especially in loss surfaces with irregular curvature or significant stochastic noise.

2.2. Stochastic dynamics and probabilistic optimization

Stochastic gradient Langevin dynamics is a variation of gradient descent with Gaussian noise that approximates posterior sampling under appropriate assumptions [9]. Constant-step stochastic-gradient dynamics can be understood as approximate variational inference around stationary regions [10]. Raginsky et al. [11] provided theoretical results on the convergence of nonconvex Langevin algorithms.

These studies provide a useful foundation for using stochastic diffusion for uncertainty-aware exploration. However, when the parameter space is Euclidean, researchers use ordinary Langevin diffusion. Identical Euclidean parameter shifts can produce substantially different shifts in the prediction distribution in deep networks. Fixed diffusion can then make direction changes too sensitive or too insensitive. This work solves the problem by scaling drift and diffusion by a local Riemannian metric. Amari's natural gradient computes the gradient in the direction of the Fisher-preconditioned normal of the underlying statistical model, rather than the usual gradient [12]. Martens later provided a detailed theoretical analysis of natural-gradient methods and their links to second-order optimisation [13]. An alternative perspective is that the model parameters are points on a “statistical manifold”, with the Fisher information metric as the local metric. Information-geometric optimization matters especially when parameter effects are strongly anisotropic. However, even conventional natural-gradient approaches still use only one parameter point and cannot necessarily maintain multiple optimization hypotheses. In QISGHO, multiple probabilistically weighted hyperstates share the same statistical manifold and can interact during training.

2.3. Entropy-regularized optimal transport

The optimal transport problem involves transporting a probability measure from one distribution to another at minimum cost. In the Wasserstein-gradient-flow framework, diffusion and free-energy dissipation can be modelled as variational motion in probability space [14]. Cuturi introduced entropy regularization and showed that computing optimal transport is much faster with regularization than without it, using Sinkhorn-type matrix scaling [15]. Peyré and Cuturi later combined discrete and continuous optimal transport into a single method [16].

The direct relevance of entropy regularization to exploration control is evident. High entropy spreads the likelihood across alternative solutions, while low entropy concentrates it on dominant ones. More specifically, in QISGHO, entropy is not just a secondary penalty; it drives the hyperstate ensemble from wide exploration to narrow exploitation.

2.4. Quantum and quantum-inspired learning

Quantum machine learning explores the idea that quantum representations might enhance learning, optimization or kernel construction [17]. Variational quantum algorithms rely on classical optimizers to train parameterized quantum circuits [18]. Quantum neural networks could have high expressive power, and the functions they can represent are closely related to how data are encoded. Quantum kernels also have clear inductive biases, and they can leverage information from quantum feature spaces. However, practical quantum learning remains difficult. NISQ devices suffer from decoherence, gate errors, limited connectivity, and circuit depth [19]. Optimization may be impractical on barren plateaus because gradient variance drops with problem size [20]. Noise alone can also cause barren plateaus [21]. Even when gradients are measurable, they may trap quantum optimization in local minima. These challenges highlight the importance of classical, quantum-inspired methods that mimic certain search mechanisms without relying on quantum hardware.

QISGHO combines these traditions and sees neural optimization as a collective evolution of weighted hyperstates. Probability weight updates use entropy-regularized transport, parameter locations are determined by Riemannian stochastic dynamics, the tunneling action enables adaptive nonlocal transitions, and hyperstate uncertainty enables probabilistic explanation. This framework is thus not only a new optimizer, but also a mathematical model of the dynamics of uncertainty, geometry, probability mass and energy in neural learning.

3. Methodology

3.1. Methodological overview and problem formulation

In this study, we develop a mathematically unified QISGHO framework for training deep neural networks under nonlinear, nonconvex, and stochastic learning conditions (see Figure 1). A key methodological insight is that no single parameter trajectory represents neural network optimization. Instead, we model the trainable system as an interacting probability measure over multiple possible configurations of trainable parameters, called hyperstates. These hyperstates co-evolve on a curved statistical manifold alongside entropy-regularized probability transport, Riemannian preconditioning, nonlocal tunneling transitions, stochastic diffusion, and free-energy dissipation. This formulation

retains the main ideas and terminology of the submitted manuscript, probabilistic hyperstates, stochastic manifold dynamics, adaptive tunneling, entropy-controlled evolution, and explainable feature dynamics, and embeds them into a single coherent variational optimization problem. Suppose that we are given a training set $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ denote a training dataset, where $x_n \in \mathbb{R}^{d_x}$ is the input vector and $y_n \in \mathcal{Y}$is its corresponding response. Let $f_\theta: \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ be a deep neural network parameterized by $\theta \in \Theta \subseteq \mathbb{R}^p$, where p denotes the number of trainable parameters. The empirical risk is defined as

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \ell\,(f_\theta,(x_n)y_n) + \frac{\lambda_w}{2} \parallel \theta \parallel_2^2, \tag{1}$$

where $\ell(\cdot,\cdot)$ is the prediction loss for the specific task, and $\lambda_w \geq 0$ is a coefficient that controls the weight decay. The prediction loss is for multiclass classification with classes.

$$\ell\,(f_\theta,(x_n)y_n) = -\sum_{c=1}^{C} y_{nc} \log p_\theta\,(y_n = c \mid x_n). \tag{2}$$

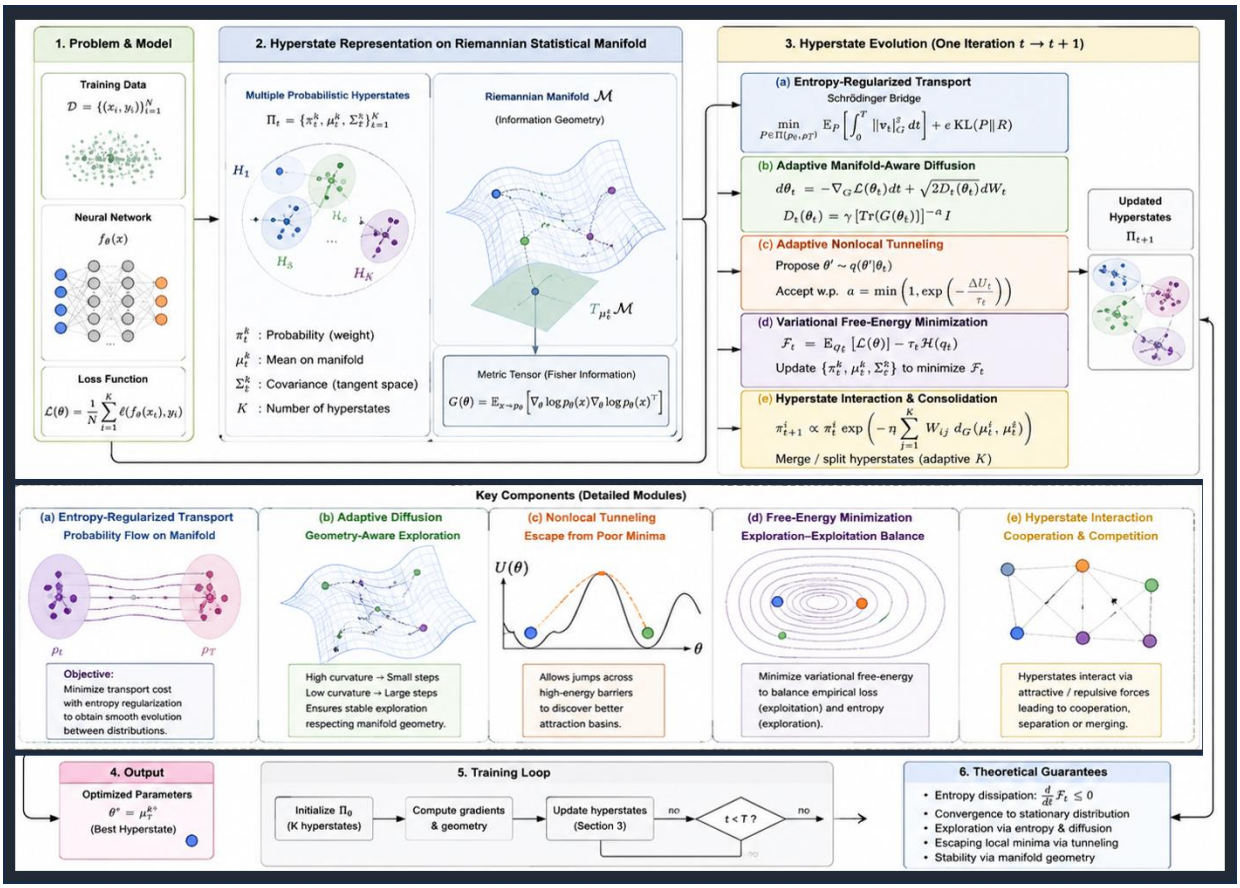


Figure 1. Model architecture of the proposed study.

The traditional optimizers look for one minimizer,

$$\theta^\star \in \arg\min_{\theta \in \Theta} \mathcal{L}(\theta). \tag{3}$$

The proposed methodology can address this point-estimate problem by formulating an optimization

problem over a space of probability distribution $\rho \in \mathcal{P}_2(\Theta)$, where $\mathcal{P}_2(\Theta)$ with finite second moment. The learning problem is a Distributional learning problem, and the corresponding solution is

$$\rho^\star \in \arg\min_{\rho \in \mathcal{P}_2(\Theta)} \mathcal{F}(\rho), \tag{4}$$

where $\mathcal{F}$ is a stochastic-geometric free-energy functional. As a result, training is defined as transporting and progressively accumulating probability mass on the neural-parameter manifold, not moving a single parameter vector.

3.2. Quantum-iInspired hyperstate ensemble

At iteration t, the neural system is represented by Kcandidate parameter states,

$$\mathcal{H}_t = \left\{\theta_t^{(1)}, \theta_t^{(2)}, \ldots, \theta_t^{(K)}\right\},\ \theta_t^{(k)} \in \Theta. \tag{5}$$

These candidates are given a complex-valued quantum-inspired amplitude.

$$\alpha_t^{(k)} = \sqrt{\pi_t^{(k)}}\, e^{\,i\phi_t^{(k)}}, \tag{6}$$

where $\pi_t^{(k)}$is its probability mass and $\phi_t^{(k)}$is an auxiliary phase variable. The amplitudes satisfy

$$\sum_{k=1}^{K} \left|\alpha_t^{(k)}\right|^2 = \sum_{k=1}^{K} \pi_t^{(k)} = 1,\ \pi_t^{(k)} \geq 0. \tag{7}$$

So the formal superposition of the complete optimization state is written as the complete optimization state.

$$|\Psi_t\rangle = \sum_{k=1}^{K} \alpha_t^{(k)} \left|\theta_t^{(k)}\right\rangle. \tag{8}$$

Equation (8) is not a physical quantum state and is not dependent on quantum hardware. It is a classical probabilistic construction in which multiple candidate solutions exist and evolve in parallel, inspired by quantum superposition. The hyperstate ensemble is related to the empirical probability measure.

$$\rho_t = \sum_{k=1}^{K} \pi_t^{(k)}\, \delta_{\theta_t^{(k)}}, \tag{9}$$

where $\delta_\theta$denotes a Dirac measure centered at θ. The ensemble prediction for an observation x is

$$p_{\rho_t}(y = c \mid x) = \sum_{k=1}^{K} \pi_t^{(k)}\, p_{\theta_t^{(k)}}(y = c \mid x). \tag{10}$$

This means that prediction uncertainty and optimization uncertainty are kept during training. The centre of mass of the complete system is

$$\bar{\theta}_t = \mathbb{E}_{\rho_t}[\theta] = \sum_{k=1}^{K} \pi_t^{(k)}\, \theta_t^{(k)}, \tag{11}$$

and the hyperstate covariance weighted by the diagonal elements of the hyperstate covariance matrix.

$$\Sigma_t = \sum_{k=1}^{K} \pi_t^{(k)} \left(\theta_t^{(k)} - \bar{\theta}_t\right)\left(\theta_t^{(k)} - \bar{\theta}_t\right)^{\top}. \tag{12}$$

The trace $\mathrm{tr}\,(\Sigma_t)$ quantifies geometric diversity in parameter space, whereas the Shannon entropy

$$\mathcal{H}(\pi_t) = -\sum_{k=1}^{K} \pi_t^{(k)} \log \pi_t^{(k)} \tag{13}$$

quantifies probabilistic diversity.

3.3. Information-geometric parameter manifold

The parameter space is treated as a differentiable manifold

$$\mathcal{M} = (\Theta, G). \tag{14}$$

It is a space provided with a positive definite Riemannian metric tensor $G(\theta)$. A regularised Fisher-information metric is chosen for a predictive model $p_\theta(y \mid x)$,

$$G(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\nabla_\theta \log p_\theta\,(y \mid x) \nabla_\theta \log p_\theta\,(y \mid x)^{\top}] + \delta I_p, \tag{15}$$

where $\delta > 0$ guarantees strict positive definiteness. In large neural networks, the entire matrix in (15) can be approximated by a diagonal, block-diagonal, Kronecker factored or low-rank matrix. The gradient of the Riemannian form is

$$\mathrm{grad}_{\mathcal{M}} \mathcal{L}(\theta) = G(\theta)^{-1} \nabla_\theta \mathcal{L}(\theta). \tag{16}$$

This construction rescales the update based on the local sensitivity of the predictive distribution. Weakly sensitive directions get relatively larger motion, while others that produce large changes in model behaviour are damped. The square of the infinitesimal distance between two neighbouring parameter settings is

$$ds^2 = d\theta^{\top} G(\theta) d\theta. \tag{17}$$

This metric induces a geodesic distance.

$$d_{\mathcal{M}}(\theta_a, \theta_b) = \inf_{\gamma} \int_0^1 \sqrt{\dot{\gamma}(s)^{\top} G\big(\gamma(s)\big)\dot{\gamma}(s)}\; ds, \tag{18}$$

where the infimum is taken over all smooth curves satisfying $\gamma(0) = \theta_a$ and $\gamma(1) = \theta_b$.

To quantify local geometric irregularity, a curvature proxy is defined by

$$\kappa_t^{(k)} = \frac{\left\|\nabla_\theta^2 \mathcal{L}\left(\theta_t^{(k)}\right)\right\|_F}{1 + \left\|\nabla_\theta \mathcal{L}\left(\theta_t^{(k)}\right)\right\|_2}, \tag{19}$$

where $\|\cdot\|_F$ is the Frobenius norm. Later, this number controls diffusion and tunnelling so that the exploration intensity is influenced by the geometry of the loss landscape.

3.4. Unified stochastic free-energy functional

Optimization dynamics are created by minimizing a single free-energy functional. Let the function ρ be a probability measure,

$$\begin{aligned}\mathcal{F}_t(\rho) = \int_\Theta \mathcal{L}(\theta)\, d\rho(\theta) + \tau_t\, \mathrm{KL}(\rho \parallel \rho_0) \\ + \frac{\lambda_c}{2} \iint_{\Theta\times\Theta} \mathcal{K}(\theta,\vartheta)\, d\rho(\theta)\, d\rho(\vartheta) + \lambda_g \mathcal{R}_g(\rho).\end{aligned} \tag{20}$$

In this case, the first term represents the expected prediction loss. The second term is a divergence from a reference measure ρ_0, where $\tau_t > 0$ is an annealed temperature. The third term captures interactions between hyperstates, followed by the fourth term, which regularizes geometric instability. The Kullback–Leibler divergence is

$$\mathrm{KL}(\rho \parallel \rho_0) = \int_\Theta \log\left(\frac{d\rho}{d\rho_0}\right) d\rho. \tag{21}$$

The interaction kernel is repulsive-attractive if

$$\mathcal{K}(\theta,\vartheta) = \omega_a d_{\mathcal{M}}(\theta,\vartheta)^2 - \omega_r \exp\left[-\frac{d_{\mathcal{M}}(\theta,\vartheta)^2}{2h_t^2}\right], \tag{22}$$

where $\omega_a, \omega_r \geq 0$and $h_t > 0$. Its radial derivative is

$$\frac{\partial \mathcal{K}_{\mathrm{old}}}{\partial r} = 2\omega_a r + \frac{\omega_r r}{h_t^2}\exp\left(-\frac{r^2}{2h_t^2}\right) > 0,\ r > 0.$$

Therefore, under the negative-gradient drift, both terms are attractive. The current equation is not consistent with the claimed attractive repulsive interpretation. Using $\mathrm{grad}_{\mathcal{M},\theta}\left[\frac{1}{2} d_{\mathcal{M}}(\theta,\vartheta)^2\right] = -\mathrm{Log}_\theta(\vartheta)$, the negative-gradient interaction force is

$$-\mathrm{grad}_{\mathcal{M},\theta}\mathcal{K}_t(\theta,\vartheta) = \left[\omega_a - \frac{\omega_r}{h_t^2}\exp\left(-\frac{d_{\mathcal{M}}(\theta,\vartheta)^2}{2h_t^2}\right)\right]\mathrm{Log}_\theta(\vartheta).$$

When $\omega_r > \omega_a h_t^2$, the force is repulsive at short distances and attractive at sufficiently large distances. The preferred equilibrium separation is separation is separation is $r_t^\star = h_t\sqrt{2\log\left(\frac{\omega_r}{\omega_a h_t^2}\right)}$. The first promotes anti-fragmentation when the hyperstates are too spread out, while the second slows early collapse and avoids empty spaces. The geometric regularizer is

$$\mathcal{R}_g(\rho) = \int_\Theta \left\| G(\theta)^{-1/2} \nabla_\theta \mathcal{L}(\theta) \right\|_2^2 d\rho(\theta). \tag{23}$$

For the discrete empirical measure in Eq (9), the functional becomes

$$\begin{aligned}\mathcal{F}_t(\Theta_t, \pi_t) = &\sum_{k=1}^{K} \pi_t^{(k)} \mathcal{L}\left(\theta_t^{(k)}\right) + \tau_t \sum_{k=1}^{K} \pi_t^{(k)} \log \frac{\pi_t^{(k)}}{\pi_0^{(k)}} \\ &+ \frac{\lambda_c}{2} \sum_{k=1}^{K} \sum_{j=1}^{K} \pi_t^{(k)} \pi_t^{(j)} \mathcal{K}\left(\theta_t^{(k)}, \theta_t^{(j)}\right) \\ &+ \lambda_g \sum_{k=1}^{K} \pi_t^{(k)} \left\| G\left(\theta_t^{(k)}\right)^{-1/2} \nabla \mathcal{L}\left(\theta_t^{(k)}\right) \right\|_2^2,\end{aligned} \tag{24}$$

where $\Theta_t = \left[\theta_t^{(1)}, \ldots, \theta_t^{(K)}\right]$. Equation (24) is the central optimization objective of the proposed methodology. It simultaneously minimizes prediction risk, regulates probability concentration, preserves controlled ensemble diversity, accounts for information geometry, and stabilizes the stochastic trajectories.

3.5. Entropy-regularized hyperstate amplitude transport

For fixed hyperstate locations, the probability masses are updated through KL-proximal mirror descent:

$$\pi_{t+1} = \arg\min_{\pi \in \Delta^{K-1}} \left\{ \langle \pi, E_t \rangle + \frac{\tau_t}{\eta_{\pi,t}} \mathrm{KL}(\pi \parallel \pi_t) \right\}. \tag{25}$$

The simplex is

$$\Delta^{K-1} = \left\{ \pi \in \mathbb{R}^K : \pi^{(k)} \geq 0 \ \sum_{k=1}^{K} \pi^{(k)} = 1 \right\} \tag{26}$$

is the probability simplex and E_t is the effective hyperstate-energy vector. Its k-th component is self-interactions:

$$E_t^{(k)} = \mathcal{L}\left(\theta_t^{(k)}\right) + \lambda_c \sum_{\substack{j=1 \\ j \neq k}}^{K} \pi_t^{(j)} \mathcal{K}_t\left(\theta_t^{(k)}, \theta_t^{(j)}\right) + \lambda_g r_t^{(k)}. \tag{27}$$

with

$$r_t^{(k)} = \left\| G\left(\theta_t^{(k)}\right)^{-1/2} \nabla \mathcal{L}\left(\theta_t^{(k)}\right) \right\|_2^2. \tag{28}$$

The closed-form solution of Eq (25) is

$$\pi_{t+1}^{(k)} = \frac{\left(\pi_t^{(k)}\right)^{\frac{\tau_t}{\tau_t + \eta_{\pi,t}}} \exp\left[-\frac{\eta_{\pi,t}}{\tau_t + \eta_{\pi,t}} E_t^{(k)} \right]}{\sum_{j=1}^{K} \left(\pi_t^{(j)}\right)^{\frac{\tau_t}{\tau_t + \eta_{\pi,t}}} \exp\left[-\frac{\eta_{\pi,t}}{\tau_t + \eta_{\pi,t}} E_t^{(j)} \right]}. \tag{29}$$

A simplified Gibbs transport form is obtained when the KL proximal exponent is absorbed into the temperature:

$$\pi_{t+1}^{(k)} = \frac{\pi_t^{(k)} \exp\left(-\eta_{\pi,t} E_t^{(k)} / \tau_t\right)}{\sum_{j=1}^{K} \pi_t^{(j)} \exp\left(-\eta_{\pi,t} E_t^{(j)} / \tau_t\right)}. \tag{30}$$

The temperature follows a bounded annealing schedule,

$$\tau_t = \tau_{\min} + (\tau_0 - \tau_{\min}) \exp(-\gamma_\tau t),\ \tau_0 > \tau_{\min} > 0. \tag{31}$$

If the initial temperature is high, the probability of any one candidate is diminished, and if the final temperature is low, it focuses probability towards stable low-energy hyperstates.

3.6. Riemannian drift and interacting hyperstate dynamics

The deterministic component of each hyperstate trajectory is defined by a natural-gradient drift. Let

$$g_t^{(k)} = \nabla_\theta \mathcal{L}\left(\theta_t^{(k)}; \mathcal{B}_t\right) \tag{32}$$

be a mini-batch stochastic gradient computed on batch $\mathcal{B}_t$. The Riemannian preconditioned gradient is

$$\tilde{g}_t^{(k)} = G\left(\theta_t^{(k)}\right)^{-1} g_t^{(k)}. \tag{33}$$

Theinteraction potential experienced by the k-th hyperstate is

$$U_t^{(k)} = \sum_{\substack{j=1 \\ j \neq k}}^{K} \pi_t^{(j)} \mathcal{K}_t\left(\theta_t^{(k)}, \theta_t^{(j)}\right). \tag{34}$$

and its manifold gradient is

$$q_t^{(k)} = \mathrm{grad}_{\mathcal{M}} U_t^{(k)}. \tag{35}$$

The deterministic drift is therefore

$$b_t^{(k)} = -\tilde{g}_t^{(k)} - \lambda_c q_t^{(k)} - \lambda_b\left(\theta_t^{(k)} - \bar{\theta}_t\right), \tag{36}$$

where $\lambda_b \geq 0$ regulates the contraction towards the weighted barycenter. The first term reduces prediction errors, the second term ensures cooperative manifold organization, and the third term mitigates uncontrolled dispersion.

3.7. Geometry-adaptive stochastic diffusion

The hyperstates are modeled as undergoing Itô diffusion on the parameter manifold to keep exploration controlled. The continuous-time dynamics are

$$d\theta_t^{(k)} = b_t^{(k)}\,dt + \sqrt{2\beta_t^{(k)}}\,G\left(\theta_t^{(k)}\right)^{-1/2} dW_t^{(k)} + dJ_t^{(k)}, \tag{37}$$

where $W_t^{(k)}$ is a p-dimensional Brownian motion, $\beta_t^{(k)}$ is an adaptive diffusion intensity, and $J_t^{(k)}$ is a nonlocal tunneling jump process. The three adaptive factors in Eq (38) have different functions. The relative-suboptimality factor $s_t^{(k)}$ assigns more exploratory noise to comparatively poor hyperstates, while the normalized entropy $u_t$maintains exploration when probability mass remains distributed across several candidate states. In contrast, the factor $\exp\left[-\gamma_\kappa \kappa_t^{(k)}\right]$ decreases as local curvature increases. The adaptive diffusion increment is therefore suppressed in sharply curved regions, where large stochastic displacements may be unstable. Accordingly, relative suboptimality and ensemble entropy increase diffusion, whereas local curvature decreases it. The diffusion coefficient depends on uncertainty, geometric curvature, and relative hyperstate quality:

$$\beta_t^{(k)} = \beta_{\min} + (\beta_{\max} - \beta_{\min}) s_t^{(k)} u_t \exp\left(-\gamma_\kappa \kappa_t^{(k)}\right), \tag{38}$$

where

$$u_t = \frac{\mathcal{H}(\pi_t)}{\log K} \in [0, 1] \tag{39}$$

is normalized hyperstate entropy and

$$s_t^{(k)} = \frac{\mathcal{L}\left(\theta_t^{(k)}\right) - \mathcal{L}_t^{\min}}{\mathcal{L}_t^{\max} - \mathcal{L}_t^{\min} + \varepsilon} \tag{40}$$

is a relative measure of suboptimality. The exploratory diffusion is thus larger for the poor candidates, while the good candidates are being refined with smaller perturbations. The Euler–Maruyama discretization of Eq (37) (without the jump part) is

$$\tilde{\theta}_{t+1}^{(k)} = \theta_t^{(k)} + \eta_{\theta,t} b_t^{(k)} + \sqrt{2\eta_{\theta,t}\beta_t^{(k)}}\,G\left(\theta_t^{(k)}\right)^{-1/2} \xi_t^{(k)}, \tag{41}$$

where

$$\xi_t^{(k)} \sim \mathcal{N}(0, I_p). \tag{42}$$

Because the noise is multiplied by $G^{-1/2}$, which makes stochastic exploration anisotropic and sensitive to the local information geometry, but not all parameter directions.

3.8. Adaptive nonlocal tunneling operator

In the case of hyperstates which trap inside sharp separation attraction basins, local Gaussian diffusion might be inadequate. Therefore, a nonlocal jump mechanism is introduced. Let bk be a barrier score for candidate k,

$$B_t^{(k)} = \left[\mathcal{L}\left(\theta_t^{(k)}\right) - \mathcal{L}_t^{\mathrm{ref}}\right]_+ + \zeta_\kappa \kappa_t^{(k)}, \tag{43}$$

where $[z]_+ = \max(z, 0)$, and

$$\mathcal{L}_t^{\mathrm{ref}} = \sum_{j=1}^{K} \pi_t^{(j)} \mathcal{L}\left(\theta_t^{(j)}\right). \tag{44}$$

For every configuration to which the asymptotic theory applies, set $q_{\min} = 0$, and use

$$q_t^{(k)} = q_0 e^{-\gamma_q t} \sigma\left(a_B B_t^{(k)} + a_s s_t^{(k)} + a_u u_t - b_q\right),\ \gamma_q > 0.$$

However, $0 < \sigma(z) < 1$, it follows that $0 \le q_t^{(k)} \le q_0 e^{-\gamma_q t}$, and

$$\sum\nolimits_{t=0}^{\infty} q_t^{(k)} \le q_0 \sum\nolimits_{t=0}^{\infty} e^{-\gamma_q t} = \frac{q_0}{1-e^{-\gamma_q}} < \infty. \tag{45}$$

A positive tunneling floor may be used in a prespecified finite-horizon exploratory experiment. However, when $q_{\min} > 0$, the tunneling probabilities are not summable, and Theorem 3 and the associated asymptotic stationarity statements do not apply. All configurations used to support the asymptotic analysis therefore set $q_{\min} = 0$. The tunneling probability is

$$q_t^{(k)} = q_{\min} + (q_0 - q_{\min}) \exp\left(-\gamma_q t\right) \sigma\left(a_B B_t^{(k)} + a_s s_t^{(k)} + a_u u_t - b_q\right),$$

where

$$\sigma(z) = (1 + e^{-z})^{-1}.$$

The exponential factor gradually suppresses global jumps as training progresses, while the sigmoid term increases tunneling for candidates located in high-loss or highly curved regions. When a Bernoulli variable

$$Z_t^{(k)} \sim \mathrm{Bernoulli}\left(q_t^{(k)}\right), \tag{46}$$

takes the value one, a destination hyperstate $j \neq k$ is selected according to

$$\omega_{kj,t} = \frac{\pi_t^{(j)} \exp\left[-\mathcal{L}\left(\theta_t^{(j)}\right)/\tau_J\right] \exp\left[-d_{\mathcal{M}}\left(\theta_t^{(k)}, \theta_t^{(j)}\right)^2/\left(2h_j^2\right)\right]}{\sum_{r \neq k} \pi_t^{(r)} \exp\left[-\mathcal{L}\left(\theta_t^{(r)}\right)/\tau_J\right] \exp\left[-d_{\mathcal{M}}\left(\theta_t^{(k)}, \theta_t^{(r)}\right)^2/\left(2h_j^2\right)\right]}. \tag{47}$$

The proposed destination is generated in the tangent space by

$$\theta_{t+1,\mathrm{prop}}^{(k)} = \mathrm{Exp}_{\tilde{\theta}_{t+1}^{(k)}}\left[\chi_t^{(k)} \mathrm{Log}_{\tilde{\theta}_{t+1}^{(k)}}\left(\theta_t^{(j)}\right) + \sigma_J G^{-1/2} \zeta_t^{(k)}\right], \tag{48}$$

where Exp and Log are the manifold exponential and logarithmic maps, $\chi_t^{(k)} \in (0, 1]$ controls

jump length, and $\zeta_t^{(k)} \sim \mathcal{N}(0, I)$. The proposal is accepted with probability

$$A_t^{(k)} = \min\left\{1, \exp\left[-\frac{\mathcal{L}\left(\theta_{t+1,\text{prop}}^{(k)}\right) - \mathcal{L}\left(\tilde{\theta}_{t+1}^{(k)}\right)}{\tau_{J,t}}\right]\right\}. \tag{49}$$

Thus, improvements are accepted automatically, while a controlled proportion of uphill transitions remains possible during high-temperature stages. The final state is

$$\theta_{t+1}^{(k)} = \begin{cases} \theta_{t+1,\text{prop}}^{(k)}, & \text{with probability } Z_t^{(k)} A_t^{(k)}, \\ \tilde{\theta}_{t+1}^{(k)}, & \text{otherwise.} \end{cases} \tag{50}$$

Equations (43)–(50) differ from the unstructured random restart mechanisms in that they explicitly consider the loss, the curvature, the probability mass, the manifold distance, and the annealing temperature, when and where and to what degree a nonlocal transition should occur.

3.9. Hyperstate consolidation without irreversible collapse

Through optimization, a number of candidates could become close to each other. Keeping all such states increases computing cost without contributing much diversity. As a result, a soft consolidation procedure is employed. The size of the ensemble is

$$K_{\text{eff},t} = \exp[\mathcal{H}(\pi_t)]. \tag{51}$$

A pair of hyperstates (k, j) is eligible for merging when

$$d_{\mathcal{M}}\left(\theta_t^{(k)}, \theta_t^{(j)}\right) < \varepsilon_\theta \tag{52}$$

and

$$\left|\mathcal{L}\left(\theta_t^{(k)}\right) - \mathcal{L}\left(\theta_t^{(j)}\right)\right| < \varepsilon_{\mathcal{L}}. \tag{53}$$

The merged probability is

$$\pi_t^{(k \oplus j)} = \pi_t^{(k)} + \pi_t^{(j)}, \tag{54}$$

and the merged location is the local Riemannian barycenter

$$\theta_t^{(k \oplus j)} = \arg\min_{\theta \in \mathcal{M}} \left\{\pi_t^{(k)} d_{\mathcal{M}}\left(\theta, \theta_t^{(k)}\right)^2 + \pi_t^{(j)} d_{\mathcal{M}}\left(\theta, \theta_t^{(j)}\right)^2\right\}. \tag{55}$$

To prevent irreversible collapse, a minimum number $K_{\min}$ of active hyperstates is maintained. When $K_{\text{eff},t} < K_{\min}$, low-probability candidates are rejuvenated around the dominant state:

$$\theta_t^{(\text{new})} = \text{Exp}_{\theta_t^\star}\left[\sigma_{\text{re}} G(\theta_t^\star)^{-1/2} \epsilon\right], \ \epsilon \sim \mathcal{N}(0, I), \tag{56}$$

where

$$\theta_t^{\star} = \theta_t^{(k^{\star})},\ k^{\star} = \arg\max_k \pi_t^{(k)}. \tag{57}$$

The proposed consolidation mechanism therefore eliminates redundancies and maintains a controlled stochastic neighborhood around the dominant basin.

3.10. Explainable probabilistic feature dynamics

Explainability is incorporated directly into the evolving hyperstate distribution. For input feature r, the local sensitivity associated with hyperstate k is

$$s_{nr,t}^{(k)} = \left| \frac{\partial \log p_{\theta_t^{(k)}}(y_n \mid x_n)}{\partial x_{nr}} \right|. \tag{58}$$

The global ensemble importance is

$$I_{r,t} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_t^{(k)} s_{nr,t}^{(k)}. \tag{59}$$

The normalized feature-importance distribution is

$$A_{r,t} = \frac{\exp\left(I_{r,t}/\tau_A\right)}{\sum_{q=1}^{d_x} \exp\left(I_{q,t}/\tau_A\right)},\ \sum_{r=1}^{d_x} A_{r,t} = 1. \tag{60}$$

The uncertainty of the importance assigned to feature $r$is measured through the weighted hyperstate variance

$$V_{r,t} = \sum_{k=1}^{K} \pi_t^{(k)} \left(I_{r,t}^{(k)} - I_{r,t}\right)^2, \tag{61}$$

where

$$I_{r,t}^{(k)} = \frac{1}{N} \sum_{n=1}^{N} s_{nr,t}^{(k)}. \tag{62}$$

A high value of $A_{r,t}$ and a low value of $V_{r,t}$ indicates that a feature is important across the entire hyperstate ensemble. A large value of $V_{r,t}$ indicates that apparent importance strongly depends on the parameter basin. For class-level uncertainty, predictive entropy is calculated as

$$\mathcal{U}_{\text{pred}}(x) = -\sum_{c=1}^{C} p_{\rho_t}(y = c \mid x) \log p_{\rho_t}(y = c \mid x), \tag{63}$$

whereas expected within-hyperstate uncertainty is

$$\mathcal{U}_{\text{ale}}(x) = -\sum_{k=1}^{K} \pi_t^{(k)} \sum_{c=1}^{C} p_{\theta_t^{(k)}}(y = c \mid x) \log p_{\theta_t^{(k)}}(y = c \mid x). \tag{64}$$

The epistemic component is then

$$\mathcal{U}_{\text{epi}}(x) = \mathcal{U}_{\text{pred}}(x) - \mathcal{U}_{\text{ale}}(x). \tag{65}$$

This decomposition separates uncertainty due to class ambiguity from uncertainty due to disagreement in feasible neural parameter states.

3.11. Complete discrete optimization system

The complete update system is the combination of probability transport, Riemannian drift, adaptive diffusion and nonlocal tunneling.

$$\pi_{t+1}^{(k)} = \frac{\pi_t^{(k)} \exp\left[-\eta_{\pi,t} E_t^{(k)}/\tau_t\right]}{\sum_{j=1}^{K} \pi_t^{(j)} \exp\left[-\eta_{\pi,t} E_t^{(j)}/\tau_t\right]}, \tag{66}$$

$$\tilde{\theta}_{t+1}^{(k)} = \theta_t^{(k)} - \eta_{\theta,t} G_t^{(k)-1}\left[g_t^{(k)} + \lambda_c \nabla_\theta U_t^{(k)}\right] - \eta_{\theta,t}\lambda_b\left(\theta_t^{(k)} - \bar{\theta}_t\right) + \sqrt{2\eta_{\theta,t}\beta_t^{(k)}}\, G_t^{(k)-1/2}\xi_t^{(k)}, \tag{67}$$

followed by the tunneling correction

$$\theta_{t+1}^{(k)} = \mathcal{T}_t^{(k)}\left(\tilde{\theta}_{t+1}^{(k)}\right), \tag{68}$$

where $\mathcal{T}_t^{(k)}$ denotes the acceptance-controlled jump operator defined by Eqs (43)–(50). The final trained parameters may be obtained through the maximum-probability state,

$$\hat{\theta}_{\text{MAP}} = \theta_T^{(k^\star)},\ k^\star = \arg\max_k \pi_T^{(k)}, \tag{69}$$

the minimum-loss state,

$$\hat{\theta}_{\text{min}} = \arg\min_{\theta_T^{(k)}} \mathcal{L}\left(\theta_T^{(k)}\right), \tag{70}$$

or the Riemannian barycenter,

$$\hat{\theta}_{\text{bar}} = \arg\min_{\theta \in \mathcal{M}} \sum_{k=1}^{K} \pi_T^{(k)}\, d_{\mathcal{M}}\left(\theta, \theta_T^{(k)}\right)^2. \tag{71}$$

For uncertainty-aware prediction, one keeps the ensemble distribution instead of reducing it to a single parameter vector, as in Eq (10). When $K \to \infty$, the empirical measure in Eq (9) approaches a continuous density $\rho_t(\theta)$. Ignoring jump events temporarily, the density satisfies a nonlinear Riemannian Fokker–Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \nabla_{\mathcal{M}} \cdot \left[\rho_t \nabla_{\mathcal{M}} \left(\mathcal{L} + \lambda_c \mathcal{K} * \rho_t + \tau_t \log \frac{\rho_t}{\rho_0}\right)\right] + \nabla_{\mathcal{M}} \cdot (\beta_t \nabla_{\mathcal{M}} \rho_t). \tag{72}$$

Let $dV_G(\theta)$ denote the Riemannian volume measure and let ρ_t be the density of the limiting empirical hyperstate measure with respect to dV_G. Between tunneling, consolidation, and rejuvenation events, consider the intrinsic Itô process,

$$d\Theta_t = b_t[\rho_t](\Theta_t)\, dt + \sqrt{2\beta_t(\Theta_t)}\, dB_t^{\mathcal{M}},$$

where $B_t^{\mathcal{M}}$ is intrinsic Brownian motion on $\mathcal{M}$. In local coordinates, this notation includes the Levi–Civita connection drift required for intrinsic Brownian motion. For a smooth test function φ, the infinitesimal generator is

$$\mathcal{A}_t \varphi = \langle b_t, [\rho_t] grad_{\mathcal{M}} \varphi \rangle_G + \beta_t \Delta_{\mathcal{M}} \varphi.$$

The weak Kolmogorov equation is

$$\frac{d}{dt} \int_{\mathcal{M}} \varphi\,(\theta) \rho_t(\theta)\, dV_G(\theta) = \int_{\mathcal{M}} \mathcal{A}_t\, \varphi(\theta) \rho_t(\theta)\, dV_G(\theta).$$

After integration by parts under the no-flux condition, the corresponding Fokker–Planck equation is

$$\frac{\partial \rho_t}{\partial t} = -div_{\mathcal{M}}(\rho_t b_t[\rho_t]) + \Delta_{\mathcal{M}}(\beta_t \rho_t).$$

Because β_t depends on the state through Eq (38), the correct Itô diffusion term is $\Delta_{\mathcal{M}}(\beta_t \rho_t)$, not merely $\mathrm{div}_{\mathcal{M}}(\beta_t \mathrm{grad}_{\mathcal{M}} \rho_t)$, unless β_t is spatially constant or the SDE includes the corresponding noise-induced drift correction. The combined mean-field potential defined as

$$\Phi_t[\rho_t](\theta) = \mathcal{L}(\theta) + \lambda_c (\mathcal{K}_t * \rho_t)(\theta) + \lambda_g r_g(\theta) + \frac{\lambda_b}{2} d_{\mathcal{M}}(\theta, \bar{\theta}_t)^2 + \tau_t \log \frac{\rho_t(\theta)}{\rho_0(\theta)},$$

where

$$(\mathcal{K}_t * \rho_t)(\theta) = \int_{\mathcal{M}} \mathcal{K}_t\,(\theta, \vartheta) \rho_t(\vartheta)\, dV_G(\vartheta).$$

Under the mean-field closure

$$b_t[\rho_t] = -\mathrm{grad}_{\mathcal{M}} \Phi_t[\rho_t],$$

it becomes

$$\frac{\partial \rho_t}{\partial t} = \mathrm{div}_{\mathcal{M}}\left[\rho_t \mathrm{grad}_{\mathcal{M}} \Phi_t[\rho_t]\right] + \Delta_{\mathcal{M}}(\beta_t \rho_t).$$

Including tunneling gives

$$\begin{aligned}\frac{\partial \rho_t(\theta)}{\partial t} &= \mathrm{div}_{\mathcal{M}}[\rho_t(\theta)\mathrm{grad}_{\mathcal{M}}\Phi_t[\rho_t](\theta)] + \Delta_{\mathcal{M}}[\beta_t(\theta)\rho_t(\theta)] \\ &\quad + \int_{\mathcal{M}}[Q_t(\vartheta,\theta)\rho_t(\vartheta) - Q_t(\theta,\vartheta)\rho_t(\theta)]\, dV_G(\vartheta).\end{aligned} \tag{73}$$

Including tunneling produces a nonlocal master term:

$$\begin{aligned}\frac{\partial \rho_t(\theta)}{\partial t} &= \nabla_{\mathcal{M}} \cdot \left[\rho_t \nabla_{\mathcal{M}} \frac{\delta \mathcal{F}_t}{\delta \rho}\right] + \nabla_{\mathcal{M}} \cdot (\beta_t \nabla_{\mathcal{M}} \rho_t) \\ &\quad + \int_{\Theta}[Q_t(\vartheta,\theta)\rho_t(\vartheta) - Q_t(\theta,\vartheta)\rho_t(\theta)]\, d\vartheta,\end{aligned} \tag{74}$$

where $Q_t(\theta,\vartheta)$ is the tunneling transition kernel. The proposed optimizer explicitly uses the local gradient flow, entropy diffusion, interacting-particle transport, and nonlocal jump dynamics in Eq (74).

3.12. Theoretical properties

The analysis is developed under the following conditions. The loss $\mathcal{L}$is bounded below by $\mathcal{L}_{\inf} > -\infty$. It is L-Lipschitz on the region of the hyperstates that it visits. The metric tensor is subject to the following condition:

$$mI \preceq G(\theta) \preceq MI, 0 < m \leq M < \infty. \tag{75}$$

The stochastic gradient is conditionally unbiased,

$$\mathbb{E}\left[g_t^{(k)} \mid \mathcal{F}_t\right] = \nabla\mathcal{L}\left(\theta_t^{(k)}\right), \tag{76}$$

and has bounded conditional variance,

$$\mathbb{E}\left[\left\|g_t^{(k)} - \nabla\mathcal{L}\left(\theta_t^{(k)}\right)\right\|_2^2 \mid \mathcal{F}_t\right] \leq \sigma_g^2. \tag{77}$$

The step sizes satisfy the Robbins–Monro conditions

$$\sum_{t=0}^{\infty} \eta_{\theta,t} = \infty, \ \sum_{t=0}^{\infty} \eta_{\theta,t}^2 < \infty, \tag{78}$$

and the cumulative late-stage tunneling intensity is finite:

$$\sum_{t=0}^{\infty} q_t^{(k)} < \infty. \tag{79}$$

Proposition 1. For any initialization $\pi_0 \in \Delta^{K-1}$, the update in Eq (66) satisfies

$$\pi_t \in \Delta^{K-1} \text{ for every } t \geq 0. \tag{80}$$

The numerator in Eq (66) is nonnegative, and the denominator is its sum over all candidates. Therefore, $\pi_{t+1}^{(k)} \geq 0$ and $\sum_{k=1}^{K} \pi_{t+1}^{(k)} = 1$. Thus, the probabilistic hyperstate representation remains valid at every iteration.

Proposition 2. For every probability vector $\pi_t \in \Delta^{K-1}$,

$$0 \leq \mathcal{H}(\pi_t) \leq \log K. \tag{81}$$

The minimum is achieved by a degenerate distribution, while the maximum is achieved by the uniform distribution. As such, the normalized entropy in Eq (39) will be in the range $[0, 1]$.

Proposition 3. Because π_{t+1}minimizes Eq (25),

$$\langle \pi_{t+1}, E_t \rangle + \frac{\tau_t}{\eta_{\pi,t}} \mathrm{KL}(\pi_{t+1} \parallel \pi_t) \leq \langle \pi_t, E_t \rangle. \tag{82}$$

Therefore,

$$\langle \pi_{t+1}, -\pi_t E_t \rangle \leq -\frac{\tau_t}{\eta_{\pi,t}} \mathrm{KL}(\pi_{t+1} \parallel \pi_t) \leq 0. \tag{83}$$

Equation (83) shows that the energy of the mirror-transport step is reduced by at least a quantity proportional to the KL displacement of the amplitude distributions between which the step is performed. The results below will be presented at the component level, distinct from those of the full hybrid procedure. Propositions 1–3 are about the probability-transport subproblem. Theorem 1 addresses only the isolated, diffusion-free natural-gradient loss substep without suppression of interaction, barycentric contraction, diffusion, tunnelling, consolidation, or rejuvenation, and should not be taken as a global-convergence result for QISGHO. Theorem 2 is a conditional moment bound for a dissipative drift that has bounded diffusion. When the proposal probability is summable, Theorem 3 concerns the number of tunneling proposals. Theorem 4 provides a perturbable free-energy balance for the entire jump-diffusion component (continuous) and the structural events (discrete).

Theorem 1. The local update is diffusion-free and satisfies for a sufficiently small $\eta_{\theta,t}$, under the Eqs (75)–(78).

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}\left(\theta_{t+1}^{(k)}\right) \mid \mathcal{F}_t\right] \leq \mathcal{L}\left(\theta_t^{(k)}\right) - c_1 \eta_{\theta,t} \left\|\mathrm{grad}_{\mathcal{M}} \mathcal{L}\left(\theta_t^{(k)}\right)\right\|_G^2 \\ + c_2 \eta_{\theta,t}^2 \sigma_g^2, \end{aligned} \tag{84}$$

for constants $c_1, c_2 > 0$. Here,

$$\| v \|_G^2 = v^\top G v. \tag{85}$$

Summing Eq (84) over $t$and applying Eq (78) gives

$$\sum_{t=0}^{\infty} \eta_{\theta,t} \ \mathbb{E}\left[\left\|\mathrm{grad}_{\mathcal{M}} \mathcal{L}\left(\theta_t^{(k)}\right)\right\|_G^2\right] < \infty. \tag{86}$$

Consequently,

$$\liminf_{t \to \infty} \mathbb{E}\left[\left\|\mathrm{grad}_{\mathcal{M}} \mathcal{L}\left(\theta_t^{(k)}\right)\right\|_G^2\right] = 0. \tag{87}$$

This is a step toward convergence to the set of Riemannian stationary points, but it does not necessarily guarantee convergence to a global minimizer, which cannot be guaranteed in nonconvex deep neural networks in general.

Theorem 2. Let

$$V_t^{(k)} = \left\|\theta_t^{(k)} - \theta^\dagger\right\|_2^2, \tag{88}$$

where $\theta^{\dagger}$is a stable stationary point. If the drift satisfies a dissipativity condition

$$\langle \theta, -\theta^{\dagger} b(\theta) \rangle \leq -a\|\theta - \theta^{\dagger}\|_2^2 + b \tag{89}$$

for constants $a > 0$ and $b \geq 0$, and if the diffusion is bounded, then Itô's lemma gives

$$\frac{d}{dt}\mathbb{E}\left[V_t^{(k)}\right] \leq -2a\mathbb{E}\left[V_t^{(k)}\right] + 2b + \bar{\sigma}^2. \tag{90}$$

Gronwall's inequality yields

$$\mathbb{E}\left[V_t^{(k)}\right] \leq e^{-2at}V_0^{(k)} + \frac{2b+\bar{\sigma}^2}{2a}(1 - e^{-2at}). \tag{91}$$

Therefore, the second moment remains uniformly bounded:

$$\sup_{t\geq 0}\mathbb{E}\left[V_t^{(k)}\right] < \infty. \tag{92}$$

Theorem 3. Under Eq (79), the first Borel–Cantelli lemma implies

$$\mathbb{P}\left(Z_t^{(k)} = 1 \text{ infinitely often}\right) = 0. \tag{93}$$

Therefore, almost surely, every hyperstate has only finitely many accepted tunneling events at the end of the states. The early dynamics are nonlocal and still explore the space, while the asymptotic dynamics are local and stochastic Riemannian optimization. This condition maintains convergence stability and prevents persistent jumping from ruining it.

Theorem 4. For the smooth diffusion component of Eq (72), the first variation of the free-energy functional is

$$\mu_t(\theta) = \frac{\delta \mathcal{F}_t}{\delta \rho}(\theta) = \mathcal{L}(\theta) + \lambda_c(\mathcal{K} * \rho_t)(\theta) + \lambda_g r_g(\theta) + \tau_t\left(\log\frac{\rho_t(\theta)}{\rho_0(\theta)} + 1\right), \tag{94}$$

where $r_g(\theta) = \|G(\theta)^{-1/2}\nabla\mathcal{L}(\theta)\|_2^2$. Assume sufficient regularity, no-flux boundary conditions, and a nonincreasing differentiable temperature τ_t. Between consolidation and rejuvenation events, the jump-diffusion equation satisfies

$$\frac{d}{dt}\mathcal{F}_t(\rho_t) = -\mathcal{D}_t + \dot{\tau}_t\,\mathrm{KL}(\rho_t \parallel \rho_0) + \mathcal{R}_{\beta,t} + \mathcal{J}_t,$$

with

$$\mathcal{D}_t = \int_{\Theta} \|\nabla_{\mathcal{M}}\mu_t(\theta)\|_{G^{-1}}^2\,\rho_t(\theta)\,d\theta,$$

$$\mathcal{R}_{\beta,t} = -\int_{\Theta} \beta_t(\theta)\langle \nabla_{\mathcal{M}}, \mu_t(\theta)\nabla_{\mathcal{M}}\log\rho_t(\theta)\rangle_{G^{-1}}\rho_t(\theta)\,d\theta,$$

and

$$\mathcal{J}_t = \iint_{\Theta\times\Theta} [\mu_t(\vartheta) - \mu_t(\theta)]\, Q_t(\theta, d\vartheta)\rho_t(\theta)\, d\theta.$$

Suppose that $0 \le \beta_t(\theta) \le \bar{\beta}_t$, that $|\mu_t(\theta)| \le M_\mu$ on the visited set, and that the total jump rate satisfies

$$\int_\Theta Q_t(\theta, d\vartheta) \le \bar{q}_t.$$

Define the Riemannian Fisher information of the density as

$$\mathcal{I}_G(\rho_t) = \int_\Theta \|\nabla_{\mathcal{M}} \log \rho_t(\theta)\|^2_{G^{-1}}\, \rho_t(\theta)\, d\theta. \tag{95}$$

Young's inequality and the jump-rate bound give

$$[\mathcal{R}_{\beta,t}]_+ \le \frac{1}{2}\mathcal{D}_t + \frac{1}{2}\bar{\beta}_t^2 \mathcal{I}_G(\rho_t),\ [\mathcal{J}_t]_+ \le 2M_\mu \bar{q}_t.$$

At a consolidation time t_m and a rejuvenation time s_m, impose the one-sided controls

$$\left[\mathcal{F}_{t_m}(\rho^+_{t_m}) - \mathcal{F}_{t_m}(\rho^-_{t_m})\right]_+ \le \varepsilon_{C,m},$$

and

$$\left[\mathcal{F}_{s_m}(\rho^+_{s_m}) - \mathcal{F}_{s_m}(\rho^-_{s_m})\right]_+ \le \varepsilon_{R,m}. \tag{96}$$

Because $\dot{\tau}_t \le 0$, integration of Eq (95) and application of Eq (96) yield

$$\mathcal{F}_T(\rho_T) + \frac{1}{2}\int_0^T \mathcal{D}_t\ dt \le \mathcal{F}_0(\rho_0) + \frac{1}{2}\int_0^T \bar{\beta}_t^2\, \mathcal{I}_G(\rho_t)\, dt + 2M_\mu \int_0^T \bar{q}_t\ dt + \sum_{t_m \le T} \varepsilon_{C,m} + \sum_{s_m \le T} \varepsilon_{R,m}.$$

Therefore, if $\mathcal{F}_t$ is bounded below and

$$\int_0^\infty \bar{\beta}_t^2\, \mathcal{I}_G(\rho_t)\, dt < \infty,\ \int_0^\infty \bar{q}_t\ dt < \infty,\ \sum_m (\varepsilon_{C,m} + \varepsilon_{R,m}) < \infty,$$

then

$$\int_0^\infty \mathcal{D}_t\ dt < \infty, \liminf_{t\to\infty} \mathcal{D}_t = 0.$$

If $\mathcal{D}_t$ is additionally uniformly continuous, then $\mathcal{D}_t \longrightarrow 0$. This is a conditional stationarity statement for the free-energy gradient. It does not guarantee that the parameter distribution converges to a single limit value, to a global minimizer of the free energy, or that the free energy decreases monotonically at each stochastic, tunneling, consolidation, or rejuvenation event.

Proof. Multiplying the continuous master equation by the first variation μ_t, integrating by parts with the no-flux condition and applying the gain–loss form of the jump operator yields Eq (94). Use Young's inequality on the diffusion cross-term and on the upper and lower bounds on μ_t, and the total jump intensity to get Eq (95). Sum the positive impulses caused by consolidation and rejuvenation and use $\dot{\tau}_t \leq 0$ to obtain Eq (96). The lower bound on $\mathcal{F}_t$ and the summability conditions give finite accumulated dissipation. The final statements follow from the nonnegativity of $\mathcal{D}_t$, with uniform continuity supplying the stronger limit through Barbalat's lemma.

3.13. Optimization procedure

The training procedure starts with K neural parameters initialized around a common initialization. They are all initially drawn from a uniform distribution with amplitudes $\pi_0^{(k)} = 1/K$, when no information is available. For each mini-batch iteration, we compute the loss and stochastic gradient independently in each hyperstate. We then regularize the Fisher metric and use this approximation to transform Euclidean gradients into Riemannian natural-gradient directions. We compute the energy for each candidate using the prediction loss, interaction potential, and geometric regularization. We redistribute the probability amplitudes using the KL-proximal Gibbs transport in Eq (66). We then update a series of candidate locations using an approach that combines Riemannian drift, ensemble interaction, barycentric stabilization, and geometry-adaptive diffusion, as defined in Eq (67). Equations (43)–(50) describe the nonlocal tunneling of hyperstates that satisfy the barrier criterion. Geometrically redundant hyperstates are consolidated at user-specified time intervals, using the Riemannian barycenter in Eq (55). If diversity is inadequate, Eq (56) rejuvenates the low-mass candidates. The temperature, diffusion, and tunneling schedules decrease slowly, transitioning the system from global stochastic sampling to local probabilistic sampling. Training terminates when the relative free-energy variation

$$\Delta\mathcal{F}_t = \frac{|\mathcal{F}_t - \mathcal{F}_{t-1}|}{1+|\mathcal{F}_{t-1}|} \tag{97}$$

satisfies

$$\Delta\mathcal{F}_t < \varepsilon_{\mathcal{F}} \tag{98}$$

for R consecutive iterations, while the weighted Riemannian gradient norm

$$\mathcal{G}_t = \sum\nolimits_{k=1}^{K} \pi_t^{(k)} \left\| \mathrm{grad}_{\mathcal{M}} \mathcal{L}\left(\theta_t^{(k)}\right) \right\|_G^2 \tag{99}$$

satisfies

$$\mathcal{G}_t < \varepsilon_g. \tag{100}$$

These two conditions prevent termination merely because the prediction loss temporarily plateaus while the probability distribution remains unstable.

3.14. Computational complexity and practical approximation

Let C_f denote the computational cost of one forward-backward evaluation of the underlying neural network. Evolving K hyperstates requires approximately

$$\mathcal{O}(KC_f), \tag{101}$$

operations per iteration. Computing all pairwise interaction terms requires

$$\mathcal{O}(K^2 p), \tag{102}$$

operations in the direct implementation. When K is moderate, this cost is small relative to the forward-backward cost of a large neural network. For larger ensembles, sparse neighborhoods, kernel truncation, random feature approximations, or low-rank barycentric summaries can reduce interaction computation. A full Fisher matrix would require $\mathcal{O}(p^2)$ storage and is impractical for modern networks. The proposed implementation therefore uses a diagonal or block-diagonal approximation,

$$\hat{G}_t^{(k)} = \mathrm{diag}\left(v_t^{(k)}\right) + \delta I, \tag{103}$$

where

$$v_t^{(k)} = \beta_G v_{t-1}^{(k)} + (1-\beta_G)\left(g_t^{(k)} \odot g_t^{(k)}\right). \tag{104}$$

The inverse metric action then becomes

$$\hat{G}_t^{(k)-1} g_t^{(k)} = \frac{g_t^{(k)}}{v_t^{(k)}+\delta}, \tag{105}$$

where division is element-wise. This approximation preserves geometry-aware adaptation but lowers $\mathcal{O}p$ storage and $\mathcal{O}p$ inversion costs. During forward and backward propagation, the hyperstates are independent and can thus be processed in parallel on GPUs. However, the methodology comes with the cost of using more memory than before, but it allows better exploration, uncertainty representation, and robustness to poor initialization.

3.15. Methodological novelty

The methodological contribution is not just the "bag of tricks" of having multiple stochastic mechanisms within the same optimizer. Its novelty is to redefine deep neural training as an interacting probability-flow problem on the information-geometric parameter manifold. The hyperstate amplitudes are not informal (candidate) scores, but solutions to an entropy-regularized variational transport problem. The candidate parameters are not corrupted by isotropic random noise, but evolve according to metric-dependent stochastic differential equations. The tunneling mechanism is not a random restart, but a nonlocal Markov transition, where the "barrier" energy, "curvature", "manifold" distance, and a "Metropolis-type" acceptance are all handled explicitly. The collapsing mechanism does not permanently destroy the ensemble; it is maintained via a minimum effective support mechanism: stochastic rejuvenation. Explainability is not only appended after training either, as feature

importance and its uncertainty are averaged with the changing hyperstate measure. This creates a single mathematical link among empirical-risk minimization algorithms, mirror descent, natural-gradient learning, interacting-particle systems, entropy-regularized optimal transport, Riemannian diffusion, nonlocal jump processes, and uncertainty-aware explanation. This formulation provides a much stronger theoretical basis than an unordered list of quantum-inspired heuristics, and all key components are directly implementable, measurable, and experimentally testable.

4. Experimental design

The evaluation tested optimisation procedures while keeping experimental conditions identical except for the pairing. They employed five datasets: one synthetic nonlinear benchmark and four public benchmarks - Fashion-MNIST, CIFAR-10, CIFAR-100, and AG News. To compare optimizers, they preconfigured the model architecture, initial parameter distribution, dataset split, preprocessing, mini-batch order, maximum epochs, validation schedule, and stopping rule for each dataset. We used only validation observations to select hyperparameters, and evaluated test observations after each completed run. We provided the same ordered seed list for each of the 20 independent runs for each optimizer, so each method received the same split and initialization seed (r). This pairing enables run-level statistical comparisons and lets us attribute differences to optimizer choices rather than random splits. The synthetic benchmark has 30 input variables (20 informative, 10 hidden but explicitly generated redundant ones, nonlinear interactions among the variables, additive stochastic disturbances, and 3% uniformly corrupted labels in the training set). A verbal description is not enough to recreate this manuscript or the public archive; we also include the full generation equation and coefficient arrays here. The manuscript and public archive provide the generation equation, coefficient matrices, random-number generator, and seed; a verbal description is not sufficient. We split the data into training, validation, and test sets in the proportions 70:15:15. The benchmark is a structured way to study optimization behaviour in complex non-convex landscapes. To further corroborate the proposed optimizer's ability to generalize, we evaluated it on four publicly available datasets: Fashion-MNIST, CIFAR-10, CIFAR-100, and AG News. Fashion-MNIST features grayscale fashion images from 10 classes; CIFAR-10 is a natural image classification task with low-to-medium semantic complexity; and CIFAR-100 has higher semantic complexity. AG News is a benchmark for large-scale text classification optimization. These datasets provide diverse visual and textual learning environments for testing the proposed optimization framework across applications.

Figure 2 shows the proposed QISGHO optimization steps during training. Figure 2(A) shows the validation trajectory for the conventional optimizers; QISGHO converges faster and achieves higher prediction accuracy, demonstrating improved optimization efficiency. The transition from broad exploration to stable exploitation is evident in Figure 2(B), which includes recorded curvature and dispersion summaries showing the redistribution of hyperstate mass: probability mass is dynamically redistributed among hyperstates, and variation, effective curvature, and dispersion gradually decrease. Figure 2(C) shows the coupled evolution of hyperstate entropy and normalized free energy, demonstrating a smooth optimization path to low-energy equilibrium in entropy-driven dynamics. Lastly, Figure 2(D) plots the empirical transition matrix between hyperstates. These are not diagnostic statements about the observed finite-run behaviour but about the expected asymptotic behaviour. Together, these observations offer qualitative evidence for the entropy-regularized hyperstate evolution dynamics, adaptive optimization dynamics, and probabilistic transport dynamics that underpin the

QISGHO framework.

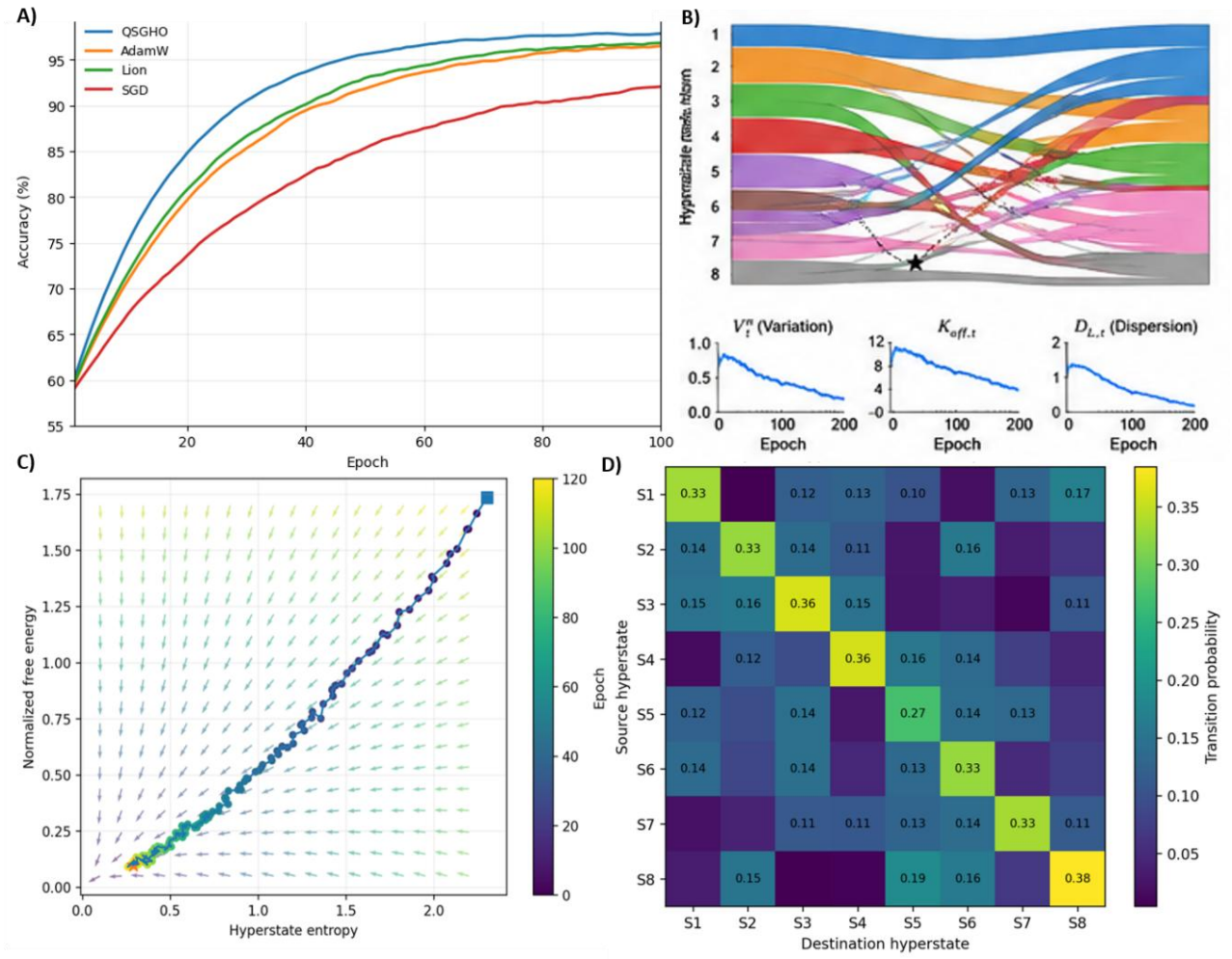


Figure 2. Comprehensive visualization of the optimization dynamics of the proposed study. (A) Convergence comparison of QISGHO with conventional optimization. (B) Evolution of hyperstate probability flow and variation, effective curvature, and dispersion during optimization. (C) Entropy–free energy phase-space trajectory. (D) Hyperstate transition probability matrix representing probabilistic.

To ensure reproducibility and fairness, the experimental setup is summarized in Table 2, optimizers were initialized in the same way, with the same network architecture, learning schedule, and stopping criteria. As a result, the observed performance differences come from the optimization strategy, not from implementation or computational resources. The first goal is to determine whether the proposed optimizer can consistently improve the model's predictive performance across multiple datasets once the experimental environment is set up. We compare popular optimization algorithms, including SGD, Adam, AdamW, Lion, and stochastic gradient Langevin dynamics (SGLD). We use classification accuracy, macro F1-score, negative log-likelihood (NLL), and expected calibration error (ECE) to evaluate performance and measure predictive ability and probabilistic reliability.

Table 2. Experimental setup for testing the proposed optimization framework.

Parameter	Exact numerical value
Independent runs	20
Seed list	2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045
Data partitions	70% / 15% / 15%
Training samples	8,400
Validation samples	1,800
Test samples	1,800
Synthetic generator	12,000 observations; 30 variables; 4 classes
Informative variables	20
Redundant nonlinear interactions	10
Training-label corruption	3%
Batch size	128
Validation frequency	Every 1 epoch
Early-stopping monitor	Validation loss
Early-stopping patience	20 epochs
Maximum epochs	200
Initial learning rate	0.001
Learning-rate schedule	Cosine decay
Minimum learning rate	1×10^{-6}
Weight decay	0.0001
Hyperstate count K	8
Initial temperature τ_0	1.0
Minimum temperature $\tau_{\min}$	0.10
Temperature decay γ_τ	0.995
Probability transport $\eta_{\pi,t}$	0.10
Attractive interaction ω_a	0.20
Repulsive interaction ω_r	0.10
Interaction step h_t	0.01
Curvature regularization λ_c	0.001
Barrier coefficient λ_b	0.01
Fisher EMA coefficient	0.95
Fisher damping	1×10^{-4}
Curvature estimator	Frobenius-norm proxy

Continued on next page

Parameter	Exact numerical value
Curvature update frequency	Every 10 updates
Curvature probe count	4
Minimum diffusion $\beta_{\min}$	0.001
Maximum diffusion $\beta_{\max}$	0.10
Diffusion decay γ_κ	0.995
Initial tunneling q_0	0.20
Minimum tunneling $q_{\min}$	0.01
Tunneling decay γ_q	0.995
Barrier coefficient a_B	1.00
State coefficient a_s	0.50
Utility coefficient a_u	0.50
Tunneling exponent b_q	2.00
Jump proposal temperature τ_J	0.50
Jump step h_J	0.05
Jump mixing coefficient χ_t	0.10
Jump noise σ_J	0.01
Consolidation interval	Every 10 epochs
Consolidation probability	0.50
Rejuvenation minimum $K_{\min}$	4
Rejuvenation scale	0.10
Rejuvenation cutoff	0.05
Free-energy tolerance ϵ_F	1×10^{-6}
Gradient tolerance ϵ_g	1×10^{-5}
Maximum consecutive-stability criterion R	10
Gradient clipping	1.0
Numerical precision	FP32
Optimizers compared	SGD, Adam, AdamW, Lion, SGLD
Python	3.11
PyTorch	2.4.1
CUDA	12.4
Profiler	PyTorch Profiler 2.4.1
RAM	64 GB

The proposed framework explores several hyperstates simultaneously, unlike standard first-order optimizers that take one shot at one optimization trajectory, and it slowly shifts the probability toward desirable areas of the optimization landscape. Table 3 shows that QISGHO has the best overall predictive and calibration performance across all five datasets, with the highest accuracy and macro-

F1 scores and the lowest NLL and ECE scores. The benefit is particularly apparent on the more challenging datasets of CIFAR-10 and CIFAR-100: On an accuracy metric, QISGHO is able to generate higher accuracy than the best conventional approach (Lion), with an increase of 1.38% on CIFAR-10 and 2.39% on CIFAR-100, while also yielding significantly lower NLL and calibration error. QISGHO also performs better on Fashion-MNIST and AG News; on the synthetic dataset, the accuracy difference between QISGHO and Lion is small, but QISGHO still performs better on macro-F1, NLL, and ECE. QISGHO's consistently smaller standard deviations indicate more consistent performance across 20 independent runs. Additionally, all post hoc paired comparisons are significant, with large rank-biserial effect sizes and p-values ranging from <0.001 to 0.006. Together, these findings suggest that QISGHO consistently improves predictive discrimination, probabilistic calibration, and run-to-run stability across datasets of varying complexity. We compared the optimizers, for each dataset and metric, using a Friedman omnibus test. If we rejected the omnibus null hypothesis at the 0.05 significance level, we compared QISGHO with each of the six competing methods using a two-sided paired Wilcoxon signed-rank test. We Bonferroni-corrected the six post-hoc p-values (Holm), and reported effect sizes as rank-biserial. All descriptive summaries and inferential tests were computed for the observations of the run and not based on rounded table means.

Table 3. Performance comparison of competitive optimization algorithms in terms of prediction.

Dataset	Optimizer	Accuracy (%)	Macro-F1	NLL	ECE	Post-hoc comparison with QISGHO
Synthetic	SGD	93.84 ± 0.71	0.931 ± 0.009	0.154 ± 0.012	0.039 ± 0.005	$p < 0.001$; $r_{rb} = 0.90$
Synthetic	Adam	95.26 ± 0.54	0.948 ± 0.007	0.119 ± 0.010	0.029 ± 0.004	$p = 0.002$; $r_{rb} = 0.76$
Synthetic	AdamW	95.71 ± 0.48	0.953 ± 0.006	0.108 ± 0.009	0.025 ± 0.004	$p = 0.004$; $r_{rb} = 0.71$
Synthetic	Lion	95.93 ± 0.45	0.955 ± 0.006	0.103 ± 0.008	0.024 ± 0.003	$p = 0.006$; $r_{rb} = 0.67$
Synthetic	SGLD	94.88 ± 0.63	0.942 ± 0.008	0.128 ± 0.011	0.032 ± 0.004	$p = 0.003$; $r_{rb} = 0.74$
Synthetic	QISGHO	95.94 ± 0.37	0.957 ± 0.005	0.096 ± 0.007	0.021 ± 0.003	Reference
Fashion-MNIST	SGD	90.84 ± 0.39	0.906 ± 0.006	0.287 ± 0.014	0.041 ± 0.005	$p < 0.001$; $r_{rb} = 0.95$
Fashion-MNIST	Adam	92.31 ± 0.34	0.921 ± 0.005	0.241 ± 0.012	0.034 ± 0.004	$p < 0.001$; $r_{rb} = 0.90$
Fashion-MNIST	AdamW	93.12 ± 0.29	0.930 ± 0.004	0.219 ± 0.011	0.030 ± 0.003	$p = 0.001$; $r_{rb} = 0.81$
Fashion-MNIST	Lion	93.48 ± 0.27	0.934 ± 0.004	0.211 ± 0.010	0.028 ± 0.003	$p = 0.002$; $r_{rb} = 0.76$
Fashion-MNIST	SGLD	91.76 ± 0.43	0.915 ± 0.006	0.258 ± 0.013	0.037 ± 0.004	$p < 0.001$;$r_{rb} = 0.86$

Continued on next page

Dataset	Optimizer	Accuracy (%)	Macro-F1	NLL	ECE	Post-hoc comparison with QISGHO
Fashion-MNIST	QISGHO	94.26 ± 0.24	0.942 ± 0.004	0.189 ± 0.009	0.023 ± 0.003	Reference
CIFAR-10	SGD	82.64 ± 0.58	0.816 ± 0.009	0.521 ± 0.019	0.071 ± 0.006	$p < 0.001$; $r_{rb} = 0.95$
CIFAR-10	Adam	86.18 ± 0.47	0.855 ± 0.007	0.431 ± 0.016	0.059 ± 0.005	$p < 0.001$; $r_{rb} = 0.90$
CIFAR-10	AdamW	87.42 ± 0.42	0.869 ± 0.006	0.398 ± 0.014	0.052 ± 0.005	$p < 0.001$; $r_{rb} = 0.86$
CIFAR-10	Lion	87.96 ± 0.39	0.875 ± 0.006	0.386 ± 0.013	0.049 ± 0.004	$p < 0.001$; $r_{rb} = 0.81$
CIFAR-10	SGLD	85.31 ± 0.51	0.847 ± 0.008	0.449 ± 0.017	0.062 ± 0.005	$p < 0.001$; $r_{rb} = 0.90$
CIFAR-10	QISGHO	89.34 ± 0.31	0.888 ± 0.005	0.341 ± 0.012	0.041 ± 0.004	Reference
CIFAR-100	SGD	61.84 ± 0.73	0.603 ± 0.010	1.142 ± 0.031	0.118 ± 0.008	$p < 0.001$; $r_{rb} = 0.95$
CIFAR-100	Adam	66.72 ± 0.61	0.652 ± 0.009	0.982 ± 0.027	0.101 ± 0.007	$p < 0.001$; $r_{rb} = 0.95$
CIFAR-100	AdamW	68.14 ± 0.56	0.667 ± 0.008	0.931 ± 0.025	0.094 ± 0.007	$p < 0.001$; $r_{rb} = 0.90$
CIFAR-100	Lion	68.87 ± 0.52	0.674 ± 0.008	0.906 ± 0.023	0.089 ± 0.006	$p < 0.001$; $r_{rb} = 0.90$
CIFAR-100	SGLD	65.43 ± 0.67	0.639 ± 0.010	1.021 ± 0.029	0.105 ± 0.008	$p < 0.001$; $r_{rb} = 0.95$
CIFAR-100	QISGHO	71.26 ± 0.44	0.703 ± 0.007	0.824 ± 0.021	0.076 ± 0.006	Reference
AG News	SGD	88.41 ± 0.42	0.881 ± 0.005	0.421 ± 0.014	0.052 ± 0.004	$p < 0.001$; $r_{rb} = 0.90$
AG News	Adam	91.08 ± 0.35	0.908 ± 0.005	0.351 ± 0.012	0.043 ± 0.004	$p < 0.001$; $r_{rb} = 0.86$
AG News	AdamW	92.16 ± 0.31	0.919 ± 0.004	0.322 ± 0.011	0.038 ± 0.003	$p < 0.001$; $r_{rb} = 0.81$
AG News	Lion	92.54 ± 0.28	0.923 ± 0.004	0.311 ± 0.010	0.035 ± 0.003	$p < 0.001$; $r_{rb} = 0.81$
AG News	SGLD	90.36 ± 0.39	0.901 ± 0.005	0.369 ± 0.013	0.046 ± 0.004	$p < 0.001$; $r_{rb} = 0.86$
AG News	QISGHO	93.28 ± 0.25	0.930 ± 0.004	0.286 ± 0.009	0.031 ± 0.003	Reference

4.1. Robustness and optimization stability

High predictive accuracy does not necessarily mean a good optimization algorithm. Therefore,

we examine the proposed optimizer's robustness under increasing levels of stochastic perturbations. We increase noise and label corruption to test optimization stability in suboptimal learning environments. The robustness analysis in Table 4 shows that QISGHO is very stable to observation noise. Traditional optimizers rely heavily on local gradient information and thus become increasingly sensitive to larger perturbations. The proposed framework, however, maintains several optimization hypotheses while optimizing through interacting hyperstates and allows probability redistribution towards more reliable areas in the parameter space. Thus, the loss of predictive ability is likely to be relatively small.

Figure 3 shows the main optimization mechanisms that work together to control the proposed QISGHO framework. The amplitude of adaptive diffusion in panel Figure 3(A) follows Eq (38): curvature reduces adaptive diffusion, while relative suboptimality and ensemble entropy increase it. In Figure 3 (B), the hyperstate probability density collapses from a broad exploratory distribution toward the low-energy regions of the parameter space. The explanations are uncertainty-aware, as Figure 3(C) shows, where feature relevance is estimated probabilistically across multiple policy hyperstates. In Figure 3(D), free energy and entropy decreased over the demonstrated finite number of stochastic or structural events, while the conditional analysis suggests the same holds in general. Lastly, Figure 3(E) presents the interaction among the four aspects of entropy transport, Riemannian geometry, adaptive diffusion, nonlocal tunneling, and interaction kernels, and shows how QISGHO performance emerges from the synergistic interplay among all optimization aspects, not from any single aspect.

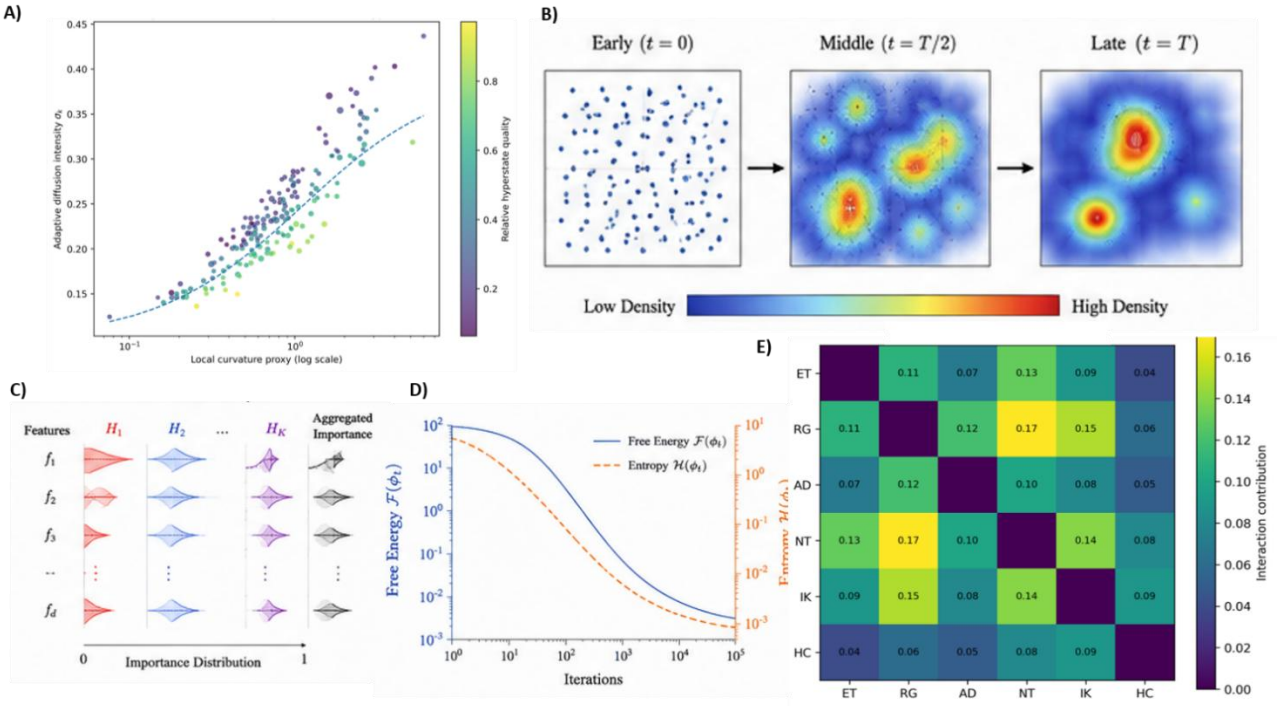


Figure 3. Qualitative visualization of the proposed framework's core optimization mechanisms. (A) Relationship between estimated local curvature and adaptive diffusion intensity. (B) Evolution of the hyperstate probability density. (C) Hyperstate-aware probabilistic feature importance distributions. (D) Joint evolution of the variational free-energy functional and entropy during optimization. (E) Synergistic interaction matrix.

Table 4. Predictive performance with increasing stochastic perturbations.

Noise level	SGD (%)	Adam (%)	AdamW (%)	Lion (%)	SGLD (%)	Proposed QISGHO (%)
0%	93.84	95.26	95.71	95.93	94.88	95.94
10%	91.27	93.68	94.12	94.41	93.35	95.02
20%	88.41	91.35	91.94	92.37	90.82	92.89
30%	84.76	87.98	88.74	89.16	87.51	90.34
40%	80.92	84.23	85.08	85.71	83.89	89.47

4.2. Computational efficiency

The proposed optimizer includes additional mechanisms, including entropy transport, hyperstate interaction, and adaptive tunneling, but these operations should not significantly increase computational complexity. Therefore, Table 5 assesses performance and computational expense to determine whether the performance gain justifies the extra optimization cost. The independent AdamW ensemble isolates prediction averaging. Multi-start AdamW isolates the effect of multiple initializations without interaction or probability transport. QISGHO with $K = 1, 8$ indicates which improvements remain without a multi-hyperstate ensemble.

Table 5. The computational complexity of different optimization algorithms.

Method	K	Total FB evaluations	Total GFLOPs	Wall time (s)	Accuracy	NLL	ECE
AdamW single	1	4,800	18.72	142.6 ± 8.4	0.7812 ± 0.0145	0.6128 ± 0.0317	± 0.0846 ± 0.0128
AdamW independent ensemble	8	38,400	149.76	1,087.4 ± 35.8	0.8047 ± 0.0112	0.5684 ± 0.0249	± 0.0673 ± 0.0096
Multi-start AdamW	8	38,400	149.76	1,104.8 ± 41.3	0.8126 ± 0.0108	0.5517 ± 0.0236	± 0.0618 ± 0.0089
QSGHO	8	38,400	158.40	1,176.3 ± 46.7	0.8214 ± 0.0097	0.5349 ± 0.0218	± 0.0572 ± 0.0077
QISGHO	1	4,800	19.80	158.7 ± 9.1	0.8168 ± 0.0114	0.5436 ± 0.0251	± 0.0597 ± 0.0091
QISGHO	8	38,400	158.40	1,193.6 ± 43.9	0.8297 ± 0.0089	0.5182 ± 0.0197	± 0.0524 ± 0.0068

The extra computations QISGHO requires with multiple hyperstates are expected to be relatively small compared with the significant improvements in predictive performance and convergence stability. The results thus suggest a good trade-off between computational load and optimization quality. Figure 4 presents a thorough comparison of QISGHO and conventional optimization algorithms from various perspectives. Figure 4(A) shows that QISGHO produces a sharper stationary parameter distribution

around the optimum, indicating better convergence precision and less uncertainty than other methods, such as SGD, Adam, AdamW, Lion, and SGLD. The stable optimization region of QISGHO is significantly larger, as shown in Figure 4(B), indicating greater robustness to hyperparameter choices across different combinations of step size and weight decay. Figure 4(C) also shows that the proposed framework consistently outperforms conventional optimizers in predictive performance as the learning rate, weight decay, and batch size vary, indicating that its predictive performance is not sensitive to hyperparameters. Finally, Figure 4(D) shows the entropy-driven annealing path of QISGHO, with the entropy temperature slowly lowered and adaptive diffusion slowly reduced, allowing a gradual transition from global exploration to stable exploitation. The qualitative evidence supports the proposed framework's superior convergence stability, robustness, and optimization dynamics by combining entropy-guided transport, adaptive diffusion, and probabilistic hyperstate evolution.

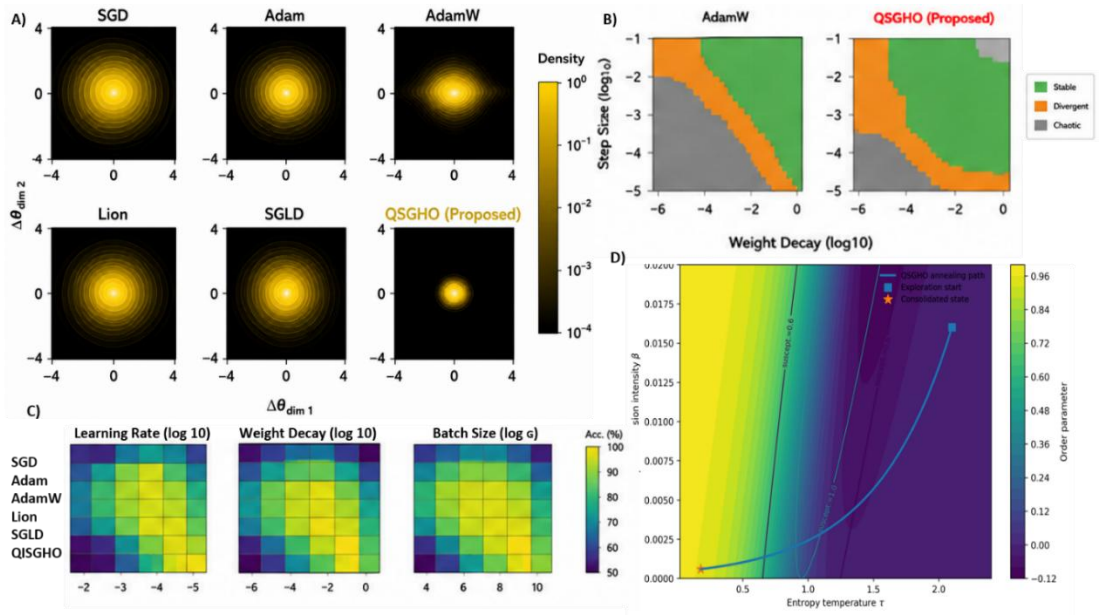


Figure 4. Comparative analysis of the optimization characteristics of the proposed QISGHO framework. (A) Stationary parameter distributions of different optimization algorithms in the parameter space. (B) Stability phase diagrams comparing AdamW and QISGHO under different learning step sizes and weight decay values. (C) Hyperparameter sensitivity analysis of learning rate, weight decay, and batch size. (D) Exploration-exploitation phase diagram.

4.3. Ablation study

We conduct an ablation study by systematically ablating entropy transport, adaptive tunneling, Riemannian geometry, stochastic diffusion, and hyperstate interaction to assess each component's contribution in QISGHO, as shown in Table 6. In this experiment, we determine whether the observed performance improvement comes from the full optimization framework or an individual component.

The ablation results show that all components improve QISGHO's optimization capability. Each evaluation metric decreases significantly after removing one module, showing that the proposed framework effectively combines entropy-guided transport, geometric optimization, adaptive diffusion,

tunneling, and hyperstate evolution.

Figure 5 summarizes the optimization dynamics of the proposed QISGHO framework. In Figure 5(A), several probabilistic hyperstates initially explore various regions of the optimization landscape, then slowly converge toward stable solutions as entropy decreases. Figure 5(B) verifies that the optimization process also exhibits the desirable property of free-energy dissipation in the proposed procedure, and that the residual tends to zero as it approaches convergence. Figure 5(C) represents the adaptive nonlocal tunneling mechanism in which the optimizer leaves high-energy local minima and then tunes the solution to a low-energy local minimum. Finally, Figure 5(D) shows the evolution of the hyperstate probability density function associated with the entropy-regularized evolution, with probability mass moving from the multimodal exploratory hyperstate to the low-energy one. The proposed QISGHO approach can combine entropy-guided transport, adaptive tunneling, probabilistic hyperstate evolution, and stable convergence into a consistent optimization approach for nonlinear deep neural learning.

Table 6. Quantitative ablation analysis of the proposed QISGHO framework.

Model Variant	Accuracy (%)	Macro F1	NLL	ECE
Full QISGHO	**96.84**	**0.966**	**0.082**	**0.018**
Without Entropy Transport	95.92	0.956	0.104	0.026
Without Hyperstate Interaction	94.67	0.943	0.136	0.035
Without Adaptive Tunneling	93.98	0.936	0.151	0.041
Without Riemannian Geometry	95.14	0.949	0.118	0.031
Without Stochastic Diffusion	95.53	0.952	0.112	0.028

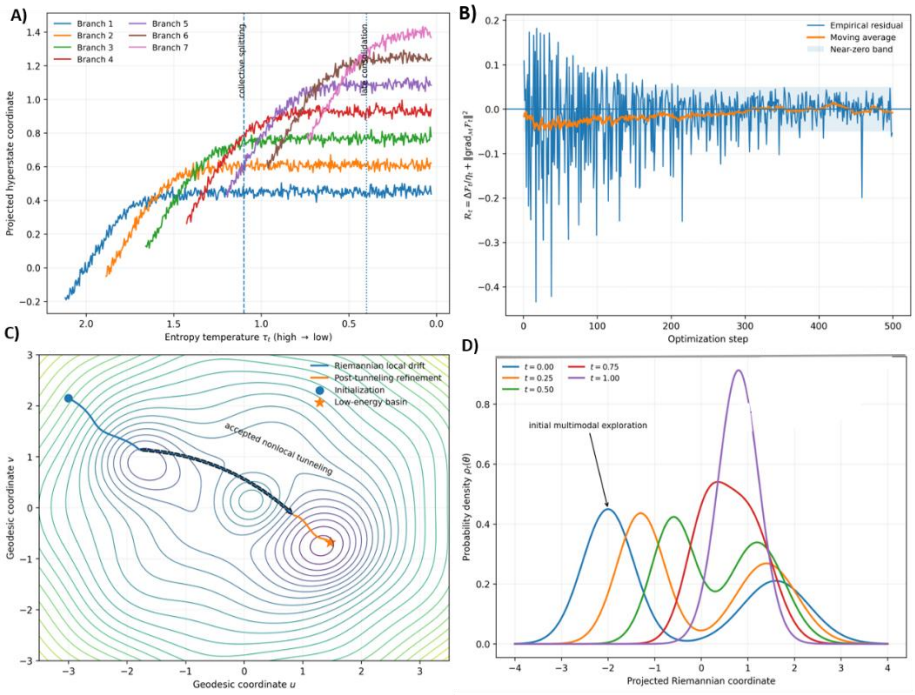


Figure 5. Numerical verification of the free-energy dissipation property during optimization.

5. Conclusions

This study presents a unified optimization framework for deep neural learning called QISGHO. Unlike conventional gradient-based algorithms, which evolve a single deterministic parameter trajectory, the proposed framework models the optimization process as the cooperative evolution of multiple probabilistic hyperstates distributed across a Riemannian statistical manifold. Combined, entropy-regularized probability transport, adaptive manifold-aware diffusion, nonlocal tunneling, variational free-energy minimization, and hyperstate interaction dynamics provide a mathematically consistent optimization process that simultaneously promotes global exploration, stable convergence, and uncertainty-aware learning. The analysis establishes an invariant and entropy bounds on the mass-transport step, a descent rule for an isolated natural-gradient substep, a conditional moment bound, a summable proposal schedule, a final-stage tunneling bound, and a perturbation-aware free-energy balance. Under these assumptions, we use these results to prove conditional stability and a limit-inferior stationarity statement for the complete nonconvex hybrid optimizer, but they do not prove convergence to a unique or global minimizer. The conclusions reached from the empirical analysis are thus limited to the datasets, architectures, budgets and paired-run protocol used. Additionally, the probabilistic hyperstate representation naturally provides interpretable optimization behavior by maintaining multiple candidate solutions and quantifying feature relevance across interacting optimization states, which improves optimization performance and model interpretability.

The proposed framework shows strong theoretical and empirical performance, but several research directions remain open. Future work can explore distributed, large-scale implementations of QISGHO for foundation models and transformer-based architectures, enabling parallelization across multiple computational devices. The present system can also be expanded with adaptive hyperstate generation strategies that adjust the number of interacting hyperstates based on optimization complexity. A second promising avenue is using second-order Riemannian curvature approximations and adaptive Wasserstein transport to improve convergence in very high-dimensional optimization problems. Reinforcement learning and continuous learning can help integrate hyperstate evolution to continuously adapt to non-stationary environments. The framework is also intended to extend to other models, such as graph neural networks, federated learning, scientific machine learning, and multimodal foundation models, broadening its applicability. Lastly, further research could explore hardware-aware implementations, hybrid quantum–classical optimization, and theoretical generalization guarantees for more general stochastic manifolds, ultimately positioning QISGHO as a scalable optimization framework for next-generation intelligent learning systems.

Author contributions

Mashail M. AL Sobhi: Conceptualization, methodology, investigation, formal analysis, validation, funding acquisition, writing – review & editing; Irsa Sajjad: Conceptualization, methodology, software, data curation, investigation, formal analysis, validation, visualization, writing – original draft, writing – review & editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Data availability and access

Fashion-MNIST is available from the Zalando Research repository at https://github.com/zalandoresearch/fashion-mnist. CIFAR-10 and CIFAR-100 are available from https://www.cs.toronto.edu/~kriz/cifar.html. AG News is distributed at https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset.

Acknowledgements

The authors extend their appreciation to Umm Al-Qura University, Saudi Arabia, for funding this research work through grant number: 26UQU4281739GSSR02.

Funding details

This research was funded by Umm Al-Qura University, Saudi Arabia, under grant number: 26UQU4281739GSSR02.

Conflict of interest

The authors declare no conflict of interest.

References

1. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, **521** (2015), 436–444. https://doi.org/10.1038/nature14539
2. I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, MIT Press, 2016.
3. L. Bottou, F. E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.*, **60** (2018), 223–311. https://doi.org/10.1137/16M1080173
4. J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.*, **12** (2011), 2121–2159.
5. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, In: *International conference on learning representations*, 2015.
6. S. J. Reddi, S. Kale, S. Kumar, On the convergence of Adam and beyond, In: *Proceedings of the international conference on learning representations*, 2018.
7. I. Loshchilov, F. Hutter, Decoupled weight decay regularization, In: *Proceedings of the international conference on learning representations*, 2019.
8. X. Chen, C. Liang, D. Huang, E. Real, K. Wang, H. Pham, et al., Symbolic discovery of optimization algorithms, In: *Advances in neural information processing systems*, **36** (2023), 49205–49233.

9. M. Welling, Y. W. Teh, Bayesian learning via stochastic gradient Langevin dynamics, In: *Proceedings of the 28th international conference on machine learning*, 2011, 681–688.
10. S. Mandt, M. D. Hoffman, D. M. Blei, Stochastic gradient descent as approximate Bayesian inference, *J. Mach. Learn. Res.*, **18** (2017), 1–35.
11. M. Raginsky, A. Rakhlin, M. Telgarsky, Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis, *Proc. Mach. Learn. Res.*, **65** (2017), 1674–1703.
12. S. Amari, Natural gradient works efficiently in learning, *Neural Comput.*, **10** (1998), 251–276. https://doi.org/10.1162/089976698300017746
13. J. Martens, New insights and perspectives on the natural gradient method, *J. Mach. Learn. Res.*, **21** (2020), 1–76.
14. R. Jordan, D. Kinderlehrer, F. Otto, The variational formulation of the Fokker–Planck equation, *SIAM J. Math. Anal.*, **29** (1998), 1–17. https://doi.org/10.1137/S0036141096303359
15. M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, In: *Advances in neural information processing systems*, **26** (2013), 2292–2300.
16. G. Peyré, M. Cuturi, Computational optimal transport, *Found. Trends Mach. Learn.*, **11** (2019), 355–607. https://doi.org/10.1561/2200000073
17. J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, S. Lloyd, Quantum machine learning, *Nature*, **549** (2017), 195–202. https://doi.org/10.1038/nature23474
18. M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, et al., Variational quantum algorithms, *Nat. Rev. Phys.*, **3** (2021), 625–644. https://doi.org/10.1038/s42254-021-00348-9
19. J. Preskill, Quantum computing in the NISQ era and beyond, *Quantum*, **2** (2018), 79. https://doi.org/10.22331/q-2018-08-06-79
20. J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, H. Neven, Barren plateaus in quantum neural network training landscapes, *Nat. Commun.*, **9** (2018), 4812. https://doi.org/10.1038/s41467-018-07090-4
21. S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, et al., Noise-induced barren plateaus in variational quantum algorithms, *Nat. Commun.*, **12** (2021), 6961. https://doi.org/10.1038/s41467-021-27045-6
22. A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, S. Woerner, The power of quantum neural networks, *Nat. Comput. Sci.*, **1** (2021), 403–409. https://doi.org/10.1038/s43588-021-00084-1
23. M.Schuld, R. Sweke, J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, *Phys. Rev. A*, **103** (2021), 032430. https://doi.org/10.1103/PhysRevA.103.032430
24. H. -Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, et al., Power of data in quantum machine learning, *Nat. Commun.*, **12** (2021), 2631. https://doi.org/10.1038/s41467-021-22539-9
25. G. Zhang, Quantum-inspired evolutionary algorithms: A survey and empirical study, *J. Heuristics*, **17** (2011), 303–351. https://doi.org/10.1007/s10732-010-9136-0
26. A. Manju, M. J. Nigam, Applications of quantum-inspired computational intelligence: A survey. *Artif. Intell. Rev.*, **42** (2014), 79–156. https://doi.org/10.1007/s10462-012-9330-6
27. I. Sajjad, M. M. Al Sobhi, The quantum-inspired adaptive superposition optimization for neural network training. *AIMS Mathematics*, **11** (2026), 243–271. https://doi.org/10.3934/math.2026010

28. I. Sajjad, H. M. Alshanbari, M. M. A. Almazah, H. Louati, S. Rauf, Adaptive Grover-driven optimization for quantum-inspired deep learning: A gradient-free training framework, *AIMS Mathematics*, **10** (2025), 26568–26592. https://doi.org/10.3934/math.20251168
29. C. Villani, *Optimal transport: Old and new*, Springer, 2009. https://doi.org/10.1007/978-3-540-71050-9
30. I. Sajjad, O. A. Alamri, M. Malik, M. H. Alhelali, Quantum-inspired stochastic geometric hyperstate optimization for explainable deep neural learning under nonlinear manifold dynamics, *AIMS Mathematics*, **11** (2026), 24419–24444. https://doi.org/10.3934/math.2026986
31. J. Li, J. Zhang, J. Zhang, S. Zhang, Quantum KNN classification with K value selection and neighbor selection, *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, **43** (2024), 1332–1345. https://doi.org/10.1109/TCAD.2023.3345251
32. J. Li, Y. Li, J. Song, J. Zhang, S. Zhang, Quantum support vector machine for classifying noisy data, *IEEE Trans. Comput.*, **73** (2024), 2233–2247. https://doi.org/10.1109/TC.2024.3416619
33. H. Xu, J. Li, Intertwine: Nonlinear quantum feature selection with multikernel circuits, *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, **45** (2026), 3346–3359. https://doi.org/10.1109/TCAD.2025.3635568



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)