FROM OPTIMIZATION TO SAMPLING: MUON-INSPIRED LANGEVIN DYNAMICS FOR BAYESIAN LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix-aware optimizers such as MUON exploit the spectral geometry of neural-network weights and have demonstrated great success in LLM training. It is thus important to investigate their extensions from optimization to Bayesian sampling. Directly inserting the MUON direction into Langevin dynamics, however, changes the sampling target. To remedy this issue, we study two alternatives: Proximal-MUON Langevin Dynamics (PMLD) uses a Moreau–Yosida spectral prior, while Annealed PMLD (APMLD) uses a finite MUON transient before Langevin sampling. We establish spectral localization for PMLD, an APMLD warm-start bound under strong convexity and smoothness, and a common finite-time sampling bound. Experiments on Bayesian classifiers and diffusion models show improved robustness for PMLD and finite-time posterior-predictive performance for APMLD.

1 INTRODUCTION

Neural-network parameters are naturally organized into matrices (Shi et al., 2026), whose singular directions describe coordinated changes across input and output coordinates. MUON exploits this structure by approximately orthogonalizing momentum updates, reducing the dominance of large singular values while retaining their singular directions (Jordan et al., 2024). For Bayesian neural networks, the goal is to represent uncertainty through a posterior distribution over these parameters. This raises a natural question: how can MUON-inspired spectral geometry inform the distribution we sample or the way we initialize sampling?

Langevin dynamics provide a direct connection between gradient updates and posterior sampling: the drift follows the negative gradient of the target potential, while Brownian noise enables exploration. This connection also imposes a constraint. Replacing the negative gradient by a MUON direction while retaining isotropic Brownian noise generally changes the invariant law. An update designed for optimization therefore requires an explicit probabilistic interpretation before it can be used for sampling. We study two routes for introducing MUON-inspired geometry: encoding spectral structure in the target and using MUON during a finite initialization phase.

The first route gives **Proximal-MUON Langevin Dynamics (PMLD)**, operationally the unadjusted Langevin algorithm (ULA) with a spectral Moreau prior. We specialize the Moreau–Yosida construction to a spectral-norm ball, penalizing excursions beyond a prescribed radius. The penalty is differentiable and admits a gradient expressed through the singular values of each parameter matrix. It defines a soft posterior whose spectral concentration can be quantified under a normalized base law. Here "proximal" describes the construction of the prior rather than the transition kernel.

The second route gives **Annealed Proximal-MUON Langevin Dynamics (APMLD)**, operationally MUON-initialized ULA. A finite MUON transient selects an iterate that is then perturbed by Gaussian noise to initialize Langevin sampling for the chosen target. The transient affects the initial distribution, while the subsequent Langevin drift is determined by the sampling potential.

A common finite-time analysis then connects both constructions to initialization, mixing, and discretization error. Table 1 places the resulting iteration bounds alongside representative Langevin and composite samplers. These guarantees retain the dimension–accuracy order of a ULA-type analysis; they make the two intervention points explicit but do not establish a faster worst-case sampling order.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Representative convergence guarantees for Langevin-type samplers. Rates are iteration complexities to accuracy ε; constants and assumptions differ across papers.

Method	Target class	Representative rate
ULA/LMC (Dalalyan, 2017; Durmus & Moulines, 2017)	smooth strongly log-concave	$O\big(\kappa^2 d\,\varepsilon^{-2}\log(\frac{1}{\varepsilon})\big)$
MALA (Dwivedi et al., 2019)	smooth strongly log-concave	$O\big(\kappa d\log(\frac{\kappa}{\varepsilon})\big)$
RGO composite sampler (Liu & Chewi, 2026)	$f+g$ strongly convex; restricted Gaussian oracle	$\widetilde{O}\Big(\kappa\sqrt{d}\,\log^4(\frac{1}{\varepsilon})\Big)$
PMLD/APMLD, ours (Theorem F.1 and Corollary F.1)	spectral Moreau target or post-MUON phase	$O\big(\kappa_g^2 d\,\varepsilon^{-2}\log^2\max\{d,\frac{1}{\varepsilon}\}\big)$

Our contributions are summarized as follows:

- **Transferring the optimization geometry to sampling.** We introduce two routes for using MUON-inspired geometry in Bayesian sampling: PMLD shapes the posterior through a spectral prior, while APMLD shapes the initialization for a chosen target. This separation allows spectral control of posterior samples or improved finite-budget initialization without relying on a MUON drift to preserve the desired sampling law. A common bound makes their roles in target specification and finite-time sampling error explicit (Proposition 4.3).
- **Controllable spectral concentration and sample-wise robustness.** PMLD makes spectral concentration controllable through the prior, yielding a conditional robustness certificate for sampled networks (Proposition 4.1 and Corollary C.1). In the Fashion-MNIST experiments, this construction delivers lower prediction error and better calibration under FGSM perturbations, with comparable clean performance (Figure 2).
- **Better predictive performance within finite sampling budgets.** APMLD improves early predictive accuracy and calibration over ULA and Adam-initialized ULA at comparable wall-clock cost in the CIFAR-100 experiment, while retaining the same post-switch target (Figure 2). Its reported FID and KID improvements across the tested diffusion settings further support the practical value of MUON initialization (Table 2).

2 Related Work

Matrix geometry and MUON. Matrix-aware optimization includes spectral descent, tensor preconditioning, and gradient orthogonalization (Carlson et al., 2015; Gupta et al., 2018; Tuddenham et al., 2022). MUON uses Newton–Schulz iterations to approximately orthogonalize momentum updates (Jordan et al., 2024). With positive decoupled weight decay, its optimization geometry has been connected to spectral-norm constraints through a KKT characterization (Chen et al., 2025). A complementary analysis interprets idealized deterministic normalized training dynamics as spectral Wasserstein flows (Peyré, 2026).

Langevin sampling and initialization. Nonasymptotic analyses of the unadjusted Langevin algorithm quantify convergence to smooth log-concave targets and its dependence on initialization and discretization (Dalalyan, 2017; Durmus & Moulines, 2017). Metropolis-adjusted Langevin methods correct discretization bias through an accept–reject step and can improve the accuracy dependence of mixing bounds (Dwivedi et al., 2019). Table 1 records representative guarantees together with their initialization requirements. APMLD controls initialization through the stationarity of the selected MUON iterate and its Gaussian perturbation. The resulting KL and moment bounds need not imply the bounded density ratio assumed by a warm-start MALA theorem.

Proximal sampling and spectral priors. Proximal MCMC methods use proximity operators and Moreau–Yosida regularization to handle nonsmooth log-concave targets (Pereyra, 2016; Brosse et al., 2017). More recent composite samplers obtain high-accuracy guarantees using restricted Gaussian oracles (RGOs), with additional access to a mode and a proximal operator (Liu & Chewi, 2026). These oracle requirements differ from evaluating PMLD's spectral envelope gradient. PMLD adopts the smooth spectral penalty as part of the Bayesian model: its soft posterior is the intended target, and localization quantifies concentration around a spectral-norm ball. Accordingly, its rate in Table 1 measures error to that soft posterior, not to the unmodified base law or a hard-constrained posterior.

3 Background and Problem Setup

Notation. A boldface uppercase letter denotes a matrix ($\mathbf{X} \in \mathbb{R}^{m\times n}$), a boldface lowercase letter denotes a vector ($\mathbf{x} \in \mathbb{R}^n$), and a plain letter denotes a scalar ($x \in \mathbb{R}$). The inner product between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is denoted by $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^n x_i y_i$. For matrices of the same shape, the Frobenius inner product is $\langle \mathbf{X}, \mathbf{Y}\rangle_F = \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{Y})$, and the Frobenius norm is $\|\mathbf{X}\|_F = \sqrt{\langle \mathbf{X}, \mathbf{X}\rangle_F}$. We use $\|\mathbf{x}\|_2$ and $\|\mathbf{X}\|_2$ for the Euclidean norm of a vector and the operator norm of a matrix, respectively. For one matrix block $\mathbf{X} \in \mathbb{R}^{m\times n}$, we write $d = mn$ and $r = \min\{m, n\}$. The nuclear norm $\|\mathbf{G}\|_*$ is the sum of its singular values. Gaussian matrix laws and Brownian motions are understood after vectorization; the resulting Euclidean norm equals the Frobenius norm $\|\cdot\|_F$.

For two probability measures μ and ν on a measurable space $(\mathbb{X}, \mathcal{B})$, the total variation (TV) distance is $\|\mu - \nu\|_{\mathrm{TV}} := \sup_{B\in\mathcal{B}} |\mu(B) - \nu(B)|$. If $\mu \ll \nu$, the Kullback–Leibler (KL) divergence is $\mathrm{KL}(\mu\|\nu) := \int_{\mathbb{X}} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu$, and $\mathrm{KL}(\mu\|\nu) = +\infty$ otherwise. When μ and ν admit densities p, q with respect to a common dominating measure, $\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\int |p - q|$.

MUON and spectral geometry. For a weight matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$ and loss $\mathcal{L}$, gradient descent uses the gradient $\mathbf{G} = \nabla_{\mathbf{W}}\mathcal{L}$ and updates $\mathbf{W} \leftarrow \mathbf{W} - \eta\mathbf{G}$, where η is the stepsize. The deterministic MUON update without momentum considered here keeps the gradient's *singular directions* but replaces its positive singular values by one via the matrix polar factor. For a full-rank matrix $\mathbf{G}$,

$$\mathrm{Ortho}(\mathbf{G}) = \begin{cases} \mathbf{G}(\mathbf{G}^\mathsf{T}\mathbf{G})^{-1/2}, & m \geq n \\ (\mathbf{G}\mathbf{G}^\mathsf{T})^{-1/2}\mathbf{G}, & m < n. \end{cases}$$

More generally, with a compact SVD $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, we define the polar partial isometry by

$$\mathrm{Ortho}(\mathbf{G}) = \mathbf{U}\mathbf{V}^\mathsf{T}.$$

Here the compact SVD contains only positive singular values, and we set $\mathrm{Ortho}(0) = 0$. Thus MUON replaces the Euclidean gradient direction with one adapted to spectral geometry: $\mathbf{W}_{\mathrm{MUON}} \leftarrow \mathbf{W}_{\mathrm{MUON}} - \eta\,\mathrm{Ortho}(\mathbf{G})$.

Langevin sampling and target preservation. Let $\pi(\boldsymbol{\theta}) \propto \exp\{-U(\boldsymbol{\theta})\}$ be a target distribution on $\mathbb{R}^d$. The overdamped Langevin diffusion is

$$\mathrm{d}\boldsymbol{\theta}_t = -\nabla U(\boldsymbol{\theta}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t, \tag{1}$$

where $\mathbf{B}_t$ denotes standard Brownian motion in $\mathbb{R}^d$. Under standard regularity conditions, π is invariant for (1). Euler–Maruyama discretization of this diffusion gives the unadjusted Langevin algorithm (ULA):

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\nabla U(\boldsymbol{\theta}_k) + \sqrt{2\eta}\,\boldsymbol{\xi}_{k+1}, \qquad \boldsymbol{\xi}_{k+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d).$$

The formal similarity between optimization updates and ULA might suggest replacing ∇U in (1) by a MUON-style direction such as $\mathrm{Ortho}(\nabla U)$. In general, however, this changes the drift without a corresponding change in the target density and therefore does not preserve the original Gibbs law. This target-preservation constraint is the central obstruction in transferring MUON directly from optimization to sampling.

4 Main Results

Based on this observation, we propose methods to address this obstruction in two different ways. PMLD transfers the spectral geometry into the target distribution itself, while APMLD confines the MUON dynamics to a finite transient before switching to a Langevin sampler for the chosen target as shown below. The distinction is the object being changed: PMLD changes the target potential and then applies ordinary ULA, whereas APMLD changes the initial law and then applies ordinary ULA to the chosen target. We first consider the target-side route and then the transient-side route.

4.1 PROXIMAL-MUON LANGEVIN DYNAMICS

While MUON modifies the singular spectrum of the gradient, Chen et al. (2025) relate MUON with positive decoupled weight decay to spectral-norm-constrained optimization through a KKT characterization, which motivates us to examine the prior geometry for sampling as well. We use this connection to define our target-side construction: rather than replacing the Langevin drift by a MUON direction, we encode the corresponding spectral geometry directly in the posterior over the model parameters.

Let $\mathcal{B}_R := \{\mathbf{W} : \|\mathbf{W}\|_2 \le R\}$ denote the spectral-norm ball. The most direct Bayesian realization of this geometry is the **posterior with a hard spectral prior:**

$$\pi_R^{\mathrm{hard}}(\mathbf{W}) \propto \exp\{-U(\mathbf{W}) - \iota_{\mathcal{B}_R}(\mathbf{W})\}, \qquad \iota_{\mathcal{B}_R}(\mathbf{W}) = \begin{cases} 0, & \|\mathbf{W}\|_2 \le R, \\ +\infty, & \text{otherwise.} \end{cases}$$

This posterior enforces the spectral constraint exactly, but its indicator potential is nonsmooth. Proximal sampling methods can accommodate such targets, including RGO-based methods (Liu & Chewi, 2026) and Prox-MALA (Pereyra, 2016), but they require either a nontrivial restricted Gaussian oracle or repeated projections within a Metropolis correction.

We instead retain the same spectral geometry through a **smooth spectral prior**. For a proper lower-semicontinuous convex function g, recall that its Moreau–Yosida envelope is defined by

$$M_{\gamma g}(\mathbf{W}) := \inf_{\mathbf{Z}} \left\{ g(\mathbf{Z}) + \frac{1}{2\gamma}\|\mathbf{W} - \mathbf{Z}\|_F^2 \right\}, \qquad \nabla M_{\gamma g}(\mathbf{W}) = \frac{1}{\gamma}\left(\mathbf{W} - \mathrm{prox}_{\gamma g}(\mathbf{W})\right),$$

where $\nabla M_{\gamma g}$ is $1/\gamma$-Lipschitz (Beck, 2017, Chapter 6). Here $\gamma > 0$, and the envelope is convex and continuously differentiable, even when g is extended-real-valued. A direct derivation of these facts is given in Appendix A.

For $g = \iota_{\mathcal{B}_R}$, the envelope has a simple closed form. Writing $\mathbf{W} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, we obtain

$$M_{\gamma \iota_{\mathcal{B}_R}}(\mathbf{W}) = \frac{1}{2\gamma}\sum_{i=1}^r (\sigma_i - R)_+^2, \quad \nabla M_{\gamma \iota_{\mathcal{B}_R}}(\mathbf{W}) = \frac{1}{\gamma}\sum_{i=1}^r (\sigma_i - R)_+ \mathbf{u}_i \mathbf{v}_i^\top.$$

We take this smooth spectral penalty as part of the Bayesian prior and define the PMLD target, which is a **posterior with a smooth spectral prior**:

$$\widetilde{\pi}_\gamma(\boldsymbol{\theta}) \propto \exp\left\{-\widetilde{U}_\gamma(\boldsymbol{\theta})\right\} := \exp\left\{-U(\boldsymbol{\theta}) - M_{\gamma \iota_{\mathcal{B}_R}}(\boldsymbol{\theta})\right\}.$$

Thus, $\widetilde{\pi}_\gamma$ is the posterior induced by the soft spectral prior rather than a numerical surrogate for the hard target. For differentiable U, the corresponding Langevin drift is $-\nabla \widetilde{U}_\gamma$, which gives the dynamics $\mathrm{d}\boldsymbol{\theta}_t = -\nabla_{\boldsymbol{\theta}} \widetilde{U}_\gamma(\boldsymbol{\theta}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t$. Its Euler–Maruyama discretization is $\widetilde{\boldsymbol{\theta}}_{k+1} = \widetilde{\boldsymbol{\theta}}_k - \eta \nabla \widetilde{U}_\gamma(\widetilde{\boldsymbol{\theta}}_k) + \sqrt{2}(\mathbf{B}_{(k+1)\eta} - \mathbf{B}_{k\eta})$ for $k \ge 0$ and the discrete iterate $\widetilde{\boldsymbol{\theta}}_k$ approximates the diffusion state $\boldsymbol{\theta}_{k\eta}$. Algorithm 1 summarizes PMLD. Its convergence analysis is given in Appendix F.

Algorithm 1 is therefore standard ULA for the explicitly modified target $\widetilde{\pi}_\gamma$; the proximal operation is used to construct and differentiate the prior, not as a separate Markov transition. Proposition 4.1 concerns localization under a normalized base law, while the finite-time ULA guarantees additionally require the global convexity and smoothness assumptions stated below.

To visualize how the prior changes the posterior geometry, we examine a two-dimensional slice of the parameter space spanned by MAP solutions (Figure 1). With the likelihood held fixed, replacing the isotropic Gaussian prior by the spectral Moreau prior visibly reshapes the energy landscape, concentrating low-energy regions according to spectral structure rather than only Frobenius shrinkage.

The main structural consequence of the PMLD target is explicit posterior localization in spectral norm. Although the Moreau prior is a soft rather than a hard constraint, posterior mass outside the prescribed spectral region decays exponentially with the squared spectral violation.

Proposition 4.1 (Soft spectral localization)**.** *Let $R > 0$, $\gamma > 0$, and let $\pi \propto e^{-U}$ be a probability measure with $\pi(\mathcal{B}_R) > 0$. Then for every $t > 0$: $\mathbb{P}_{\mathbf{W}\sim\widetilde{\pi}_\gamma}\left(\|\mathbf{W}\|_2 > R + t\right) \le \frac{1}{\pi(\mathcal{B}_R)} \exp\left(-\frac{t^2}{2\gamma}\right)$.*

Algorithm 1 Proximal-MUON Langevin Dynamics (PMLD)

Require: Initial model parameters $\boldsymbol{\theta}_0 = \mathbf{W}_0 \in \mathbb{R}^{m\times n}$, spectral cap $R \geq 0$, envelope parameter $\gamma > 0$, stepsize $\eta > 0$, and differentiable base potential U.

1: **for** $t = 1, \ldots$ **do**
2: Compute gradient: $\mathbf{G}_t \leftarrow \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_{t-1})$
3: Compute SVD of parameters: $\boldsymbol{\theta}_{t-1} = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^\mathsf{T}$
4: Compute envelope gradient: $\mathbf{E}_t = \nabla M_{\gamma \iota_{\mathcal{B}_R}}(\boldsymbol{\theta}_{t-1}) \leftarrow \frac{1}{\gamma} \sum_i (\sigma_i - R)_+ \mathbf{u}_i \mathbf{v}_i^\mathsf{T}$
5: Sample: $\boldsymbol{\theta}_t \sim \mathcal{N}\big(\boldsymbol{\theta}_{t-1} - \eta(\mathbf{G}_t + \mathbf{E}_t), 2\eta \mathbf{I}\big)$
6: **end for**
7: **return** samples of $\{\boldsymbol{\theta}_t\}_t$

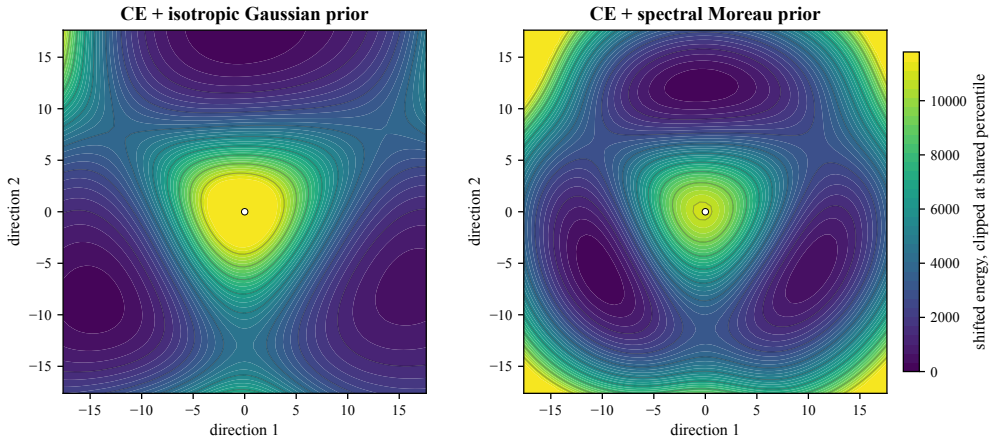


Figure 1: **Visualization of the posterior geometry induced by the PMLD prior.** We evaluate the negative log-posterior on the two-dimensional affine slice $\theta(\alpha, \beta) = \theta_0 + \alpha d_1 + \beta d_2$, where d_1, d_2 are orthonormal directions constructed from independently trained MAP solutions. Each point (α, β) therefore represents a network along the same parameter-space slice, with color indicating its shifted posterior energy. The two panels share the same likelihood and slice and differ only in the prior: the isotropic Gaussian prior (left) and the spectral Moreau prior used by PMLD (right).

More generally, let $N_t(\mathbf{W}) := \#\{i : \sigma_i(\mathbf{W}) > R + t\}$ *. Then for every integer* $q \geq 1$*, we have* $\mathbb{P}_{\mathbf{W}\sim\widetilde{\pi}_\gamma}\left(N_t(\mathbf{W}) \geq q\right) \leq \frac{1}{\pi(\mathcal{B}_R)} \exp\left(-\frac{qt^2}{2\gamma}\right)$.

The spectral localization result directly yields a high-probability robustness certificate for networks sampled from the PMLD posterior. In particular, posterior control of the layerwise spectral norms bounds the network Lipschitz constant and, together with a positive classification margin, certifies invariance of the predicted class to sufficiently small input perturbations. A detailed corollary for BNN posterior is given in Appendix C.

4.2 Annealed Proximal-MUON Langevin Dynamics

PMLD deliberately modifies the posterior in order to encode spectral structure. This raises a complementary question: can MUON improve the finite-time behavior of a sampler?

APMLD confines MUON to a finite initialization phase for the chosen smooth potential $U_g := U + M_{\gamma g}$. Starting from $\boldsymbol{\theta}_0 = \mathbf{W}_0$, it forms $T - 1$ MUON updates, selects the smallest index minimizing $\|\nabla U_g(\boldsymbol{\theta}_t)\|_*$ over $0 \leq t \leq T-1$, and injects independent Gaussian noise at that selected center. The original-posterior warm-start theorem below takes $g = 0$. After injection, the algorithm uses Euler–Maruyama sampling for U_g. The corresponding diffusion has the chosen invariant law; the fixed-step chain has discretization error. We summarize the resulting procedure as APMLD in Algorithm 2.

Algorithm 2 Annealed Proximal-MUON Langevin Dynamics (APMLD)

Require: Initial model parameters $\boldsymbol{\theta}_0 = \mathbf{W}_0 \in \mathbb{R}^{m\times n}$; chosen target potential $U_g = U + M_{\gamma g}$; annealing time $T \geq 1$; MUON stepsize $\eta_{\mathrm{MUON}} \geq 0$; injected variance $\sigma^2 > 0$; sampling stepsize $\eta > 0$.

1: **for** $t = 1, \ldots, T-1$ **do**
2: $\quad \boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta_{\mathrm{MUON}}\, \mathrm{Ortho}(\nabla U_g(\boldsymbol{\theta}_{t-1}))$
3: **end for**
4: $t^* \leftarrow \arg\min_{0\leq t\leq T-1} \|\nabla U_g(\boldsymbol{\theta}_t)\|_*$
5: Draw $\boldsymbol{\theta}_T \sim \mathcal{N}(\boldsymbol{\theta}_{t^*}, \sigma^2 \mathbf{I}_d)$ independently of future sampling noise.
6: **for** $k = 1, 2, \ldots$ **do**
7: $\quad \boldsymbol{\theta}_{T+k} \sim \mathcal{N}(\boldsymbol{\theta}_{T+k-1} - \eta \nabla U_g(\boldsymbol{\theta}_{T+k-1}), 2\eta \mathbf{I})$
8: **end for**
9: **return** samples of $\{\boldsymbol{\theta}_t\}_{t\geq T}$

For proper lower-semicontinuous convex g and differentiable U, $\nabla U_g = \nabla U + \gamma^{-1}(I - \mathrm{prox}_{\gamma g})$. No gradient of g itself is required. The choice of g is flexible here: with $g = 0$, the sampling phase targets $\pi \propto e^{-U}$; with $g = \iota_{\mathcal{B}_R}$, it uses exactly the spectral Moreau drift of PMLD. The transient and sampling stepsizes are separate parameters.

The remaining question is whether the randomized iterate produced by the MUON transient constitutes a useful warm start. Under strong convexity and smoothness, the stationarity attained during the transient can be converted into an explicit distributional bound.

Assumption 4.1 (Strong convexity and smoothness)**.** Throughout this assumption, U is continuously differentiable on $\mathbb{R}^{m\times n}$ and $0 < M \leq L < \infty$. Assume U is M-strongly convex with unique minimizer $\mathbf{W}^\star$, and it has L-Lipschitz gradient:

$$\text{(i)}\ \langle \nabla U(\mathbf{X}) - \nabla U(\mathbf{Y}), \mathbf{X} - \mathbf{Y}\rangle_F \geq M\|\mathbf{X}-\mathbf{Y}\|_F^2, \qquad \text{(ii)}\ \|\nabla U(\mathbf{X}) - \nabla U(\mathbf{Y})\|_F \leq L\|\mathbf{X}-\mathbf{Y}\|_F.$$

Proposition 4.2 (Original-posterior APMLD warm start)**.** *Under Assumption 4.1, let $d = mn$, $r = \min\{m, n\}$, and let $\mathbf{W}_0$ be a deterministic initial matrix. Use Algorithm 2 with $g = 0$, exact compact polar factors, integer $T \geq 1$, and $\sigma^2 = 1/L$. Write $D = U(\mathbf{W}_0) - U(\mathbf{W}^\star)$ and $\kappa = L/M$. Let $\overline{D} \geq D$ be a known finite upper bound. If $\overline{D} > 0$, set $\eta_{\mathrm{MUON}} = \sqrt{2\overline{D}/(LrT)}$. If $\overline{D} = 0$, leave the transient iterates fixed at $\mathbf{W}^\star$. For the law $\widehat{\pi}$ of the injected $\mathbf{Y}_0$, we have* $\|\widehat{\pi} - \pi\|_{\mathrm{TV}} \leq \min\left\{1, \sqrt{\frac{\kappa\overline{D}r}{2T} + \frac{d}{4}\log\kappa}\right\}, \mathbb{E}\|\mathbf{Y}_0 - \mathbf{W}^\star\|_F^2 \leq \frac{2\overline{D}Lr}{M^2T} + \frac{d}{L}$.

The squared TV bound retains a Gaussianization term $d\log\kappa/4$. It can be vacuous when this term is large; increasing κ does not improve it. The r/T term alone does not prove an improvement over a different initialization or establish asymptotic acceleration. This is an initialization bound, not a permanent sampling-error floor. Thus Proposition 4.2 analyzes a particular randomized initializer rather than a new stationary sampler.

4.3 Finite-Time Sampling Error

The preceding constructions intervene at different stages of Bayesian sampling: PMLD modifies the target distribution, whereas APMLD modifies the initialization for a chosen target. After these interventions, however, both methods reduce to Langevin sampling from a prescribed smooth target. We therefore conclude with a common finite-time view of the two constructions.

Let $\pi_g(\boldsymbol{\theta}) \propto \exp\{-U_g(\boldsymbol{\theta})\}$, where $U_g := U + M_{\gamma g}$, denote the target of the Langevin phase, and let ρ_0 denote its initial distribution. For PMLD, $g = \iota_{\mathcal{B}_R}$ and $\pi_g = \widetilde{\pi}_\gamma$; for APMLD targeting the original posterior, $g = 0$ and $\rho_0 = \widehat{\pi}$. The explicit discretization bound and resulting iteration complexity are given in Appendix F.

Proposition 4.3 (Finite-time sampling bound)**.** *Suppose U_g is M-strongly convex and L_g-smooth, with $0 < M \leq L_g$. For $\eta > 0$ and integer $k \geq 0$, let $\nu_{k\eta}$ denote the law after k Euler–Maruyama steps, initialized from ρ_0, and let $\mu_{k\eta}$ denote the continuous-time Langevin law with the same initial*

distribution. Then

$$\|\nu_{k\eta} - \pi_g\|_{\mathrm{TV}} \le \underbrace{\|\nu_{k\eta} - \mu_{k\eta}\|_{\mathrm{TV}}}_{\text{discretization}} + \underbrace{\sqrt{\frac{1}{2}\,\mathrm{KL}(\rho_0\|\pi_g)}e^{-M\eta k}}_{\text{initialization and mixing}},$$

where the PMLD and APMLD constructions affect complementary components of this bound: PMLD determines the target π_g, while APMLD controls the initial distribution ρ_0 entering the mixing term.

For the spectral target $\widetilde{\pi}_\gamma$, let $\nu_{k\eta}$ be the ULA law from any initialization, including APMLD with $g = \iota_{\mathcal{B}_R}$. Under Proposition 4.1's assumptions, the defining event bound for TV gives, for $t > 0$,

$$\mathbb{P}_{\nu_{k\eta}}(\|W\|_2 > R + t) \le \min\left\{1,\ \underbrace{\frac{e^{-t^2/(2\gamma)}}{\pi(\mathcal{B}_R)}}_{\text{equilibrium localization}} + \underbrace{\|\nu_{k\eta} - \widetilde{\pi}_\gamma\|_{\mathrm{TV}}}_{\text{finite-time sampling error}}\right\}. \tag{2}$$

PMLD controls the first term through the spectral prior; APMLD changes the initialization entering the second, bounded above under Proposition 4.3's additional assumptions. Smaller γ sharpens the localization bound but raises the conservative smoothness constant $L + 1/\gamma$, affecting discretization. For a fixed quadratic target, Appendix E shows how the remaining horizon determines which center errors matter. This explains how the routes complement each other without asserting that MUON always improves initialization or that a warm start changes equilibrium tails.

5 Experiments

Equation 2 separates target localization from finite-time sampling error. We probe these roles separately through PMLD robustness and APMLD early posterior-predictive performance. The nonconvex networks and approximate polar factors fall outside the convex guarantees; predictive metrics do not directly verify either term of the bound or the quadratic mechanism in Appendix E. Metric and attack definitions are in Appendix G.

5.1 Bayesian MLP

5.1.1 PMLD: spectral landscape and robustness on Fashion-MNIST

We evaluate PMLD on Fashion-MNIST using a fully Bayesian $784-256-256-10$ ReLU MLP, with $10{,}000$ training and $2{,}000$ test images. We compare a Gaussian-prior ULA baseline with PMLD, which replaces the Gaussian weight penalty by the aggregated Moreau-envelope prior $\sum_\ell \frac{1}{2\gamma}\,\mathrm{dist}_F^2\left(W_\ell, \{W : \|W\|_2 \le R_\ell\}\right)$. Both chains start from the same MAP solution, with $R_\ell = 0.50\|W_{\ell,\mathrm{MAP}}\|_2$ and $\gamma = 10^{-4}$. We use step sizes 2×10^{-7} for ULA and 2×10^{-8} for PMLD, run each chain for $3{,}000$ steps, and average over three seeds. Further implementation details are deferred to the appendix.

Figure 2 compares clean and FGSM-perturbed posterior-predictive performance. While the methods are comparable on clean prediction, the advantage of PMLD becomes pronounced under perturbation: PMLD maintains lower NLL, classification error, and ECE as the attack strength increases. This behavior is qualitatively consistent with the spectral-control motivation. Figure 2 evaluates the averaged predictor under an ℓ_∞ FGSM perturbation constructed from mean per-network cross-entropy.

5.1.2 APMLD: finite-time initialization on CIFAR-100

We evaluate APMLD on CIFAR-100 using frozen ResNet-18 features and Bayesian inference over a $d_{\mathrm{feat}}-1024-1024-100$ ReLU classification head. We use $20{,}000$ training and $5{,}000$ test observations with an isotropic Gaussian prior of precision 10^{-4}.

We compare ULA, Adam+ULA, and APMLD under the same post-switch posterior target. Adam+ULA and APMLD use 800 warm-start steps followed by $1{,}200$ ULA steps, with APMLD

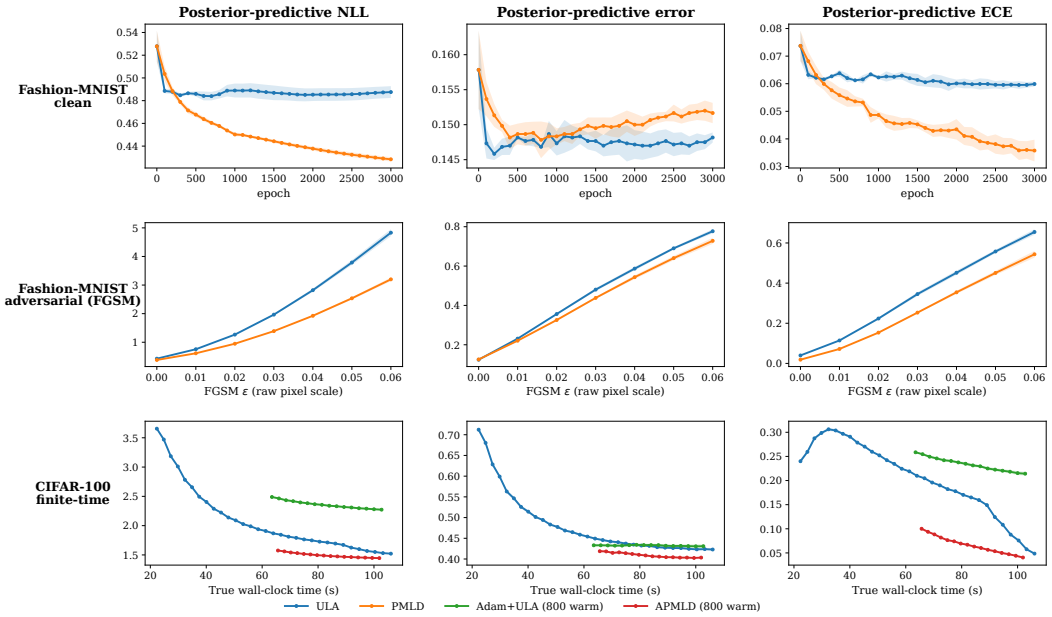


Figure 2: Bayesian MLP experiments for PMLD and APMLD. The first row compares Gaussian-prior ULA and spectral PMLD on clean Fashion-MNIST prediction as a function of sampling epoch. The second row evaluates the same two methods under FGSM perturbations on Fashion-MNIST. The third row compares ULA, Adam+ULA, and APMLD on CIFAR-100 frozen ResNet-18 features as a function of corrected wall-clock time. Columns report posterior-predictive NLL, classification error, and expected calibration error; lower is better in all panels.

implementing its matrix-aware updates through a Newton–Schulz approximation to the polar factor. The ULA step size is 10^{-6}, and results are averaged over five seeds.

Consequently, this experiment evaluates the practical warm-start hypothesis rather than Proposition 4.2's KL or stationarity bound: the classification-head posterior is nonconvex and the polar factor is approximate, while the proposition assumes global strong convexity and an exact compact polar factor. All methods nevertheless use the same post-switch target, so differences after the switch isolate finite-time initialization behavior under the reported implementations. Figure 2 reports performance against corrected wall-clock time. APMLD reaches lower posterior-predictive NLL, classification error, and ECE at comparable wall-clock cost. Its advantage over Adam+ULA, particularly in calibration, indicates that in this setting the observed finite-time benefit is not reproduced by the tested conventional optimization warm start.

5.2 Bayesian Diffusion Models

We next test whether the two constructions extend beyond classification. We Bayesianize selected weight matrices of pretrained diffusion models while keeping the remaining parameters fixed. PMLD samples these weights under the spectral prior, whereas APMLD applies a finite MUON warm start before Langevin sampling. Generated images are evaluated using retained posterior weight samples.

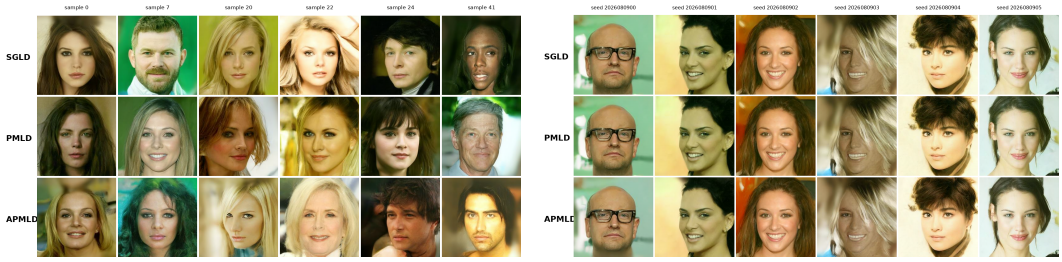
CIFAR-10. We evaluate a DDPM at deterministic pretraining checkpoints of 20k and 50k updates. All methods start from the same pretrained network and use $1{,}000$ posterior-sampling updates, 10 retained weight samples, and 5,000 generated images.

CelebA-HQ. We additionally evaluate a pretrained 256×256 CelebA-HQ DDPM, Bayesianizing a selected hidden convolutional weight matrix in the final upsampling stage while freezing the remaining denoiser. Five posterior samples are retained. We report a 1,000-image/250-step protocol and a more expensive 5,000-image/1,000-step protocol.

Table 2: Posterior-sampler comparison for Bayesian diffusion models. CIFAR-10 results use 5,000 generated images and 10 retained posterior weight samples; 20k and 50k denote deterministic DDPM pretraining checkpoints. CelebA-HQ results use five retained posterior weight samples and are reported under both the 1k-image/250-step and 5k-image/1000-step reverse-diffusion protocols. Lower FID and KID and higher diversity and quality-adjusted diversity (QAD) are preferred; Lag-1 is a descriptive generated-image statistic. Best values within each setting are shown in bold.

Dataset	Setting	Method	FID ↓	KID ↓	Diversity ↑	Lag-1 ↓	QAD ↑
CIFAR-10	20k pretrain	SGLD	56.6649	0.04047	**0.2323**	0.2160	0.00403
		PMLD	50.8699	0.03525	0.2197	**0.1678**	0.00423
		APMLD	**48.7582**	**0.02950**	0.2230	0.2210	**0.00448**
	50k pretrain	SGLD	56.8123	0.03908	**0.2418**	0.2688	0.00418
		PMLD	50.4025	0.03296	0.2301	**0.1856**	0.00448
		APMLD	**48.4159**	**0.02913**	0.2252	0.2519	**0.00456**
CelebA-HQ	1k / 250 steps	SGLD	64.9292	0.04782	0.05454	0.23605	0.000827
		PMLD	63.3931	0.04597	**0.05455**	0.23517	**0.000847**
		APMLD	**63.1381**	**0.04520**	0.05159	**0.10310**	0.000804
	5k / 1000 steps	SGLD	46.3556	0.04634	0.06747	0.22320	0.001425
		PMLD	46.5777	0.04645	0.06747	0.22367	0.001418
		APMLD	**44.9385**	**0.04468**	**0.07320**	**0.02478**	**0.001593**

Table 2 shows consistent improvements from APMLD: it achieves the best FID and KID at both CIFAR-10 checkpoints and under both CelebA-HQ protocols. PMLD also improves FID and KID over SGLD on CIFAR-10 and yields lower lag-1 correlation, while its advantage is less consistent on CelebA-HQ. Together, these results suggest that the finite MUON transient extends beyond classification, while the benefit of the spectral prior can depend on the modeling setting. These reported scores measure generated-image quality, not distance to the parameter posterior; without repeated-run uncertainty they do not establish the significance of the differences.



(a) Representative samples (b) Matched-seed samples

Figure 3: CelebA-HQ-256 qualitative results.

6 Conclusion

In this paper, we demonstrate two complementary benefits of MUON-inspired geometry for Bayesian learning: controllable posterior spectral structure and improved prediction within finite sampling budgets. PMLD provides a smooth spectral prior with explicit localization bounds and conditional samplewise robustness certificates. APMLD confines MUON to initialization while retaining the chosen target for subsequent Langevin sampling. Under the stated assumptions, our finite-time analysis connects these routes by separating equilibrium spectral localization from sampling error. Empirically, PMLD improves prediction and calibration under the tested FGSM perturbations with comparable clean performance. APMLD improves early predictive performance over ULA and Adam-initialized ULA at comparable wall-clock cost and achieves the lowest reported FID and KID among the tested diffusion samplers. These results demonstrate practical benefits beyond the convex setting without establishing faster worst-case mixing. They support a concrete principle for transferring optimization geometry to sampling: spectral priors control where posterior mass concentrates, while matrix-aware initialization helps turn limited sampling budgets into better predictions.

Reproducibility statement

We provide a supplementary ZIP containing experiment implementations, documented settings, per-seed and aggregate result files, selected sampled weights, plotting scripts, and file-integrity checks. The package's `README.md` provides commands for regenerating the classification summary figure and exporting the tabulated generative metrics from the supplied result files. `EXPERIMENTS.md` maps the experiments to their implementations and outputs, and `VALIDATION.md` describes the software requirements and checks performed. The archive contains the experiment-specific inputs needed for these result-processing workflows. Input requirements and artifacts not included for checkpoint-dependent evaluations and full training or sampling runs are documented in `ARTIFACTS.md`.

Ethics statement

This work is methodological and studies posterior sampling algorithms for Bayesian learning. The experiments use standard benchmark datasets and pretrained generative models; no new human subject data are collected. As with other Bayesian deep-learning and generative-model methods, improved sampling may support beneficial uncertainty quantification and robustness analysis, but could also be used in downstream systems whose predictions affect people. The reported robustness and calibration results are benchmark-level evaluations and should not be interpreted as deployment guarantees. Experiments involving face imagery are used only as standard generative-model benchmarks; the methods are not proposed for identification, surveillance, or demographic inference. We encourage users to evaluate dataset bias, privacy constraints, and downstream harms before applying these samplers in real systems.

AI use statement

In this work, we used generative AI tools to assist with code writing, code execution, debugging, repository organization, reproducibility checks, and editing of implementation and documentation artifacts. We have not used generative AI tools for mathematical derivation, theorem proving, algorithm design, conceptualization of the proposed methods, or formulation of the main scientific claims, and other disclosure categories not related to code or documentation support are not applicable to this work. Additionally, we used generative AI tools for light copyediting and release-engineering support for the supplementary code package. We reviewed all AI-assisted work: generated or edited code was inspected and tested by the authors, execution outputs were checked against the released artifacts and expected experimental records, and documentation changes were reviewed for accuracy and consistency with the actual package contents. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

References

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997.

Nicolas Brosse, Alain Durmus, Eric Moulines, and Marcelo Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 319–342. PMLR, 2017.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted boltzmann machines. In *Artificial intelligence and statistics*, pp. 111–119. PMLR, 2015.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017. doi: 10.1111/rssb.12183.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017. doi: 10.1214/16-AAP1238.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

Ivan Gentil. Logarithmic sobolev inequality for diffusion semigroups. *arXiv preprint arXiv:1009.3421*, 2010.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Linghai Liu and Sinho Chewi. A proximal gradient algorithm for composite log-concave sampling. *arXiv preprint arXiv:2605.12461*, 2026.

Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4): 745–760, 2016.

Gabriel Peyré. Muon dynamics as a spectral wasserstein flow. *arXiv preprint arXiv:2604.04891*, 2026.

Haoming Shi, Eric C. Chi, and Hengrui Luo. Asymptotic optimism for tensor regression models with applications to neural network compression. *arXiv preprint arXiv:2603.26048*, 2026.

Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*, 2022.

APPENDIX

Appendix A establishes the envelope projection identities. Appendix B proves Proposition 4.1, and Appendix C states and proves Corollary C.1, covering spectral localization and the sampled-network robustness certificate. Appendix D proves Proposition 4.2 using the Gaussian KL estimate and the MUON descent argument. Appendix F proves Proposition 4.3 and develops the mixing, discretization, and iteration-complexity bounds for PMLD and APMLD. Appendix G provides supplementary experimental definitions.

A ENVELOPE AND SPECTRAL PROJECTION DETAILS

The following arguments use the Frobenius inner product, or equivalently the Euclidean inner product after vectorization. For a proper lower-semicontinuous convex g and $\gamma > 0$, the proximal objective has a unique minimizer $\operatorname{prox}_{\gamma g}(x)$: an affine lower bound for g gives coercivity, and the quadratic term gives strict convexity. Optimality gives $(x - \operatorname{prox}_{\gamma g}(x))/\gamma \in \partial g(\operatorname{prox}_{\gamma g}(x))$. Monotonicity of the subdifferential implies

$$\langle x - y, \operatorname{prox}_{\gamma g}(x) - \operatorname{prox}_{\gamma g}(y)\rangle \geq \| \operatorname{prox}_{\gamma g}(x) - \operatorname{prox}_{\gamma g}(y)\|^2.$$

Consequently both $\operatorname{prox}_{\gamma g}$ and $I - \operatorname{prox}_{\gamma g}$ are nonexpansive. Evaluating the envelope at y using the minimizer at x gives

$$M_{\gamma g}(y) \leq M_{\gamma g}(x) + \langle \gamma^{-1}(x - \operatorname{prox}_{\gamma g}(x)), y - x\rangle + \frac{\|y - x\|^2}{2\gamma}.$$

The reverse comparison, together with continuity of $I - \operatorname{prox}_{\gamma g}$, proves differentiability with $\nabla M_{\gamma g} = \gamma^{-1}(I - \operatorname{prox}_{\gamma g})$ and Lipschitz constant γ^{-1}. The envelope is convex because it is the infimal convolution of convex g and the quadratic norm; alternatively, use their joint convexity and take infima over the second coordinate.

B Proof of Proposition 4.1

Proof of Proposition 4.1. Recall

$$M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W}) = \frac{1}{2\gamma}\operatorname{dist}_F^2(\mathbf{W}, \mathcal{B}_R) = \frac{1}{2\gamma}\sum_i \big(\sigma_i(\mathbf{W}) - R\big)_+^2.$$

Recall that the soft spectral posterior can be written as the exponential tilt

$$\widetilde{\pi}_\gamma(\mathrm{d}\mathbf{W}) = \frac{\exp\{-M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W})\}}{\mathbb{E}_\pi\left[\exp\{-M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W})\}\right]}\,\pi(\mathrm{d}\mathbf{W}),$$

Since $M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W}) = 0$ for every $\mathbf{W} \in \mathcal{B}_R$, we have

$$\mathbb{E}_\pi\left[\exp\{-M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W})\}\right] \geq \int_{\mathcal{B}_R} 1\,\pi(\mathrm{d}\mathbf{W}) = \pi(\mathcal{B}_R).$$

Now fix $t > 0$ and define

$$A_t := \{\mathbf{W} : \|\mathbf{W}\|_2 > R + t\}.$$

For every $\mathbf{W} \in A_t$, the largest singular value satisfies $\sigma_1(\mathbf{W}) > R + t$, and therefore

$$M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W}) = \frac{1}{2\gamma}\sum_i \big(\sigma_i(\mathbf{W}) - R\big)_+^2 \geq \frac{t^2}{2\gamma}.$$

Hence

$$\exp\{-M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W})\} \leq \exp\left(-\frac{t^2}{2\gamma}\right) \qquad \text{for all } \mathbf{W} \in A_t.$$

It follows that

$$\begin{aligned}\mathbb{P}_{\mathbf{W}\sim\widetilde{\pi}_\gamma}(A_t) &= \frac{\mathbb{E}_\pi[\mathbf{1}_{A_t} e^{-M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W})}]}{\mathbb{E}_\pi[e^{-M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W})}]} \\ &\leq \frac{e^{-t^2/(2\gamma)}\pi(A_t)}{\pi(\mathcal{B}_R)} \leq \frac{e^{-t^2/(2\gamma)}}{\pi(\mathcal{B}_R)}.\end{aligned}$$

For the multiplicity statement, define

$$N_t(\mathbf{W}) := \#\{i : \sigma_i(\mathbf{W}) > R + t\}.$$

On the event $\{N_t(\mathbf{W}) \geq q\}$, at least q singular values satisfy $\sigma_i(\mathbf{W}) - R > t$, and thus

$$M_{\gamma\iota_{\mathcal{B}_R}}(\mathbf{W}) \geq \frac{qt^2}{2\gamma}.$$

Repeating the same argument yields

$$\mathbb{P}_{\mathbf{W}\sim\widetilde{\pi}_\gamma}\left(N_t(\mathbf{W}) \geq q\right) \leq \frac{1}{\pi(\mathcal{B}_R)}\exp\left(-\frac{qt^2}{2\gamma}\right).$$

□

C BNN posterior robustness under spectral localization

Corollary C.1. *Assume the joint base law $\pi \propto e^{-U}$ is a probability measure, all R_ℓ, γ_ℓ are positive, and the network is a composition of affine layers and ReLU activations. Biases, if present, are fixed or included in a normalized joint base law.*

Consider an L-layer ReLU classifier with weight matrices $\mathbf{W}_1, \ldots, \mathbf{W}_L$, and suppose that PMLD uses the layerwise spectral penalty $\sum_{\ell=1}^L \frac{1}{2\gamma_\ell}\operatorname{dist}_F^2(\mathbf{W}_\ell, \mathcal{B}_{R_\ell})$. Let $p_{\mathbf{R}} :=$

$\pi\left(\bigcap_{\ell=1}^{L}\{\|\mathbf{W}_\ell\|_2 \le R_\ell\}\right) > 0$. *Then, for any* $t_1, \dots, t_L > 0$, *with probability at least* $1 - \frac{1}{p_{\mathbf{R}}}\sum_{\ell=1}^{L}\exp\left(-\frac{t_\ell^2}{2\gamma_\ell}\right)$ *under the soft PMLD posterior, the sampled network satisfies*

$$\|\operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\boldsymbol{\delta}) - \operatorname{Logit}_{\mathbf{W}}(\mathbf{x})\|_2 \le \left[\prod_{\ell=1}^{L}(R_\ell + t_\ell)\right]\|\boldsymbol{\delta}\|_2. \tag{3}$$

Moreover, let i *be the predicted class at* $\mathbf{x}$ *and define* $m_{\mathbf{W}}(\mathbf{x}, i) := \operatorname{Logit}_{\mathbf{W}}(\mathbf{x})^{(i)} - \max_{j\neq i}\operatorname{Logit}_{\mathbf{W}}(\mathbf{x})^{(j)}$. *If* $m_{\mathbf{W}}(\mathbf{x}, i) > 0$, *then on the same event every perturbation satisfying*

$$\|\boldsymbol{\delta}\|_2 < \frac{m_{\mathbf{W}}(\mathbf{x}, i)}{2\prod_{\ell=1}^{L}(R_\ell + t_\ell)} \tag{4}$$

preserves the predicted class.

Proof of Corollary C.1. Write $H(\mathbf{W}) = \sum_{\ell=1}^{L}\operatorname{dist}_F^2(\mathbf{W}_\ell, \mathcal{B}_{R_\ell})/(2\gamma_\ell)$ and $Z_H = \mathbb{E}_\pi e^{-H}$. On the intersection of the unexpanded spectral balls, $H = 0$, so $Z_H \ge p_{\mathbf{R}} > 0$. For $A_\ell = \{\|\mathbf{W}_\ell\|_2 > R_\ell + t_\ell\}$, nonnegativity of every layer penalty gives $H \ge t_\ell^2/(2\gamma_\ell)$ on A_ℓ. It follows that

$$\widetilde{\pi}_\gamma(A_\ell) = Z_H^{-1}\mathbb{E}_\pi[1_{A_\ell}e^{-H}] \le p_{\mathbf{R}}^{-1}e^{-t_\ell^2/(2\gamma_\ell)}.$$

The union bound therefore gives the stated lower probability for $\mathcal{E} = \bigcap_\ell A_\ell^c$. No independence between layers is required. It therefore suffices to establish the stated robustness bounds on $\mathcal{E}$.

Since ReLU is 1-Lipschitz with respect to the Euclidean norm, the perturbation δ propagates through the network according to

$$\|\operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\delta) - \operatorname{Logit}_{\mathbf{W}}(\mathbf{x})\|_2 \le \left(\prod_{\ell=1}^{L}\|\mathbf{W}_\ell\|_2\right)\|\delta\|_2.$$

On $\mathcal{E}$, this gives

$$\|\operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\delta) - \operatorname{Logit}_{\mathbf{W}}(\mathbf{x})\|_2 \le \left[\prod_{\ell=1}^{L}(R_\ell + t_\ell)\right]\|\delta\|_2.$$

Now let i be the predicted class at $\mathbf{X}$ and define

$$\Delta = \operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\delta) - \operatorname{Logit}_{\mathbf{W}}(\mathbf{x}).$$

For any $j \neq i$,

$$\begin{aligned}&\operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\delta)^{(i)} - \operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\delta)^{(j)}\\ &\ge m_{\mathbf{W}}(\mathbf{x}, i) - |\Delta^{(i)}| - |\Delta^{(j)}| \ge m_{\mathbf{W}}(\mathbf{x}, i) - 2\|\Delta\|_2.\end{aligned}$$

Hence, if

$$\|\delta\|_2 < \frac{m_{\mathbf{W}}(\mathbf{x}, i)}{2\prod_{\ell=1}^{L}(R_\ell + t_\ell)},$$

then

$$2\|\Delta\|_2 < m_{\mathbf{W}}(\mathbf{x}, i),$$

and consequently

$$\operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\delta)^{(i)} > \operatorname{Logit}_{\mathbf{W}}(\mathbf{x}+\delta)^{(j)} \qquad \text{for all } j \neq i.$$

Thus the predicted class is preserved on $\mathcal{E}$, which occurs with the stated probability. □

D Proof of Proposition 4.2

Lemma D.1 (Gaussian localization and relative entropy)**.** *Let $U : \mathbb{R}^{m\times n} \to \mathbb{R}$ be continuously differentiable, M-strongly convex and L-smooth in the Frobenius norm, where $0 < M \le L$, and put $d = mn$. Let $\pi(\mathrm{d}\mathbf{W}) = Z_U^{-1} e^{-U(\mathbf{W})}\,\mathrm{d}\mathbf{W}$ and let $\mathbf{W}^\star$ minimize U. Suppose that a random matrix $\widehat{\mathbf{W}}$ satisfies $\mathbb{E}\|\nabla U(\widehat{\mathbf{W}})\|_F^2 \le c < \infty$. For an independent standard Gaussian matrix $\mathbf{Z}$, set $\mathbf{Y} = \widehat{\mathbf{W}} + \sigma\mathbf{Z}$, $\sigma^2 > 0$, and $\widehat{\pi} = \mathcal{L}(\mathbf{Y})$. Then*

$$\mathrm{KL}(\widehat{\pi}\|\pi) \le \mathbb{E}\,\mathrm{KL}\big(\mathcal{N}(\widehat{\mathbf{W}}, \sigma^2\mathbf{I}_d)\|\pi\big) \le \frac{c}{2M} + \frac{d}{2}\{L\sigma^2 - 1 - \log(M\sigma^2)\}.$$

In particular, when $\sigma^2 = 1/L$, the last bound is $c/(2M) + (d/2)\log(L/M)$. For a fixed center $\widehat{\mathbf{W}}$, the intermediate bound before applying strong convexity is $U(\widehat{\mathbf{W}}) - U(\mathbf{W}^\star)$ plus the same variance term. The result also applies to U_g, with gradient Lipschitz constant L_g, target π_g, and minimizer $\mathbf{X}^\star$.

Proof. For every fixed center $\bar{\mathbf{W}}$, smoothness and the zero mean of $\mathbf{Z}$ give

$$\mathbb{E}[U(\bar{\mathbf{W}} + \sigma\mathbf{Z})] \le U(\bar{\mathbf{W}}) + \frac{Ld\sigma^2}{2}.$$

Strong convexity gives

$$Z_U \le e^{-U(\mathbf{W}^\star)}(2\pi/M)^{d/2}.$$

The Gaussian log-density has expectation $-(d/2)\log(2\pi e\sigma^2)$. Thus

$$\mathrm{KL}(\mathcal{N}(\bar{\mathbf{W}}, \sigma^2\mathbf{I}_d)\|\pi) \le U(\bar{\mathbf{W}}) - U(\mathbf{W}^\star) + \frac{d}{2}\{L\sigma^2 - 1 - \log(M\sigma^2)\}.$$

Minimizing the strong-convexity quadratic lower bound for U at $\bar{\mathbf{W}}$ yields $U(\bar{\mathbf{W}}) - U(\mathbf{W}^\star) \le \|\nabla U(\bar{\mathbf{W}})\|_F^2/(2M)$. Also strong monotonicity and Cauchy–Schwarz imply $\|\bar{\mathbf{W}} - \mathbf{W}^\star\|_F \le \|\nabla U(\bar{\mathbf{W}})\|_F/M$. Consequently $\widehat{\mathbf{W}}$ has finite second moment about $\mathbf{W}^\star$, and the preceding quadratic expectations are finite. To justify the first inequality, form joint laws of $(\widehat{\mathbf{W}}, \mathbf{Y})$ with the same first marginal $\mathcal{L}(\widehat{\mathbf{W}})$ and conditional second marginals $\mathcal{N}(\widehat{\mathbf{W}}, \sigma^2\mathbf{I}_d)$ and π. Their relative entropy is the expected conditional entropy. Projecting onto $\mathbf{Y}$ and using relative-entropy data processing gives the first inequality. Equivalently this follows from convexity of relative entropy in its first argument. Integrating the fixed-center estimate proves the claim. □

Proof of Proposition 4.2. Suppose first $\overline{D} > 0$. For $\mathbf{G}_t = \nabla U(\boldsymbol{\theta}_t)$ and $\mathbf{O}_t = \mathrm{Ortho}(\mathbf{G}_t)$, the compact SVD gives

$$\langle \mathbf{G}_t, \mathbf{O}_t\rangle_F = \|\mathbf{G}_t\|_*, \qquad \|\mathbf{O}_t\|_F^2 = \mathrm{rank}(\mathbf{G}_t) \le r.$$

Here $\mathrm{Ortho}(0) = 0$. Smoothness therefore gives

$$U(\boldsymbol{\theta}_t - \eta_{\mathrm{MUON}}\mathbf{O}_t) \le U(\boldsymbol{\theta}_t) - \eta_{\mathrm{MUON}}\|\mathbf{G}_t\|_* + \frac{Lr\eta_{\mathrm{MUON}}^2}{2}.$$

The algorithm forms $\boldsymbol{\theta}_0 = \mathbf{W}_0, \ldots, \boldsymbol{\theta}_{T-1}$. For the analysis only, define $\mathbf{V}_T = \boldsymbol{\theta}_{T-1} - \eta_{\mathrm{MUON}}\mathbf{O}_{T-1}$; this is not the injected Gaussian state and is not eligible for selection. Telescope the $T-1$ actual descent inequalities and this last virtual inequality, then use $D \le \overline{D}$ and $U(\mathbf{V}_T) \ge U(\mathbf{W}^\star)$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathbf{G}_t\|_* \le \frac{\overline{D}}{T\eta_{\mathrm{MUON}}} + \frac{Lr\eta_{\mathrm{MUON}}}{2} = \sqrt{\frac{2\overline{D}Lr}{T}}.$$

For the selected $t^* \in \{0, \ldots, T-1\}$,

$$\|\nabla U(\boldsymbol{\theta}_{t^*})\|_F^2 \le \|\nabla U(\boldsymbol{\theta}_{t^*})\|_*^2 \le \frac{2\overline{D}Lr}{T}.$$

Lemma D.1 with $c = 2\overline{D}Lr/T$ and $\sigma^2 = 1/L$ gives the KL bound. Pinsker's inequality gives the TV bound. Strong monotonicity gives $\|\boldsymbol{\theta}_{t^*} - \mathbf{W}^\star\|_F^2 \le \|\nabla U(\boldsymbol{\theta}_{t^*})\|_F^2/M^2$. Independence and centering of the injected Gaussian then give

$$\mathbb{E}\|\mathbf{Y}_0 - \mathbf{W}^\star\|_F^2 = \|\boldsymbol{\theta}_{t^*} - \mathbf{W}^\star\|_F^2 + d/L.$$

If $\overline{D} = 0$, strong convexity implies $\mathbf{W}_0 = \mathbf{W}^\star$. The fixed transient has zero gradient and the same conclusions follow directly from the lemma and Gaussian second moment. □

Remark D.1. The displayed upper bound retains the term $(d/2)\log\kappa$ in KL as T grows. It may be uninformative in TV when dimension or condition number is large. It does not establish acceleration relative to another initialization or sampler.

E WARM-START GEOMETRY AT A FINITE SAMPLING HORIZON

The following exact diagnostic concerns a fixed quadratic target; the spectral Moreau posterior is not generally quadratic.

Proposition E.1 (Quadratic warm-start)**.** *Let $\pi = \mathcal{N}(x^\star, H^{-1})$, where $0 \prec H \preceq LI$, and initialize ULA from $\mathcal{N}(c, \sigma^2 I)$ with deterministic c, $\sigma^2 > 0$, and $0 < \eta \le 1/L$. For integer $k \ge 0$, define $A = I - \eta H$ and $C_k = \sigma^2 A^{2k} + 2\eta\sum_{j=0}^{k-1} A^{2j}$ (empty sum zero). Then*

$$\begin{aligned}\mathrm{KL}(\mathcal{L}(Y_k)\|\pi) = &\underbrace{\tfrac{1}{2}(c-x^\star)^\top H A^{2k}(c-x^\star)}_{J_k(c)} \\ &+ \underbrace{\tfrac{1}{2}\{\mathrm{tr}(HC_k) - d - \log\det(HC_k)\}}_{B_k}.\end{aligned} \tag{5}$$

Thus matched deterministic centers differ only through J_k.

Proof. Iterating the affine Gaussian recursion gives $Y_k \sim \mathcal{N}(x^\star + A^k(c - x^\star), C_k)$ with $C_k \succ 0$. The Gaussian KL formula and $AH = HA$ give the identity. □

In an orthonormal Hessian eigenbasis $Hv_i = \lambda_i v_i$,

$$J_k(c) = \frac{1}{2}\sum_{i=1}^{d} \lambda_i(1-\eta\lambda_i)^{2k}\langle c - x^\star, v_i\rangle^2. \tag{6}$$

The initial objective gap J_0 therefore need not rank centers as J_k does; moving the center cannot change the shared covariance term B_k. For random centers the identity holds conditionally, not generally for the unconditional Gaussian mixture.

For $W \in \mathbb{R}^{2\times 2}$, take

$$U(W) = \tfrac{1}{2}W_{11}^2 + 50(W_{12}^2 + W_{21}^2 + W_{22}^2), \qquad W_0 = I_2.$$

Here $M = 1$, $L = 100$, $r = 2$, and $D = 50.5$. Proposition 4.2's $T = 2$, $\overline{D} = D$ rule gives $W_{\mathrm{M}} = bI_2$, where $b = 1 - \sqrt{101/400}$; its nuclear gradient $101b < 101$ ensures selection. One $1/L$ GD step gives $W_{\mathrm{G}} = \mathrm{diag}(0.99, 0)$. With matched injection variance $1/100$ and ULA stepsize $1/200$, Equation 6 gives

	W_{M}	W_{G}
J_0	12.4994	0.49005
J_3	0.31346	0.47553

Values are rounded. Since B_3 is shared, MUON has lower endpoint KL despite higher initial loss: it reduces the slow residual more, while ULA rapidly removes its larger high-curvature residual. Conversely, for $U(W) = 50\|W\|_F^2$ with the same W_0, T and $\overline{D} = D = 100$, the prescribed MUON center is $(1 - 1/\sqrt{2})I_2$ and GD reaches zero; the KL difference is $100(1 - 1/\sqrt{2})^2\, 4^{-k} > 0$ under the same injection and sampling settings. These compare the specified update, not equal computation or optimally tuned baselines.

F Convergence Analysis

The finite-time sampling error for PMLD has two components:

- $\|\mu_{k\eta} - \widetilde{\pi}_\gamma\|_{\mathrm{TV}}$ (mixing error): the distance between the distribution $\mu_{k\eta}$ after finite-time evolution of the continuous Langevin dynamics and the target distribution $\widetilde{\pi}_\gamma$;
- $\|\nu_{k\eta} - \mu_{k\eta}\|_{\mathrm{TV}}$ (discretization error): the distance between the distribution $\nu_{k\eta}$ of the discretized Langevin dynamics and that $\mu_{k\eta}$ of the continuous Langevin dynamics.

For APMLD targeting π, reset time at the independent Gaussian draw $\mathbf{Y}_0 \sim \widehat{\pi}$. The warm start enters through the initial relative entropy and second moment in these same two error terms; it need not be added as a separate nondecaying TV error.

Proof of Proposition 4.3. Both laws start from ρ_0, independently of future Brownian increments. Apply Lemma F.1 to U_g and then the triangle inequality. Note that identical stepsizes are needed for the explicit discretization estimate, not for this triangle inequality. □

F.1 Full Proof for PMLD and APMLD

We first present two useful lemmas used to prove Theorem F.1.

Lemma F.1 (Entropy mixing for the chosen target)**.** *Let $U_g : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, M-strongly convex and have globally L_g-Lipschitz gradient, where $0 < M \le L_g < \infty$. Let $\pi_g \propto e^{-U_g}$ and let μ_t be the law of the continuous Langevin dynamics $d\mathbf{W}_t = -\nabla U_g(\mathbf{W}_t)dt + \sqrt{2}\, d\mathbf{B}_t$, with initial law ρ_0. If $K = \mathrm{KL}(\rho_0\|\pi_g) < \infty$, then*

$$\mathrm{KL}(\mu_t\|\pi_g) \le e^{-2Mt}K, \qquad \|\mu_t - \pi_g\|_{\mathrm{TV}} \le \sqrt{K/2}\, e^{-Mt} \quad (t \ge 0).$$

Proof. Since U_g is M-strongly convex, $\pi_g \propto e^{-U_g}$ satisfies the logarithmic Sobolev inequality

$$\mathrm{Ent}_{\pi_g}(f) \le \frac{1}{2M} \int \frac{\|\nabla f\|_F^2}{f}\, d\pi_g$$

for normalized positive densities f with respect to π_g; for smooth potentials this follows from the Bakry–Émery criterion (see, e.g., Gentil, 2010, Theorem 3.3). The extension to the space of continuously differentiable functions $C^{1,1}$ setting assumed here, together with the regularity needed for the entropy-dissipation argument below, is justified after the proof.

Let $(P_t)_{t\ge 0}$ be the Markov semigroup associated with the Langevin diffusion,

$$P_t\varphi(w) = \mathbb{E}[\varphi(\mathbf{W}_t) \mid \mathbf{W}_0 = w]\,.$$

Write

$$f_0 = \frac{d\rho_0}{d\pi_g}, \qquad f_t = \frac{d\mu_t}{d\pi_g}.$$

Since π_g is invariant and the Langevin diffusion is reversible with respect to π_g, we have $f_t = P_t f_0$. Moreover,

$$\mathrm{KL}(\mu_t\|\pi_g) = \mathrm{Ent}_{\pi_g}(f_t).$$

The entropy-dissipation identity for the Langevin semigroup gives

$$\frac{d}{dt}\,\mathrm{Ent}_{\pi_g}(f_t) = -\int \frac{\|\nabla f_t\|_F^2}{f_t}\, d\pi_g.$$

Combining this identity with the logarithmic Sobolev inequality yields

$$\frac{d}{dt}\,\mathrm{Ent}_{\pi_g}(f_t) \le -2M\,\mathrm{Ent}_{\pi_g}(f_t).$$

Hence, by Grönwall's inequality,

$$\mathrm{KL}(\mu_t\|\pi_g) \le e^{-2Mt}\,\mathrm{KL}(\rho_0\|\pi_g) = e^{-2Mt}K.$$

Finally, Pinsker's inequality gives

$$\|\mu_t - \pi_g\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\,\mathrm{KL}(\mu_t\|\pi_g)} \le \sqrt{\frac{K}{2}}\,e^{-Mt}.$$

The extension from regular initial densities to arbitrary ρ_0 with finite relative entropy follows from the approximation argument given below. □

Lemma F.2 (Forward relative entropy of Euler discretization)**.** *Let $U_g : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, M-strongly convex and L_g-smooth, $0 < M \le L_g$. Let $\mathbf{X}^\star$ minimize U_g. Initialize the diffusion $d\mathbf{W}_t = -\nabla U_g(\mathbf{W}_t)dt + \sqrt{2}\,d\mathbf{B}_t$ and explicit Euler chain*

$$\mathbf{Y}_{j+1} = \mathbf{Y}_j - \eta\nabla U_g(\mathbf{Y}_j) + \sqrt{2}(\mathbf{B}_{(j+1)\eta} - \mathbf{B}_{j\eta})$$

with the same law ρ_0, independent of future Brownian increments. Suppose $E_0 = \mathbb{E}\|\mathbf{Y}_0 - \mathbf{X}^\star\|_F^2 < \infty$, $0 < \eta \le 2/(M + L_g)$, and $k \ge 1$. Let $\mu_{k\eta} = \mathcal{L}(\mathbf{W}_{k\eta})$ and $\nu_{k\eta} = \mathcal{L}(\mathbf{Y}_k)$. With $\kappa_g = L_g/M$,

$$\mathrm{KL}(\nu_{k\eta}\|\mu_{k\eta}) \le \frac{(1+\kappa_g)L_g^3\eta^2}{24}(E_0 + 2d\eta k) + \frac{dL_g^2\eta^2 k}{4}.$$

For the spectral target one may take $g = \iota_{\mathcal{B}_R}$, so $U_g = \widetilde{U}_\gamma$, $L_g = L + \gamma^{-1}$, and $d = mn$.

Proof. Let $\bar{\mathbf{Y}}_t = \mathbf{Y}_j - (t - j\eta)\nabla U_g(\mathbf{Y}_j) + \sqrt{2}(\mathbf{B}_t - \mathbf{B}_{j\eta})$ for $j\eta \le t < (j+1)\eta$. Write Q for its path law and P for the diffusion path law on $[0, k\eta]$. Both have the same initial marginal. Global Lipschitz drift gives nonexplosion and finite second moments on every finite time interval. Girsanov's formula with localization, followed by data processing at time $k\eta$, gives

$$\mathrm{KL}(\nu_{k\eta}\|\mu_{k\eta}) \le \mathrm{KL}(Q\|P) \le \frac{1}{4}\sum_{j=0}^{k-1}\int_{j\eta}^{(j+1)\eta}\mathbb{E}_Q\|\nabla U_g(\bar{\mathbf{Y}}_t) - \nabla U_g(\mathbf{Y}_j)\|_F^2 dt. \tag{*}$$

Here is a localization justification of the second inequality. Stop the canonical paths when their norm first reaches N, and stop both drifts and the common diffusion there. Before stopping, the two predictable drifts and their difference are bounded, so the exponential martingale has unit expectation and the stopped path entropy is one quarter of the stopped drift energy under the Euler law. Both stopped path laws converge to the respective unstopped laws by nonexplosion as $N \to \infty$. Lower semicontinuity of relative entropy and monotone convergence of the nonnegative stopped energies give (*). Paths initially outside the stopping ball are stopped at zero; their common initial distribution contributes no relative entropy.

Put $a_j = \mathbb{E}\|\mathbf{Y}_j - \mathbf{X}^\star\|_F^2$. For $u = t - j\eta$, conditioning on the past at $j\eta$ gives

$$\mathbb{E}\|\bar{\mathbf{Y}}_t - \mathbf{Y}_j\|_F^2 = u^2\mathbb{E}\|\nabla U_g(\mathbf{Y}_j)\|_F^2 + 2du.$$

Indeed the future Brownian increment has conditional mean zero and second moment du; all cross terms are integrable by finite second moments and linear growth of ∇U_g. Using smoothness in (*) and integrating yields

$$\mathrm{KL}(\nu_{k\eta}\|\mu_{k\eta}) \le \frac{L_g^2\eta^3}{12}\sum_{j=0}^{k-1}\mathbb{E}\|\nabla U_g(\mathbf{Y}_j)\|_F^2 + \frac{dL_g^2\eta^2 k}{4}. \tag{**}$$

Tonelli justifies the nonnegative expectation/time integrals.

Strong convexity and smoothness imply

$$\langle y - \mathbf{X}^\star, \nabla U_g(y)\rangle \ge \frac{ML_g}{M+L_g}\|y - \mathbf{X}^\star\|_F^2 + \frac{1}{M+L_g}\|\nabla U_g(y)\|_F^2.$$

Hence, with $\rho = 1 - 2\eta ML_g/(M+L_g)$, the Euler moment identity and $\eta \le 2/(M+L_g)$ give

$$a_{j+1} \le \rho a_j + 2d\eta, \qquad 0 \le \rho < 1.$$

Induction, including $j = 0$, gives

$$a_j \le \rho^j E_0 + 2d\eta\sum_{i=0}^{j-1}\rho^i.$$

The empty sum at $j = 0$ is zero. Therefore

$$\sum_{j=0}^{k-1} \mathbb{E}\|\nabla U_g(\mathbf{Y}_j)\|_F^2 \le L_g^2 \sum_{j=0}^{k-1} a_j \le \frac{L_g^2}{1-\rho}(E_0 + 2d\eta k).$$

Substitution of $1 - \rho = 2\eta M L_g/(M + L_g)$ into (**) proves the displayed bound. □

The interpolation inequality used above follows by applying co-coercivity to $U_g - M\|\cdot\|_F^2/2$, which is convex and $(L_g - M)$-smooth when $L_g > M$, and expanding the result. When $L_g = M$, gradient differences equal M times the input difference, and the inequality holds with equality. For $g = \iota_{\mathcal{B}_R}$ the envelope is convex with γ^{-1}-Lipschitz gradient, so $U_g = \widetilde{U}_\gamma$ is M-strongly convex and L_γ-smooth, with $L_\gamma = L + \gamma^{-1}$ and $\kappa_\gamma = L_\gamma/M$.

Theorem F.1 (Combined error, PMLD and the chosen smooth target)**.** *Let U_g, $\mathbf{X}^\star$, $\mu_{k\eta}$ and $\nu_{k\eta}$ satisfy Lemma F.2, with common initial law $\rho_0 = \nu_0 = \mu_0$. Assume $\mathrm{KL}(\nu_0\|\pi_g) < \infty$ and put $E_0 = \mathbb{E}_{\nu_0}\|\mathbf{Y}_0 - \mathbf{X}^\star\|_F^2 < \infty$. For $0 < \eta \le 2/(M + L_g)$ and $k \ge 1$,*

$$\|\nu_{k\eta} - \pi_g\|_{\mathrm{TV}} \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}(\nu_0\|\pi_g)}e^{-M\eta k} + \sqrt{\frac{(1+\kappa_g)L_g^3\eta^2}{48}(E_0 + 2d\eta k) + \frac{dL_g^2\eta^2 k}{8}}.$$

For PMLD, $\pi_g = \widetilde{\pi}_\gamma$, $L_g = L_\gamma = L + \gamma^{-1}$, and $\kappa_g = \kappa_\gamma$. For APMLD with $g = 0$, $\pi_g = \pi$, $L_g = L$ and $\mathbf{X}^\star = \mathbf{W}^\star$.

Let $0 < \varepsilon \le 1/\mathrm{e}$, $\ell = \log\max\{d, 1/\varepsilon\}$, and assume

$$\mathrm{KL}(\nu_0\|\pi_g) \le K d^p, \qquad E_0 \le C_0 d^2,$$

where $K \ge 1$, $p \ge 0$, $C_0 \ge 0$ are independent of (d, ε). Set $q = p + 6 + \log K$ and

$$c = \min\left\{\frac{M}{9qL_g^2}, \sqrt{\frac{16M}{3(1+\kappa_g)L_g^3(C_0M + 2q)}}\right\},$$
$$\eta = \frac{c\varepsilon^2}{d\ell}, \qquad k = \left\lceil \frac{q\ell}{2M\eta} \right\rceil.$$

Then $\eta < 1/L_g \le 2/(M + L_g)$, $\|\nu_{k\eta} - \pi_g\|_{\mathrm{TV}} \le \varepsilon$, and

$$k \le (9 + \sqrt{C_0 M})\kappa_g^2(p + 6 + \log K)^2 \frac{d}{\varepsilon^2}\log^2\max\{d, 1/\varepsilon\}.$$

Thus the original $\mathcal{O}(d\varepsilon^{-2}\log^2\max\{d, 1/\varepsilon\})$ rate holds for fixed M, L_g, K, p, C_0. In the PMLD case these constants may depend on γ; this is not a uniform statement as $\gamma \downarrow 0$.

Proof. Pinsker applied to the forward KL bound of Lemma F.2, followed by the triangle inequality and Lemma F.1, gives the displayed TV bound. For the complexity statement, $\ell \ge 1$, $q \ge 6$, and $cM \le 1/(9q)$. Hence $x = q\ell/(2M\eta) > 1$ and $x \le k \le 2x = q\ell/(M\eta)$; also $\eta \le c < 1/L_g \le 2/(M + L_g)$. The mixing term is bounded by

$$\sqrt{Kd^p/2}e^{-M\eta k} \le 2^{-1/2}e^{-3\ell} \le 2^{-1/2}\varepsilon^3 \le \varepsilon/3,$$

using $d^p \le e^{p\ell}$, $K \ge 1$ and $\varepsilon \le 1/\mathrm{e}$. The two squared discretization terms satisfy

$$\frac{dL_g^2\eta^2 k}{8} \le \frac{qL_g^2 c}{8M}\varepsilon^2 \le \varepsilon^2/72,$$
$$\frac{(1+\kappa_g)L_g^3\eta^2}{48}(E_0 + 2d\eta k) \le \frac{(1+\kappa_g)L_g^3 c^2}{48M}(C_0M + 2q)\varepsilon^4 \le \varepsilon^4/9.$$

Thus the total is at most $\varepsilon/3 + \varepsilon/\sqrt{8} < \varepsilon$. Finally, $k \le qd\ell^2/(cM\varepsilon^2)$. The two branches of c respectively give

$$\frac{q}{cM} = 9\kappa_g^2 q^2, \qquad \frac{q}{cM} \le \frac{\sqrt{6}}{4}\kappa_g^2 q\sqrt{C_0M + 2q} \le (9 + \sqrt{C_0M})\kappa_g^2 q^2,$$

where the last inequality uses $q \ge 6$ and $\kappa_g \ge 1$. This proves the stated explicit bound in both branches. □

Remark F.1 (Target and deterministic initialization). At fixed finite γ, a guarantee relative to the original π requires the additional bias $\|\widetilde{\pi}_\gamma - \pi\|_{\mathrm{TV}}$; longer simulation does not remove that change of target. A deterministic start has infinite initial KL to a continuous target. For such a start, a sufficient initialization procedure is one Euler step of size $\eta_0 = 1/L_g$, followed by the theorem's sampling step η. Write $\mathbf{Z} = \mathbf{W}_0 - \nabla U_g(\mathbf{W}_0)/L_g$ and $\nu_1 = \mathcal{N}(\mathbf{Z}, 2L_g^{-1}\mathbf{I}_d)$. The fixed-center bound of Lemma D.1 and descent give

$$\mathrm{KL}(\nu_1\|\pi_g) \le \frac{L_g}{2}\|\mathbf{W}_0 - \mathbf{X}^\star\|_F^2 + \frac{d}{2}[1 + \log^+(\kappa_g/2)],$$
$$\mathbb{E}_{\nu_1}\|\mathbf{Y}_1 - \mathbf{X}^\star\|_F^2 \le \|\mathbf{W}_0 - \mathbf{X}^\star\|_F^2 + 2d/L_g.$$

Here $\log^+ z = \max\{0, \log z\}$ only enlarges the entropy bound. The second inequality follows by expanding the squared gradient-step distance and using the interpolation inequality in Lemma F.2, with $1/L_g \le 2/(M + L_g)$. If $\|\mathbf{W}_0 - \mathbf{X}^\star\|_F^2 \le B_0 d^2$, take

$$K = \max\{1, L_g B_0/2 + [1 + \log^+(\kappa_g/2)]/2\}, \quad p = 2, \quad C_0 = B_0 + 2/L_g.$$

Apply the theorem after resetting the clock at ν_1. This procedure uses $k + 1$ total steps and is an explicitly stated initialization variant, rather than a claim of finite KL at a point mass.

Corollary F.1 (Combined error, APMLD)**.** *Use Algorithm 2 with $g = 0$, the nuclear-norm selection rule and all hypotheses of Proposition 4.2. Suppose $0 < M \le L$ and the integer $T \ge 1$ are fixed, and that a known bound $\overline{D} \ge D$ satisfies $\overline{D}r \le C_D d^2$ for a fixed $C_D \ge 0$. For $0 < \varepsilon \le 1/\mathrm{e}$, choose the post-switch η, k from Theorem F.1 with*

$$L_g = L, \quad \kappa_g = \kappa = L/M, \quad K = \max\{1, \kappa C_D/T + \tfrac{1}{2}\log\kappa\}, \quad p = 2,$$
$$C_0 = 2\kappa C_D/(MT) + 1/L.$$

Then $\|\nu_{k\eta} - \pi\|_{\mathrm{TV}} \le \varepsilon$ and

$$k \le (9 + \sqrt{C_0 M})\kappa^2(8 + \log K)^2 \frac{d}{\varepsilon^2}\log^2\max\{d, 1/\varepsilon\}.$$

The hidden constant in the resulting $\mathcal{O}(d\varepsilon^{-2}\log^2\max\{d, 1/\varepsilon\})$ bound may depend on M, L, T, C_D. These are post-switch steps, in addition to $T - 1$ MUON updates, T gradient evaluations and T nuclear-norm evaluations for selection. The last polar update is virtual and need not be computed.

Proof. For $\mathbf{Y}_0 \sim \widehat{\pi}$, Proposition 4.2 gives

$$\mathrm{KL}(\widehat{\pi}\|\pi) \le \kappa\overline{D}r/T + (d/2)\log\kappa \le Kd^2,$$
$$\mathbb{E}\|\mathbf{Y}_0 - \mathbf{W}^\star\|_F^2 \le 2\overline{D}Lr/(M^2T) + d/L \le C_0 d^2.$$

Apply Theorem F.1 with $g = 0$ and reset the sampling clock at the injected state. The MUON step η_{MUON} is not used as the sampling step. □

Remark F.2. A fixed finite MUON phase changes the initialization constants in this bound, but no improvement in the dimension/accuracy order is proved. The sufficient growth hypothesis is substantive; for example, $\overline{D} \le C_D d^{3/2}$ implies it since $r \le \sqrt{d}$. It does not follow from $T = \mathcal{O}(1)$ alone.

G Supplementary Experimental Details

Input and feature dimensions in this appendix are local to the experimental definitions and are distinct from the parameter dimension $d = mn$ used in the sampling theory.

G.1 Bayesian MLP Classifiers

Let K denote the number of classes and let

$$p_i(c; W) = p(y_i = c \mid x_i, W), \qquad c \in \{1, \dots, K\},$$

be the predictive probability assigned by network parameter W. Given retained posterior samples $W^{(1)}, \ldots, W^{(S)}$, we use the posterior-predictive probability

$$\bar{p}_i(c) = \frac{1}{S} \sum_{s=1}^{S} p_i(c; W^{(s)}).$$

The posterior-predictive classifier is

$$\hat{y}_i = \arg \max_{c \in \{1,\ldots,K\}} \bar{p}_i(c),$$

and its confidence is

$$\hat{q}_i = \max_{c \in \{1,\ldots,K\}} \bar{p}_i(c).$$

The posterior-predictive negative log-likelihood is

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log \bar{p}_i(y_i).$$

We report classification error as

$$\mathrm{Err} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{y}_i \neq y_i\},$$

equivalently,

$$\mathrm{Err} = 1 - \mathrm{Acc}, \qquad \mathrm{Acc} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{y}_i = y_i\}.$$

For calibration, we use expected calibration error (ECE). Let

$$0 = a_0 < a_1 < \cdots < a_B = 1$$

be a partition of the confidence interval $[0, 1]$, and define the confidence bin

$$I_b = \{i : \hat{q}_i \in (a_{b-1}, a_b]\}, \qquad b = 1, \ldots, B.$$

The empirical accuracy and confidence within bin I_b are

$$\mathrm{acc}(I_b) = \frac{1}{|I_b|} \sum_{i \in I_b} \mathbf{1}\{\hat{y}_i = y_i\}, \qquad \mathrm{conf}(I_b) = \frac{1}{|I_b|} \sum_{i \in I_b} \hat{q}_i,$$

with empty bins omitted. The ECE is

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{n} \left|\mathrm{acc}(I_b) - \mathrm{conf}(I_b)\right|.$$

For the adversarial Fashion-MNIST experiment, we evaluate robustness under ℓ_∞-bounded fast gradient sign method (FGSM) perturbations. Given retained posterior samples $\{W^{(s)}\}_{s=1}^{S}$, define the posterior-predictive class probability as

$$\bar{p}(y \mid x) = \frac{1}{S} \sum_{s=1}^{S} p(y \mid x; W^{(s)}).$$

We construct attacks using the posterior-predictive negative log-likelihood,

$$\mathcal{L}_{\mathrm{attack}}(x, y) = -\log \bar{p}(y \mid x).$$

Thus, probabilities are averaged before taking the negative logarithm. Given perturbation radius ϵ, FGSM forms

$$x_{\mathrm{adv}} = \Pi_{[0,1]^d} \left(x + \epsilon \ \mathrm{sign}\left(\nabla_x \mathcal{L}_{\mathrm{attack}}(x, y)\right)\right),$$

where x and ϵ are expressed in raw pixel coordinates and $\Pi_{[0,1]^d}$ clips the perturbed image to the valid pixel range. Consequently,

$$\|x_{\mathrm{adv}} - x\|_\infty \leq \epsilon, \qquad x_{\mathrm{adv}} \in [0, 1]^d.$$

After constructing x_{adv}, we recompute the posterior-predictive probabilities

$$\bar{p}_{\mathrm{adv},i}(c) = \frac{1}{S} \sum_{s=1}^{S} p(y_i = c \mid x_{\mathrm{adv},i}, W^{(s)}),$$

and evaluate NLL, classification error, and ECE using $\bar{p}_{\mathrm{adv},i}$ in place of $\bar{p}_i$.

G.2 DDPM

For the DDPM experiments, we evaluate generated samples using Fréchet inception distance (FID), kernel inception distance (KID), diversity, and lag-1 autocorrelation. Let

$$x_1, \dots, x_n \sim p_{\text{data}}$$

denote real images and

$$\tilde{x}_1, \dots, \tilde{x}_m \sim p_{\text{gen}}$$

denote generated images. Let $\phi(x) \in \mathbb{R}^{2048}$ denote the Inception-v3 feature representation used for FID and KID. For CIFAR-10, we use the `torchvision` implementation with `Inception_V3_Weights.DEFAULT`, resizing images to 299×299 and applying ImageNet normalization. For CelebA-HQ, FID and KID use the Clean-FID implementation in `clean` mode with its corresponding feature extraction and preprocessing. For both datasets, diversity and the lag-1 diagnostic use the torchvision feature representation, denoted below by ψ.

The Fréchet inception distance (FID) compares Gaussian approximations to the feature distributions of real and generated images. Let

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i), \qquad \mu_g = \frac{1}{m}\sum_{j=1}^{m} \phi(\tilde{x}_j),$$

and let Σ_r and Σ_g denote the corresponding empirical covariance matrices. The FID is

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2}\right)^{1/2}\right).$$

Smaller FID indicates greater similarity between the real and generated feature distributions.

Kernel inception distance (KID) measures the squared maximum mean discrepancy between the real and generated feature distributions. Given a positive-definite kernel k, its unbiased empirical estimator is

$$\begin{aligned}\text{KID} = &\frac{1}{n(n-1)}\sum_{i\neq i'} k(\phi(x_i), \phi(x_{i'})) + \frac{1}{m(m-1)}\sum_{j\neq j'} k(\phi(\tilde{x}_j), \phi(\tilde{x}_{j'})) \\ &- \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(\phi(x_i), \phi(\tilde{x}_j)).\end{aligned}$$

We use the standard polynomial kernel

$$k(u, v) = \left(\frac{1}{d}u^\top v + 1\right)^3.$$

Reported KID values average this estimator over randomly selected real and generated feature subsets. As with FID, smaller KID indicates closer agreement between the real and generated feature distributions.

To quantify variability across retained posterior weight samples, we use a fixed-noise image bank. Let $\tilde{x}_{s,j}$ denote the image generated using posterior weight sample s and noise index j, with noise inputs matched across weight samples, and define $f_{s,j} = \psi(\tilde{x}_{s,j}) \in \mathbb{R}^d$. We use $S = 10$ retained weight samples for CIFAR-10 and $S = 5$ for CelebA-HQ, with $J = 8$ noise inputs in both cases. Diversity is the average pairwise root-mean-square feature distance across weight samples at matched noise:

$$\text{Diversity} = \frac{2}{S(S-1)J}\sum_{1\le a<b\le S}\sum_{j=1}^{J}\frac{\|f_{a,j} - f_{b,j}\|_2}{\sqrt{d}}.$$

Larger values indicate greater feature variation across posterior weight samples under matched noise. This diagnostic is computed from a separate fixed-noise bank, rather than all images used for FID and KID.

Finally, we report a signed lag-1 feature-trajectory correlation diagnostic. For each retained weight sample, average features over the fixed noise inputs:

$$z_s = \frac{1}{J}\sum_{j=1}^{J} f_{s,j}, \qquad s = 1, \dots, S,$$

where weight samples are ordered by retention time. For coordinate k, define the means of the two shifted sequences as

$$\bar{z}_k^- = \frac{1}{S-1}\sum_{s=1}^{S-1} z_{s,k}, \qquad \bar{z}_k^+ = \frac{1}{S-1}\sum_{s=1}^{S-1} z_{s+1,k},$$

and let

$$D_k = \sqrt{\left[\sum_{s=1}^{S-1}(z_{s,k} - \bar{z}_k^-)^2\right]\left[\sum_{s=1}^{S-1}(z_{s+1,k} - \bar{z}_k^+)^2\right]}.$$

The coordinatewise correlation is

$$\rho_{1,k} = \frac{\sum_{s=1}^{S-1}(z_{s,k} - \bar{z}_k^-)(z_{s+1,k} - \bar{z}_k^+)}{D_k}.$$

We average the signed correlations over valid coordinates:

$$\text{Lag-1} = \frac{1}{|\mathcal{V}|}\sum_{k\in\mathcal{V}} \rho_{1,k}, \qquad \mathcal{V} = \{k : D_k > 10^{-12}\}.$$

The diagnostic is undefined if $\mathcal{V}$ is empty. Because correlations are signed and averaged across coordinates, a smaller value does not by itself establish weaker dependence or better mixing.

We additionally report quality-adjusted diversity as

$$\text{QAD} = \frac{\text{Diversity}}{1 + \text{FID}},$$

using the corresponding dataset and evaluation protocol's FID.