DON'T STOP ME *Yet*: SAMPLING LOSS MINIMA VIA DISSIPATIVE RIEMANNIAN MECHANICS

Anonymous authors
Paper under double-blind review

ABSTRACT

The minima of modern neural network loss functions are typically not isolated, rather they form connected components of reparameterization invariant solutions on the training data. Analytically characterizing these solutions is a hard problem, but sampling approaches are feasible. By construction, existing methods either spread over low-loss regions, and thus do not sample reparameterization invariant solutions exactly, or are inherently local, which limits exploration of other minima valleys. We propose sampling such reparameterization invariant models using a dynamical system based on kinetic energy, subject to a gravitational pull and a friction term that dissipates energy from the system. Our proposed sampler, DIMS, is guaranteed to sample exactly from the minimum level sets and depends on physically motivated hyperparameters which allows control over the exploration capabilities of the sampler. We consider out-of-distribution detection through uncertainty quantification as the motivating problem and observe improved performance compared to previously proposed approaches.

1 INTRODUCTION

Training a deep neural network results in finding a single point estimate with minimum loss within a vast, high-dimensional parameter space. Yet the loss landscape has a rich geometric structure and a single point estimate can only capture local information about it, which can have important shortcomings for downstream tasks. The architecture and the data together shape the loss landscape such that several parameter configurations perform equally well on the training data; the minimum level sets of modern overparameterized networks form extended, continuous submanifolds for which the model is *reparameterization invariant* on the observed data (Dinh et al., 2017; Li et al., 2018; Garipov et al., 2018; Draxler et al., 2018).

In this paper we ask *"how should we navigate the space of good solutions"* rather than *"where is a good solution"*. We propose the **Di**ssipative **M**inima **S**ampler (DIMS), formulated as a dynamical system that navigates the loss landscape into low-loss regions while guaranteeing convergence to the submanifold of minimum solutions. As a result, the gathered samples correspond to a set of functions that agree on the training data but vary in their predictions elsewhere.

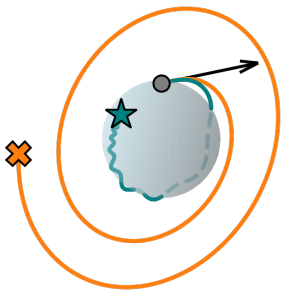


Figure 1: Two trajectories through parameter space via different dynamical systems. The sphere constitutes a submanifold of parameters that minimize $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{K} \theta_i^2 - 1$. While a geodesic on the loss surface (—) is the straightest path, it never converges to a minimum solution, even when starting from one. Our modified dynamics (—) ensure energy dissipation which leads the particle to stop at a minimum solution.

Recent works have focused on sampling low-loss regions by respecting the inherent geometry of the problem (Bergamin et al., 2023; Yu et al., 2024; Fadel et al., 2025), and even restricting samples to the reparameterization invariant set (Roy et al., 2024). Our proposed method differs by guaranteeing convergence to the minimum manifold while allowing for escaping the initial loss basin to find other connected minima – a combination that neither optimization-based methods nor existing sampling-based methods are designed for.

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.

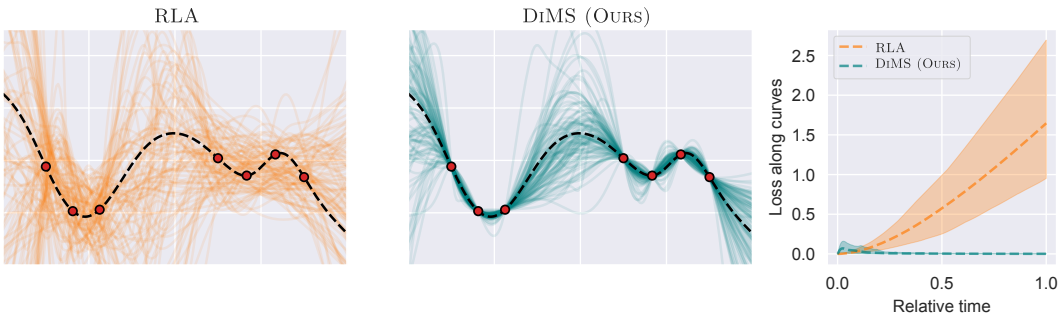


Figure 2: We fit a neural network to $N = 7$ training points and show function-space samples from two geometry-aware sampling schemes formulated as continuous dynamical systems. The Riemannian Laplace approximation (*left*) explores low-loss regions but does not interpolate the training data. Our sampler DIMS (*center*) leverages an alternative formulation of the system and guarantees convergence to minimum training loss solutions, ensuring all function-space samples to interpolate the training data, and remaining in low-loss regions along paths generated by the dynamics (*right*).

Our contributions include:

1. **A deterministic sampler for the minimum submanifolds**, continuously visiting distinct points on the minimum submanifolds of the training loss. We formalize geometry-driven dynamics on Riemannian manifolds induced by the loss with physically interpretable hyperparameters controlling the exploration behavior of the sampler. The resulting samples can be considered as coming from an implicit approximate Bayesian posterior concentrated on the minimum manifold.
2. **Theoretical guarantees** that our dynamics explore the loss surface before converging to the minimum manifold in finite time.
3. **Empirical validation** that posterior approximations concentrated on the minimum manifold generalize well, and can improve predictive performance, calibration, and especially out-of-distribution (OOD) detection through uncertainty quantification.

Intuition: The sampling process (Figure 1) starts by shooting out a particle from a fixed point $\boldsymbol{\theta}^*$ that minimizes the loss. The direction and initial speed of the particle is randomly sampled from a distribution, $\boldsymbol{v} \sim q(\boldsymbol{v})$, and determines the initial kinetic energy of the particle. By introducing external forces well-known from physics, we guide the particle toward minima, eventually stopping at a minimum level set. Every intersection with this set corresponds to a model that optimally fits the training data, yet exhibits different behavior in other data regions. This leads to high functional variation in regions away from observed data (Figure 2), which is desired for OOD detection based on uncertainty quantification. DIMS applies to any choice of smooth loss function which naturally includes posterior sampling.

2 BACKGROUND

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ denote a dataset with N paired input-output observations of $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^D$ and $y_i \in \mathcal{Y}$ where $\mathcal{Y} \subseteq \mathbb{R}$ for regression or $\mathcal{Y} = \{1, \ldots, C\}$ for classification among C classes. We consider a neural network $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ with parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^K$, trained by minimizing a loss objective $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$. The geometric framework we develop throughout the paper applies to any smooth loss; a common choice is mean reduction, i.e., $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^N \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$, over some per-sample loss $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. This loss can be used when sampling with DIMS.

Since several uncertainty quantification methods are motivated from Bayesian inference, we will often consider the natural choice of loss given by the negative log-joint,

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) := -\log p(\mathcal{D}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) + \log p(\mathcal{D}) = -\log p(\boldsymbol{\theta}) - \sum_{i=1}^N \log p\left(y_i | f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right) + C, \quad (1)$$

which equals the negative log-posterior up to a constant. Since the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is intractable for neural networks, existing methods consider an approximate posterior $q(\boldsymbol{\theta})$ from which sampling is easy, while other sampling algorithms require finer tuning, especially in high dimensions.

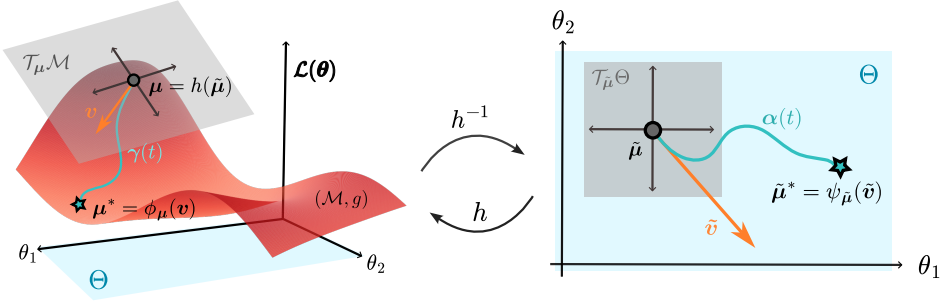


Figure 3: The loss surface viewed as a Riemannian manifold $(\mathcal{M}, g)$ embedded in $\mathbb{R}^{K+1}$. The parameter space inherits the geometry of the loss surface via the pullback metric $\mathbf{G}(\boldsymbol{\theta})$. A curve on the manifold $\boldsymbol{\gamma}_{\boldsymbol{v}}(t) \in \mathcal{M}$ starting at $\boldsymbol{\mu}$ is generated by the initial velocity $\boldsymbol{v}$ and can be expressed through its corresponding path in the parameter space $\boldsymbol{\alpha}_{\tilde{\boldsymbol{v}}}(t) \in \Theta$ induced by $\tilde{\boldsymbol{v}}$. We denote the mapping from an initial velocity to a curve at time t by $\phi_{\boldsymbol{\mu},t} : \mathcal{T}_{\boldsymbol{\mu}}\mathcal{M} \to \mathcal{M}$ and $\psi_{\tilde{\boldsymbol{\mu}},t} : \mathcal{T}_{\tilde{\boldsymbol{\mu}}}\Theta \to \Theta$.

2.1 Defining the loss surface as a Riemannian manifold

We briefly review Riemannian geometry of loss surfaces, which forms the foundation of our approach. For notational convenience we suppress dependency on the data $\mathcal{D}$ unless explicitly needed. Given a smooth loss function $\mathcal{L} : \mathbb{R}^K \to \mathbb{R}$, the graph of the function over the parameter space defines a smooth, K-dimensional submanifold embedded in $\mathbb{R}^{K+1}$, often called the loss surface:

$$\mathcal{M} = \{h(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\} \subseteq \mathbb{R}^{K+1}, \qquad \text{with} \qquad h(\boldsymbol{\theta}) = (\theta_1, \ldots, \theta_K, \mathcal{L}(\boldsymbol{\theta}))\,. \tag{2}$$

This definition of $\mathcal{M}$ naturally induces a Riemannian metric on the parameter space, as we now establish. For a configuration $\boldsymbol{\theta}$, the ambient tangent space centered at $h(\boldsymbol{\theta})$ is spanned by the Jacobian $\mathbf{J}_h(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) = [\mathbb{I}_K, \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})]^\top \in \mathbb{R}^{(K+1)\times K}$, and we can express two tangent space vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{T}_{\boldsymbol{p}}\mathcal{M}$ in parameter space coordinates as $\boldsymbol{v}_1 = \mathbf{J}_h(\boldsymbol{\theta})\tilde{\boldsymbol{v}}_1$ and $\boldsymbol{v}_2 = \mathbf{J}_h(\boldsymbol{\theta})\tilde{\boldsymbol{v}}_2$ where $\tilde{\boldsymbol{v}}_1, \tilde{\boldsymbol{v}}_2 \in \mathcal{T}_{\boldsymbol{\theta}}\mathcal{M}$ denote the corresponding tangent vectors in parameter space coordinates. We can then define inner products on the manifold via the metric tensor $g_{\boldsymbol{p}} : \mathcal{T}_{\boldsymbol{p}}\mathcal{M} \times \mathcal{T}_{\boldsymbol{p}}\mathcal{M} \to \mathbb{R}$ as:

$$g_{\boldsymbol{p}}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{v}_1^\top \boldsymbol{v}_2 = \tilde{\boldsymbol{v}}_1^\top \mathbf{J}_h(\boldsymbol{\theta})^\top \mathbf{J}_h(\boldsymbol{\theta})\tilde{\boldsymbol{v}}_2 = \tilde{\boldsymbol{v}}_1^\top \mathbf{G}(\boldsymbol{\theta})\tilde{\boldsymbol{v}}_2. \tag{3}$$

Here $\mathbf{G}(\boldsymbol{\theta}) = \mathbb{I}_K + \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^\top$ is the symmetric positive-definite (SPD) metric matrix, making $(\mathcal{M}, g)$ a Riemannian manifold whose inner products are locally distorted by curvature information from the loss gradient. This definition of the inner product is the *pullback metric* as it pulls the Euclidean inner product on the tangent space $\mathcal{T}_{\boldsymbol{p}}\mathcal{M}$ to the parameter space Θ, in which all geometric computations on $\mathcal{M}$ can therefore be carried out. Note that while the embedding h from Equation (2) captures only the loss geometry, the theory extends to any smooth embedding function.

2.2 Loss-minimizing submanifolds

For a non-convex, data-dependent loss function $\mathcal{L} : \Theta \to \mathbb{R}$, we define the loss level set at l as $S_l(\mathcal{D}) = \bigsqcup_{k\in\mathcal{K}} \{\boldsymbol{\theta} \in \mathcal{B}_k : \mathcal{L}(\boldsymbol{\theta}) = l \geq \mathcal{L}^*(\mathcal{D})\}$, where $\mathcal{B}_k$ is the basin of attraction of the k-th local minimum, $\mathcal{K}$ is the index set over all local minima and $\mathcal{L}^*(\mathcal{D})$ is the global minimum loss on data $\mathcal{D}$. We define the *loss-minimizing submanifold* as the disjoint union of minimum level sets within basins:

$$S_*(\mathcal{D}) = \bigsqcup\nolimits_{k\in\mathcal{K}} \{\boldsymbol{\theta} \in \mathcal{B}_k : \mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_k^*(\mathcal{D})\}\,. \tag{4}$$

Here $\mathcal{L}_k^*(\mathcal{D}) = \min_{\boldsymbol{\theta}'\in\mathcal{B}_k} \mathcal{L}(\boldsymbol{\theta}'; \mathcal{D})$ denotes the minimum loss value of the k-th basin, so $S_*(\mathcal{D})$ is essentially the collection of all local minima. A special case of the loss level submanifold is $S_{\mathcal{L}^*(\mathcal{D})}$ for which all solutions have minimal loss and are functionally similar on all points in $\mathcal{D}$. In the interpolation regime where $\mathcal{L}^*(\mathcal{D}) = 0$, all parameters in S_0 exactly interpolate the training data $\mathcal{D}$, yet may induce functionally distinct predictions on unseen inputs, as depicted in Figure 2. We remark that in the overparameterized regime where $K \gg N$, the loss imposes at most N constraints on the K-dimensional parameter space. With far more parameters than data points, there are locally at least $K - N$ directions in which the parameters can move without changing the loss value.

3 GEOMETRIC MECHANICS

Classical mechanics is fundamentally the study of motion and thus naturally describes the evolution of a dynamical system over time. In the following, we recall the Lagrangian formulation for a single-particle system evolving in the intrinsic coordinates of a Riemannian manifold, which in our case is the parameter space Θ. The trajectories will serve as the foundation for our sampler.

The path of a particle traveling on the loss manifold $\boldsymbol{\gamma} : [t_0, t_1] \to \mathcal{M}$ can also be expressed as a parameter space curve $\boldsymbol{\alpha} : [t_0, t_1] \to \Theta$, such that $\boldsymbol{\gamma}(t) = h(\boldsymbol{\alpha}(t))$, illustrated in Figure 3. At any time along the path, we characterize the energy of the particle by its kinetic and potential energies:

$$E_{\text{kin}} = T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) = \frac{1}{2}\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} \qquad E_{\text{pot}} = V(\boldsymbol{\alpha}) = \kappa \cdot \big(\mathcal{L}(\boldsymbol{\alpha}) - \mathcal{L}(\boldsymbol{\theta}^*_{\text{glob}})\big) \tag{5}$$

where $\dot{\boldsymbol{\alpha}} = \frac{d}{dt}\boldsymbol{\alpha}(t)$ is the velocity, and we follow standard notation conventions and neglect the explicit time-dependency unless needed. Note that $\boldsymbol{\theta}^*_{\text{glob}}$ denotes any global minimum and that $\kappa > 0$ is a constant which encodes the coupling of mass (which we set to unity) and gravitational acceleration. The particle dynamics are governed by the Euler-Lagrange (EL) equations, which are derived from the Lagrangian defined as the difference between kinetic and potential energy. The EL equations simplify under the definition of the manifold (see Appendix B), and we get a second-order ordinary differential equation (ODE) governing the acceleration:

$$\ddot{\boldsymbol{\alpha}} = -\left(\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} + \kappa\right) \cdot \operatorname{grad}\mathcal{L}(\boldsymbol{\alpha}). \tag{6}$$

where $\mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha}) := \nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})$ is the Hessian of the loss and $\operatorname{grad}\mathcal{L}(\boldsymbol{\alpha}) = \mathbf{G}(\boldsymbol{\alpha})^{-1}\nabla_\theta\mathcal{L}(\boldsymbol{\alpha})$ is the Riemannian gradient, hence guiding motion toward the steepest descent, which corresponds to a gravitational pull on the particle traveling on the manifold. Remark that the Hessian term relates to the kinetic energy and that the κ term relates to the potential energy. In practice, we can efficiently solve the second-order ODE numerically using automatic differentiation.

In the following two paragraphs, we highlight key properties, and the main issue with the dynamical system developed so far for exploring the submanifold of minima. In Figure 4 we show the behavior of the system as is, and under our improved dynamics presented in Subsection 3.1.

The straightest paths on $\mathcal{M}$.

The Lagrangian dynamics presented in Equation (6) confirm the well-known fact that geodesics – which are the straightest paths on the manifold – are curves that extremize the kinetic energy integral, since the dynamics coincide with the geodesic equations for $\kappa = 0$, i.e. when we disregard the potential energy. A particle with dynamics governed only by kinetic energy is a *free particle*, meaning that it has zero tangential acceleration and constant velocity along the path. Moreover in the case of $\kappa > 0$, the resulting trajectories can be interpreted as geodesics on a Finsler manifold under a modified metric:

$$F(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) = \sqrt{\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} + 2\kappa \cdot \mathcal{L}(\boldsymbol{\alpha})}. \tag{7}$$

For further details on Finslerian mechanics we refer to textbooks on such topics by e.g. Bucataru & Miron (2007).

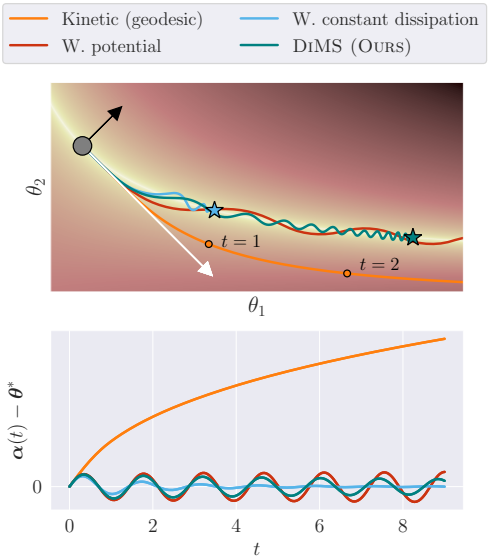


Figure 4: *Top:* With the white initial velocity, the systems affected by gravity remain closer to the minimum level set, yet without dissipation the particle never converges. With dissipation via a friction force, convergence is guaranteed, and our speed-dependent dissipation function allows traveling further than constant friction. *Bottom:* With the black initial velocity, curves affected by gravity oscillates around the minimum, while the dissipative systems converge steadily at different rates. The free particle can in contrast travel forever.

Particles that never rest. A free particle ($\kappa = 0$) has constant kinetic energy and no potential energy, and can therefore travel forever on a complete manifold – its trajectory length is unbounded and scales with integration time. Introducing a potential energy term can be interpreted as a gravitational pull on the particle, which affects the individual energy components along the path, yet conserves the total energy of the system; motion up- or downhill on the loss surface trades kinetic for potential energy and vice versa. Hence a particle with non-zero initial velocity and dynamics governed by Equation (6) never comes to rest and instead oscillates around the minimum level set. This can be problematic in high-dimensional settings where the dynamics are both computationally expensive to solve and difficult to simulate over long time horizons. See Appendix B for details.

3.1 Our solution: energy dissipation through speed-dependent friction

An obvious way to remove energy from the system - or formally to *dissipate* energy - is through the introduction of external forces, e.g. (Minguzzi, 2015; Cline, 2017). In the following we break energy conservation by introducing *friction*; a dissipative force that continuously removes energy from the system. For a dissipation function $\eta(t) > 0$, the alternative dynamical system is:

$$\ddot{\boldsymbol{\alpha}} = -\left(\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} + \kappa\right) \cdot \operatorname{grad} \mathcal{L}(\boldsymbol{\alpha}) - \eta(t) \cdot \dot{\boldsymbol{\alpha}}, \tag{8}$$

for which we describe key properties of the total energy of the system in Lemma 1 and Theorem 1. We provide proofs for all theorems and lemmas in Appendix C.

Lemma 1 (Energy dissipation and boundedness)**.** *Let $H(t) = T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) + V(\boldsymbol{\alpha}, t)$ be the total energy of a particle moving along $\boldsymbol{\alpha}(t)$ at any time $t > t_0$ where $T, V \geq 0$ and assume that $H(t_0) < \infty$. Under the dissipative dynamics in Equation (8) where $\eta(t) > 0$ is the dissipation function and holds for all $t \geq t_0$, the total energy dissipates over time as $\dot{H} = -2\eta(t)T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)$ and converges to $H_\infty := \lim_{t\to\infty} H(t)$. For all $t \geq t_0$ the total energy is bounded below as:*

$$0 \leq H_\infty \leq H(t) \leq H(t_0) < \infty$$

Theorem 1 (Convergence of motion)**.** *Under the assumptions in Lemma 1, and further assuming that the sublevel set $\Omega = \{(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) : H(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) \leq H(t_0)\}$ is compact, all trajectories converge by LaSalle's invariance principle to the largest invariant set*

$$\mathcal{I} = \{(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) : \dot{\boldsymbol{\alpha}} = \mathbf{0},\ \operatorname{grad} \mathcal{L}(\boldsymbol{\alpha}) = \mathbf{0}\},$$

consisting of equilibria of the loss. Assuming that $\mathcal{L}$ satisfies the strict saddle property, the attracting elements of $\mathcal{I}$ are almost surely local minima and converge to any $\boldsymbol{\theta} \in S_*(\mathcal{D})$.

A speed-dependent dissipation function. We suggest defining the dissipation function such that the friction force scales with the speed of the particle, i.e. the square root of the kinetic energy, thereby dissipating energy more at higher speed:

$$\eta(t) = \eta_0 \sqrt{T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)} = \eta_0/\sqrt{2} \cdot \|\dot{\boldsymbol{\alpha}}(t)\| \sqrt{1 + \|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\alpha}(t))\|^2 \cos^2(\omega)} \tag{9}$$

where η_0 is a hyperparameter controlling the effect of the dissipative force and ω denotes the angle between the velocity $\dot{\boldsymbol{\alpha}}(t)$ and the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\alpha}(t))$. Together with Equation (8), this dissipation function is what constitutes the sampling dynamics used by DIMS.

We remark that the dissipation strength directly depends on the Euclidean norm of the intrinsic velocity, $\|\dot{\boldsymbol{\alpha}}(t)\|$, revealing that the particle slows down faster at high speeds. Furthermore, the dissipation strength depends explicitly on the angle ω between the velocity of the particle and the steepness of the surface expressed by the gradient. As such, the energy dissipation is maximized when the velocity is proportional to the gradient, i.e. $\dot{\boldsymbol{\alpha}}(t) \propto \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\alpha}(t))$, while

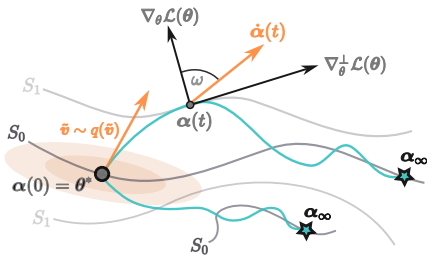


Figure 5: Dynamics of the proposed improved system. Depending on the initial velocity sample $\tilde{\boldsymbol{v}} \sim q(\tilde{\boldsymbol{v}})$, DIMS is capable of sampling distinct minimum level sets by dissipating energy particularly when moving proportional to the gradient, and eventually stop. In contrast the geodesic path is unconstrained and requires defining a stop time.

dissipation only depends on the speed of the Euclidean parameter space curve when the velocity is perpendicular to the gradient, i.e. $\dot{\boldsymbol{\alpha}}(t) \perp \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\alpha}(t))$. In other words, the particle is slowed down the least when moving in the direction $\nabla_{\theta}^{\perp} \mathcal{L}(\boldsymbol{\alpha}(t))$ that traverses the current level set.

In Theorems 2-3 we extend the previous convergence analysis to this specific dissipation function, from which we get a maximal convergence time and an exploration guarantee before convergence.

Theorem 2 (Convergence time)**.** *For a particle that traces the curve $\boldsymbol{\alpha} : [t_0, t_1] \to \Theta$ which converges to $\boldsymbol{\alpha}_\infty \in \mathcal{I}$ as $t_1 \to \infty$ according to Theorem 1, let $H(t)$ be the total energy at time t described in Lemma 1. Let furthermore $t_\epsilon = \inf\{t \geq t_1 : T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) < \epsilon\}$ be the proxy convergence time at which kinetic energy is sufficiently low to consider the particle not moving much. Then for dissipation function $\eta(t) = \eta_0 \sqrt{T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)} \geq 0$, an upper bound on the proxy convergence time is*

$$t_\epsilon \leq t_0 + H(t_0) \left(2\eta_0 \epsilon^{3/2}\right)^{-1}.$$

Theorem 3 (Lower bound on the $\boldsymbol{\gamma}$-arc-length)**.** *Under the assumptions in Theorem 2, and further defining $\Delta\mathcal{L} = \mathcal{L}(\boldsymbol{\alpha}(t_0)) - \mathcal{L}(\boldsymbol{\alpha}_\infty)$, the arc-length of the curve $\boldsymbol{\gamma}(t) = h(\boldsymbol{\alpha}) \in \mathcal{M}$ satisfies*

$$\left(T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t_0) + \kappa \Delta\mathcal{L}\right) \left(\sqrt{2}\eta_0 H(t_0)\right)^{-1} \leq \textsc{Length}(\boldsymbol{\gamma}).$$

From Theorem 2 we conclude that the time, t_ϵ, at which the particle converges to a low energy-state is finite and depends on the initial total energy, friction coefficient and the allowed speed-tolerance ϵ. While the bound may not be tight, it establishes that the dynamics converge in finite time, which is of practical and theoretical significance. Theorem 3 reveals that the particle travels at least a certain distance and that this distance is also based on the initial energy of the system. Hence, sampling with our proposed dynamics guarantees some exploration of the manifold before convergence.

3.2 How do we obtain samples from $S_*(\mathcal{D})$?

Essentially, to generate samples with DIMS, it suffices to appropriately initialize the dynamical system. Since we want to exploit the modified dynamical system defined in Equation (8) and Equation (9) to sample the submanifold of minima $S_*(\mathcal{D})$, we typically set the particle's initial position $\tilde{\boldsymbol{\mu}}$ to a minimum solution $\boldsymbol{\theta}^* \in S_*(\mathcal{D})$ obtained by training a neural network. This choice is practical as it reduces run time but we remark that convergence to $S_*(\mathcal{D})$ is guaranteed for any initial position.

A point estimate obtained via optimization corresponds to a stable equilibrium of the loss function, meaning that it can not move without an external force acting on it. We therefore need an initial velocity $\tilde{\boldsymbol{v}}$ along which motion is initialized. The particle path $\boldsymbol{\alpha}_{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{v}}, \eta_0} : [0, t_1] \to \Theta$ is then computed by numerically integrating the second-order ODE with initial conditions $\boldsymbol{\alpha}(0) = \tilde{\boldsymbol{\mu}}$ and $\dot{\boldsymbol{\alpha}}(0) = \tilde{\boldsymbol{v}}$. We treat the initial velocity as a random variable $\tilde{\boldsymbol{v}} \sim q(\tilde{\boldsymbol{v}})$ to give variation over the constructed paths. We can freely choose the distribution $q(\tilde{\boldsymbol{v}})$, so we define it heuristically as the Laplace approximation; a Gaussian respecting the local loss geometry at $\boldsymbol{\theta}^*$, which is common in approximate Bayesian inference (MacKay, 1992; Daxberger et al., 2021; Immer et al., 2021b; Bergamin et al., 2023; Yu et al., 2024).

Our sampler generates an implicit distribution over the submanifold of minima as follows:

$$\tilde{\boldsymbol{v}} \sim q(\tilde{\boldsymbol{v}}) = \mathcal{N}\left(\mathbf{0}, \mathbf{H}_{\tilde{\mathcal{L}}}(\tilde{\boldsymbol{\mu}})^{-1}\right) \quad \boldsymbol{\theta}^{(s)} = \textsc{DiMS}(\tilde{\boldsymbol{v}} | t_1, \eta_0, \tilde{\boldsymbol{\mu}}, \mathcal{D}) := \boldsymbol{\alpha}_{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{v}}, \eta_0}(t_1) \in S_*(\mathcal{D}) \tag{10}$$

where convergence to $S_*(\mathcal{D})$ is guaranteed for large enough t_1. As pointed out in Theorem 2 this time depends on the hyperparameter η_0 that controls exploration. Flipping the relation, we argue that selecting η_0 only depends on the computational budget available and that smaller η_0 are preferred. We describe how to obtain samples with DIMS computationally in Appendix D.

4 Related work

Loss landscape geometry. Understanding the geometry of the loss landscape has been a fundamental research topic in modern machine learning. Dinh et al. (2017) showed that neural network parameter spaces contain reparameterization invariant sets over which loss properties such as flatness are non-constant, motivating geometry-aware methods that are invariant to such reparameterizations.

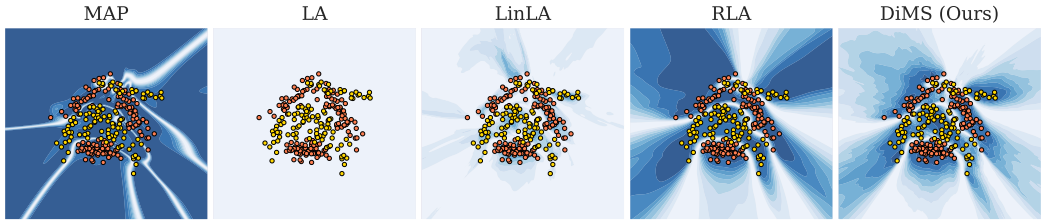


Figure 6: Binary classification on the banana dataset. The trained MAP model fits the training data well, yet gives unreliable uncertainty estimates. LA and LINLA induce predictive distributions over the input space, yet their approximations are too crude to fit the data properly. In contrast the geometry-aware methods fit the data well while providing better uncertainty estimates than the MAP, especially DIMS that also captures OOD uncertainty better than RLA. Uncertainty is given by the variance of the predictive distribution and we provide calibration metrics in Appendix D.2.

More recently Garipov et al. (2018) and Draxler et al. (2018) revealed that independently trained neural networks are connected by low-loss paths, suggesting that the reparameterization invariant set that minimizes the loss is well-described as a connected component. Entezari et al. (2021) conjectured that this connectivity is explained by permutation symmetries within the parameter space, with Zhao et al. (2025) extending this line of work to continuous symmetries.

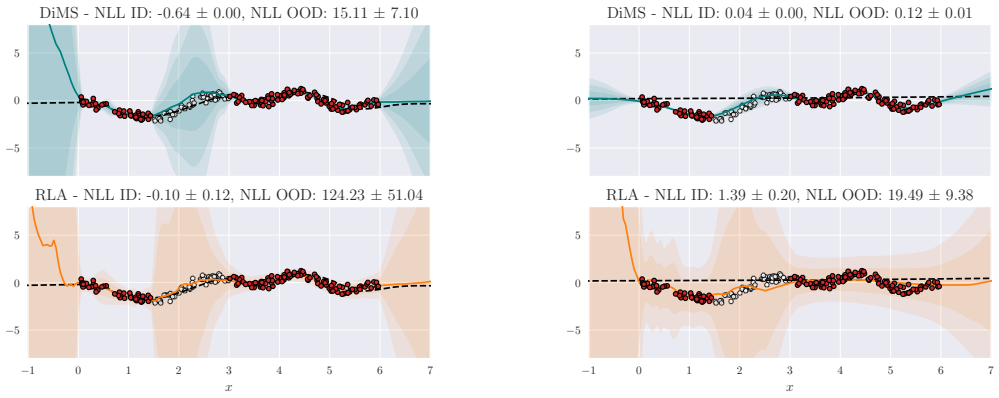
Mechanical systems in optimization and sampling theory. First-order optimization methods with momentum such as Polyak's heavy-ball (Polyak, 1964) and Nesterov's accelerated gradient (Nesterov, 1983) correspond to discretizations of second-order dissipative ODEs, a connection clarified by Su et al. (2016) and Kovachki & Stuart (2021). Our proposed continuous-time dynamics are a Riemannian generalization of Polyak's heavy-ball, which we recover exactly in the Euclidean setting with constant friction. Sampling algorithms based on mechanics – such as stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011), Hamiltonian Monte Carlo (HMC) (Neal, 2011) or its Lagrangian reformulation (Lan et al., 2012), along with their Riemannian variants (Girolami & Calderhead, 2011; Chen et al., 2014; Hartmann et al., 2022) – are guaranteed to recover the full posterior in the limit of infinite sampling time, which is rarely achieved in practice.

Sampling via posterior approximations. A conceptually simple approximate posterior technique is the Laplace approximation (MacKay, 1992) that defines the approximate distribution as a Gaussian with covariance given by the inverse Hessian. In neural network parameter spaces this assumption is often impractical and follow up works has instead used the the Generalized Gauss-Newton as the covariance (Daxberger et al., 2021) and linearized the function space samples (Immer et al., 2021b). Trajectory-based methods such as SWAG (Maddox et al., 2019) and deep ensembles provide scalable uncertainty estimates but remain agnostic to the geometric structure of the loss landscape. A more principled line of work suggests exploiting the local geometry when defining the approximate posterior; Bergamin et al. (2023) introduced the Riemannian Laplace approximation that defines a distribution in low-loss regions using the pullback metric defined previously, whereas Yu et al. (2024) corrected an inherent bias of the method by replacing the metric with the Fisher metric. In contrast our approach adjusts the bias by modifying the dynamical system. Other works (Roy et al., 2024; Fadel et al., 2025; Reichlin et al., 2025) followed alternative approaches that used a diffusion-like exploration of the reparameterization invariant solutions in the sampling process. While geometrically principled, these approaches are local and confined to a single loss basin.

5 EXPERIMENTS

We evaluate DIMS on uncertainty quantification tasks and compare with methods based on the Laplace approximation (LA), including the linearized (LINLA) and Riemannian (RLA) versions. We tune their prior precision and likelihood parameters (if any) using the marginal log-likelihood approach by Daxberger et al. (2021). Further details and code are provided in Appendix D.

Binary classification. We train a fully-connected network with 4 hidden layers of 16 units to fit the "banana" dataset in Figure 6, using `SiLU` activations, 50% dropout and label smoothing of 0.05, and we take $S{=}100$ samples with each method. We observe that DIMS exhibits high OOD uncertainty



(a) Trained initial $\tilde{\boldsymbol{\mu}} = \boldsymbol{\theta}^* \in S_*(\mathcal{D})$ (b) Random initial position $\tilde{\boldsymbol{\mu}} \notin S_*(\mathcal{D})$

Figure 7: Function space samples obtained by DIMS give high OOD estimates without breaking the fit on training data, which RLA does not guarantee. Even when initialized from a suboptimal position, DIMS provides well-behaved function space samples. Red and white dots are ID and OOD data, respectively, and the black line is the function induced by the initial position parameters. Uncertainty ranges are based on standard deviation and colored lines represent the ensemble averages.

Table 1: ID and OOD AUROC (↑) performance on images using a low-rank Hessian approximation.

		ACCURACY ↑	NLL ↓	BRIER ↓	ECE ↓	MCE ↓
MNIST	LA	0.564 ± 0.009	11.502 ± 0.507	0.083 ± 0.002	0.408 ± 0.009	0.614 ± 0.006
	LINLA	0.860 ± 0.007	2.608 ± 0.151	0.026 ± 0.001	0.127 ± 0.006	0.508 ± 0.007
	RLA	$\mathbf{0.953 \pm 0.001}$	0.609 ± 0.009	0.009 ± 0.0001	0.042 ± 0.0005	0.472 ± 0.008
	DIMS (OURS)	0.950 ± 0.0002	$\mathbf{0.271 \pm 0.002}$	$\mathbf{0.008 \pm 4 \cdot 10^{-5}}$	$\mathbf{0.035 \pm 0.0002}$	$\mathbf{0.379 \pm 0.009}$
FMNIST	LA	0.147 ± 0.003	53.519 ± 1.298	0.167 ± 0.001	0.825 ± 0.003	0.866 ± 0.003
	LINLA	0.385 ± 0.009	11.636 ± 0.270	0.115 ± 0.002	0.555 ± 0.008	0.647 ± 0.006
	RLA	0.767 ± 0.002	9.210 ± 1.266	0.045 ± 0.0004	0.220 ± 0.002	0.520 ± 0.006
	DIMS (OURS)	$\mathbf{0.814 \pm 0.002}$	$\mathbf{1.443 \pm 0.010}$	$\mathbf{0.033 \pm 0.0003}$	$\mathbf{0.149 \pm 0.001}$	$\mathbf{0.414 \pm 0.021}$

Trained on:	MNIST			FMNIST		
Evaluated on:	EMNIST	FMNIST	KMNIST	EMNIST	KMNIST	MNIST
LA	0.638 ± 0.006	0.745 ± 0.006	0.681 ± 0.007	0.432 ± 0.005	0.436 ± 0.005	0.433 ± 0.005
LINLA	0.728 ± 0.005	0.829 ± 0.006	0.775 ± 0.006	0.427 ± 0.003	0.427 ± 0.003	0.430 ± 0.003
RLA	0.846 ± 0.004	0.939 ± 0.003	0.882 ± 0.004	0.556 ± 0.004	0.586 ± 0.004	0.574 ± 0.004
DIMS (OURS)	$\mathbf{0.961 \pm 0.0002}$	$\mathbf{0.971 \pm 0.0003}$	$\mathbf{0.979 \pm 0.0002}$	$\mathbf{0.715 \pm 0.002}$	$\mathbf{0.779 \pm 0.001}$	$\mathbf{0.781 \pm 0.003}$

while remaining confident in regions supported by training data and preserving fine-grained details of the data distribution. While RLA performs similarly, it is less uncertain in a large part of the input space. The remaining methods either underfit the data or provide unreliable OOD uncertainty.

Toy regression. We train a fully connected neural network with 2 hidden layers of 16 hidden units and `GELU` activations. We use 20% dropout during training to increase the number of active neurons. We set the friction hyperparameter sufficiently low ($\eta_0 = 0.1$) and run DIMS samples until convergence, whereas RLA per definition stops at $t_1 = 1$. The resulting predictive uncertainty based on $S = 100$ Monte Carlo (MC) samples is shown in Figure 7a. Though RLA is highly uncertain out-of-distribution (OOD) it comes at the cost of fitting the data worse in-distribution (ID).

DIMS is position-agnostic in the sense that the gravitational pull forces the particle toward $S_*(\mathcal{D})$ no matter the starting position. Initializing DIMS randomly thus resemble continuous-time gradient descent with momentum and non-zero initial velocity that determines the direction in which such an optimizer is initialized. We conduct an experiment as such by initializing DIMS randomly, setting $\eta_0 = 1.0$ and running until reaching a minimum. Samples from DIMS still converge to solutions that fit the training data well with variation OOD while RLA samples perform worse under this initialization (Figure 7b). DIMS samples have limited OOD variation under such initial conditions, suggesting that maximum exploration happens when the kinetic energy term dominates over the gradient flow dynamics.

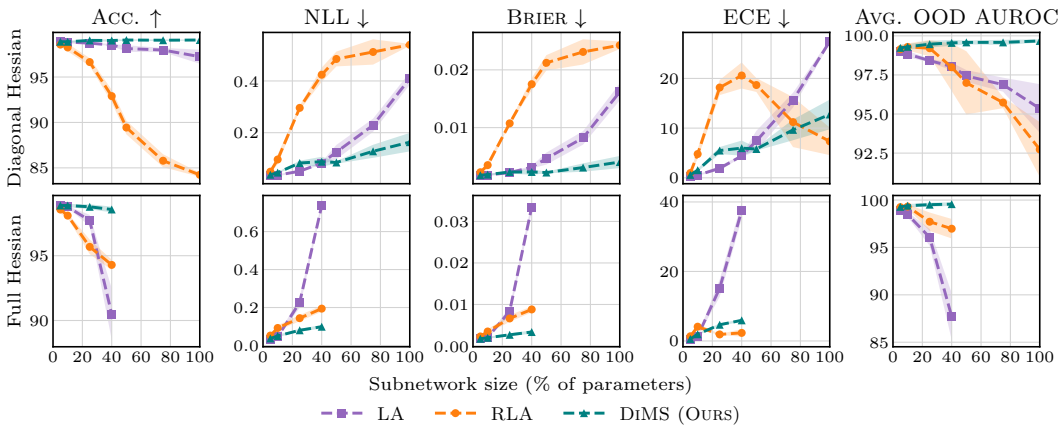


Figure 8: Performance metrics for subnetwork inference. OOD AUROC is averaged over datasets.

Image classification. We train a convolutional neural network (the LeNet architecture (LeCun et al., 1989)) with approximately $K = 44{,}000$ parameters on MNIST and Fashion MNIST and evaluate each method on the test sets along with the test sets of Extended MNIST and KMNIST. We provide the ID and OOD performance in Table 1. To define the Laplace approximations we compute the Hessian based on a randomly selected subset containing $N = 1{,}000$ of the training samples. We draw $S_B = 20$ parameter samples from each method and repeat the sampling process over $B = 5$ batches, resulting in a total of $S = 100$ samples per method using a friction coefficient of $\eta_0 = 0.1$. Computing the full Hessian is infeasible for a high-dimensional neural network as it requires inverting a $K \times K$ matrix, hence we use a low-rank approximation of it.

Subnetwork inference. Prior work by Sharma et al. (2023) revealed that neural networks do not need to be fully stochastic for properly quantifying uncertainty. For the image model considered, we follow this strategy and use DiMS to sample the parameters of a subnetwork of the full network. All training and sampling configurations remain fixed and we compare DiMS with LA and RLA. As this limits network size, subnetwork inference allows to compute the full Hessian for small enough subnetworks, yet we also consider subnetwork inference using the diagonal approximation of the Hessian as this is easier to scale. In both settings, DiMS is robust to the approximation and superior to the competing methods on most metrics (Figure 8), most likely due to its convergence guarantees.

6 Conclusion

We proposed DiMS, a sampler grounded in dissipative mechanics on Riemannian manifolds that is guaranteed to explore the minimum-loss submanifolds of neural networks, i.e., parameter space regions where models perform well on training data while retaining high predictive uncertainty on out-of-distribution inputs. We provided theoretical guarantees that the dynamics converge to these minimum-loss submanifolds in finite time regardless of the initial conditions, and validated the sampler empirically across various settings. DiMS consistently improves over or performs on par with competing Laplace-based methods on uncertainty quantification tasks, while being robust to its single physically-motivated hyperparameter – the friction coefficient – which governs energy dissipation in the dynamical system. We hope this work encourages further research on geometry and physics inspired approaches to minima exploration and uncertainty quantification in deep learning.

Limitations and future work. DiMS is, by design, a geometry-based approach and therefore computationally demanding, as is typical for such methods. The main cost stems from solving an ODE in a high-dimensional space, which costs $\mathcal{O}(SNK)$ where N, K and S are the number of data points, parameters and solver steps, respectively. While N and K can be reduced by mini-batching and subnetwork inference, the number of solver steps can be large for adaptive step-size methods in stiff systems. This can potentially be mitigated using tailored or stochastic solvers, but remains future work. Although DiMS is robust to the friction hyperparameter, a position-dependent formulation could be considered. Lastly, sampling reparameterization invariant models requires the architecture and the parameter space to induce a sufficiently expressive function space, and it is of interest to characterize the conditions under which this holds.

AI USE STATEMENT

In this work, we mainly used generative AI tools in the early stage of the project for research ideation and assistance with finding relevant prior literature in order to distinguish our initial idea from what was already established facts in the field. Thus the AI tools were used only as a form of search engine in this context and we verified existence of all papers manually. Later, we experimented with proof-checking our theory using AI tools, yet not in a particularly structured way. We instead relied on multiple authors confirming the statements independently. We did not use generative AI tools to build figures nor for writing the main body of the text, although we at times used such tools for polishing phrasings, etc.. For larger experiments, we furthermore leveraged AI tools for managing the job queue and handling configuration files, however we verified all final results and configurations manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provided a detailed explanation of our algorithm in Section 3. We additionally provided all details regarding the derivation of the algorithm and related concepts in Appendix A, B and C, including clearly stating the assumptions in the main manuscript. Regarding implementation details, we describe the main hyperparameter configurations in the main text and refer to Appendix D for further details, including tips and tricks related to the implementation of our method. Lastly, we included a link to an anonymized code base with the main implementation and the examples on binary classification and regression in Appendix D.

REFERENCES

Federico Bergamin, Pablo Moreno-Muñoz, Søren Hauberg, and Georgios Arvanitidis. Riemannian Laplace approximations for Bayesian neural networks. *Neural Information Processing Systems (NeurIPS)*, 2023.

Ioan Bucataru and Radu Miron. *Finsler-Lagrange geometry: Applications to dynamical systems*. Editura Academiei Române Buchurest, 2007.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Neural Information Processing Systems (NeurIPS)*, 2018.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*, 2014.

Douglas Cline. *Variational principles in classical mechanics*. University of Rochester River Campus Librarie, 2017.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless Bayesian deep learning. *Neural Information Processing Systems (NeurIPS)*, 2021.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, 2017.

Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*. Springer, 1992.

John R. Dormand and Peter J. Prince. A family of embedded Runge-Kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning (ICML)*, 2018.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

Samuel G Fadel, Hrittik Roy, Nicholas Krämer, Yevgen Zainchkovskyy, Stas Syrota, Alejandro Valverde Mahou, Carl Henrik Ek, and Søren Hauberg. VIKING: Deep variational inference with stochastic projections. *arXiv preprint arXiv:2510.23684*, 2025.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2011.

Marcelo Hartmann, Mark Girolami, and Arto Klami. Lagrangian manifold monte carlo on monge patches. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning (ICML)*, 2021a.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of Bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021b.

Hassan K Khalil and Jessy W Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, NJ, 2002.

Nikola B Kovachki and Andrew M Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research (JMLR)*, 2021.

Shiwei Lan, Vassilios Stathopoulos, Babak Shahbaba, and Mark Girolami. Lagrangian Dynamical Monte Carlo. *arXiv preprint arXiv:1211.3759*, 2012.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989.

John M Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the Loss Landscape of Neural Nets. In *Neural Information Processing Systems (NeurIPS)*, 2018.

David JC MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 1992.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. *Neural Information Processing Systems (NeurIPS)*, 2019.

Ettore Minguzzi. Rayleigh's dissipation function at work. *European Journal of Physics*, 2015.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2011.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl akad nauk Sssr*, 1983.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 1964.

Gunnar Rätsch, Takashi Onoda, and K-R Müller. Soft margins for adaboost. *Machine learning*, 42 (3):287–320, 2001.

Alfredo Reichlin, Miguel Vasco, and Danica Kragic Jensfelt. Walking on the Fiber: A Simple Geometric Approximation for Bayesian Neural Networks. *Transactions on Machine Learning Research (TMLR)*, 2025.

Hrittik Roy, Marco Miani, Carl Henrik Ek, Philipp Hennig, Marvin Pförtner, Lukas Tatzel, and Søren Hauberg. Reparameterization invariance in approximate Bayesian inference. *Neural Information Processing Systems (NeurIPS)*, 2024.

Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic? In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18, 2005.

Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research (JMLR)*, 2016.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011.

Hanlin Yu, Marcelo Hartmann, Bernardo Williams Moreno Sanchez, Mark Girolami, and Arto Klami. Riemannian Laplace Approximation with the Fisher Metric. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Understanding Mode Connectivity via Parameter Space Symmetry. In *International Conference on Machine Learning (ICML)*, 2025.

A Details on the loss manifold

A.1 Quantities of the pullback metric required for the derivations

The metric is given by:

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{I}_K + \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^\top, \tag{A.1}$$

The inverse metric and its square root follows from the Sherman-Morisson formula:

$$\mathbf{G}(\boldsymbol{\theta})^{-1} = \mathbb{I}_K - \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^\top}{1 + \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\|^2}, \tag{A.2}$$

$$\mathbf{G}(\boldsymbol{\theta})^{-\frac{1}{2}} = \mathbb{I}_K - \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^\top}{1 + \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\|^2 + \sqrt{1 + \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\|^2}}. \tag{A.3}$$

The Riemannian gradient can then be expressed as:

$$\operatorname{grad}\mathcal{L}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})}{1 + \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\|^2}. \tag{A.4}$$

The Christoffel symbols are needed to solve the Euler-Lagrange equations. These describe how coordinate basis vectors change along the manifold and determine how tangent vectors are corrected when moving on the manifold. We refer to classic textbooks on differential geometry, e.g. Do Carmo & Flaherty Francis (1992); Lee (2018), for further reading. For the specific definition of the manifold, the Christoffel symbols simplify to:

$$\Gamma^k_{ij}(\boldsymbol{\theta}) = \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{L}(\boldsymbol{\theta}) \cdot \sum_m^K \mathbf{G}^{-1}_{km}(\boldsymbol{\theta})\frac{\partial}{\partial\theta_m}\mathcal{L}(\boldsymbol{\theta})\right). \tag{A.5}$$

When computing the Christoffel symbols, we leveraged the closed form expression of the derivative of the metric that is the outer product of the Hessian columns with the gradient of the loss. By letting $\mathbf{H}^{(k)}_{\mathcal{L}}(\boldsymbol{\theta})$ denote the k'th column of the Hessian of $\mathcal{L}(\boldsymbol{\theta})$, the derivative of the metric is:

$$\frac{\partial}{\partial\theta_k}\mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}^{(k)}_{\mathcal{L}}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^\top + \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\mathbf{H}^{(k)}_{\mathcal{L}}(\boldsymbol{\theta})^\top. \tag{A.6}$$

For defining how the energy of a particle evolves over time, we need a description of how the metric evolves along the curve. We therefore establish that the time-derivative of the metric along $\boldsymbol{\alpha}(t)$ is:

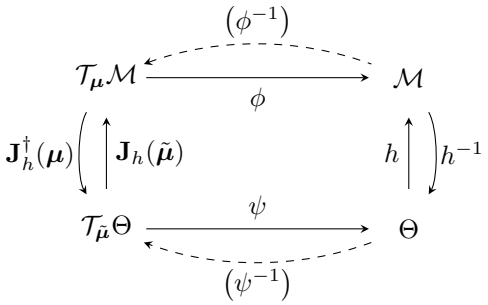
$$\dot{\mathbf{G}}(\boldsymbol{\alpha}) = \sum_{k=1}^K \dot{\alpha}_k \frac{\partial}{\partial\theta_k}\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})^\top + \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}^\top\mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha}) \tag{A.7}$$

Furthermore, the $\mathbf{G}$-norm of the gradient is:

$$\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\|_{\mathbf{G}(\boldsymbol{\theta})} = \sqrt{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})^\top\mathbf{G}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})} = \sqrt{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\|^2 + \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\|^4}. \tag{A.8}$$

A.2 Geometry-driven posterior approximations on the loss manifold

We consider the manifold $\mathcal{M}$ defined as the graph of the negative log-posterior loss on the training data as in Equation (1). For posterior approximations defined as the pushforward of a Gaussian in the tangent space via some map, we provide a schematic overview of the relations between the ambient tangent space $\mathcal{T}_{\boldsymbol{\mu}}\mathcal{M}$, intrinsic tangent space $\mathcal{T}_{\tilde{\boldsymbol{\mu}}}\Theta$, parameter space Θ and the manifold $\mathcal{M}$. The schematic is meant as a side-kick to what we depicted in Figure 3. We remark that we do not requires the pushforward map to be a diffeomorphism, yet remark that for non-bijective maps, the inverse mapping from samples to the initial position does not exist. This limits obtaining an explicit approximate posterior via the change-of-variables formula, yet admits an implicit approximate posterior obtained through sampling.



The speed of a curve $\boldsymbol{\gamma}(t) \in \mathcal{M}$ at time t is given by the norm of its velocity, $\|\dot{\boldsymbol{\gamma}}(t)\|$. Recall that an element of the ambient tangent space can be expressed in intrinsic coordinates as $\mathbf{J}_h(\boldsymbol{\theta})\tilde{\boldsymbol{v}}$, and using the Moore-Penrose inverse, $\mathbf{J}_h(\boldsymbol{\theta})^\dagger = \left(\mathbf{J}_h(\boldsymbol{\theta})^\top \mathbf{J}_h(\boldsymbol{\theta})\right)^{-1} \mathbf{J}_h(\boldsymbol{\theta})^\top = \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{J}_h(\boldsymbol{\theta})^\top$ we can hence map velocity samples from the ambient tangent space to their parameter space coordinates, since the mapping is preserved:

$$\tilde{\boldsymbol{v}} = \mathbf{J}_h(\boldsymbol{\theta})^\dagger \boldsymbol{v} = \mathbf{J}_h(\boldsymbol{\theta})^\dagger \mathbf{J}_h(\boldsymbol{\theta})\tilde{\boldsymbol{v}} = \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{G}(\boldsymbol{\theta})\tilde{\boldsymbol{v}} = \tilde{\boldsymbol{v}}. \tag{A.9}$$

Hence, mapping a vector from the parameter space coordinates to its tangent space coordinates preserves the mapping. We then define the distribution of initial velocities directly in the parameter space coordinates as:

$$\tilde{\boldsymbol{v}} = \mathbf{G}(\boldsymbol{\theta})^{-\frac{1}{2}}\boldsymbol{\epsilon}, \qquad q(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{A.10}$$

due to linearity of Gaussians. We see that the initial velocity samples from this distribution are invariant to the position $\boldsymbol{\mu}$, since:

$$\|\boldsymbol{v}\|^2 = \|\tilde{\boldsymbol{v}}\|^2_{\mathbf{G}(\boldsymbol{\theta})} = \boldsymbol{\epsilon}^\top \underbrace{\mathbf{G}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})^{-\frac{1}{2}}}_{=\mathbb{I}_K} \boldsymbol{\epsilon} = \|\boldsymbol{\epsilon}\|^2, \tag{A.11}$$

which is not the case if we disregard preconditioning by the square-root of the inverse metric, see:

$$\|\boldsymbol{v}\|^2 = \|\tilde{\boldsymbol{v}}\|^2_{\mathbf{G}(\boldsymbol{\theta})} = \tilde{\boldsymbol{v}}^\top \mathbf{G}(\boldsymbol{\theta})\tilde{\boldsymbol{v}}. \tag{A.12}$$

Remark that at a local minimum $\boldsymbol{\theta}^*$ the preconditioning has no effect since $\mathbf{G}(\boldsymbol{\theta}^*)^{-\frac{1}{2}} = \mathbb{I}_K$ by definition. For geometry-driven approximate posteriors that initialize the dynamics at a minimum - which is the case for any such distribution based on the Laplace approximation - the preconditioning does not need to be accounted for.

A.3 Which loss manifold?

In the approximate Bayesian inference framework, the loss function is given by the negative log-posterior Equation (1), which is optimized to obtain the MAP estimate. We further consider optimizing the prior precision and likelihood parameters (if any) using the marginal log-likelihood approach of Daxberger et al. (2021). This loss is then used to construct the Laplace approximation variants considered in this paper, namely, LA, LinLA, and RLA. Note that the (optimized) prior is critical for these approaches, as it influences the quality of the generated samples. For example, an ill-conditioned Hessian can result in very large initial velocities, which can cause the geodesics of RLA to traverse high-loss regions, while also requiring more computational time to solve. The same phenomenon is amplified for LA and LinLA, as is well known.

When sampling with DiMS we can freely choose the loss manifold while maintaining stable dynamics due to the convergence guarantees. In particular, one such example is to use only the log-likelihood component of the posterior loss, which is sufficient to induce the 0-loss set of reparameterization invariant solutions, thereby excluding the prior. The reason is that including the prior favors reparameterization-invariant solutions with an extra constraint on the weights, which limits the size of the minimum submanifolds and hence expressibility of DiMS, which by construction is able to travel further from the MAP estimate and more effectively explore the reparameterization-invariant sets. Note that removing the prior shifts the 0-loss set slightly away from the MAP estimate. Nevertheless, DiMS is shown to converge to this set regardless of this shift (Theorem 1). In addition, as shown in the experiments, DiMS is robust to the choice of initial velocities, while its behavior can be controlled through the friction parameter η.

B Derivations related to curve energy

For a unit-mass particle traveling on the loss manifold, its energy is fully captured by a kinetic and a potential energy component. The kinetic energy of a particle moving on the manifold can equally be formulated in its intrinsic coordinates, here referring to the parameter space. These formulations relate as:

$$E_{\text{kin}} = T(\boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}) = \frac{1}{2} \cdot g_{\boldsymbol{\gamma}}(\dot{\boldsymbol{\gamma}}, \dot{\boldsymbol{\gamma}}) = \frac{1}{2} \cdot \dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha}) \dot{\boldsymbol{\alpha}} = T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) \geq 0, \tag{B.1}$$

Similarly, the potential energy of such a system is described as:

$$E_{\text{pot}} = V(\boldsymbol{\gamma}) = \kappa \cdot \left(\mathcal{L} \circ h^{-1}\right)(\boldsymbol{\gamma}) - V^* = \kappa \cdot \left(\mathcal{L}(\boldsymbol{\alpha}) - \mathcal{L}(\boldsymbol{\theta}^*_{\text{glob}})\right) = V(\boldsymbol{\alpha}) \geq 0, \tag{B.2}$$

where $\boldsymbol{\theta}^*_{\text{glob}}$ is a parameter configuration associated to the minimum potential V^* and $\kappa > 0$ is a positive constant summarizing the gravitational acceleration and the particle's mass. The particle dynamics are then described through the Lagrangian that encodes the energy balance between the kinetic and potential energies:

$$L(\boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}) = T(\boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}) - V(\boldsymbol{\gamma}) = T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) - V(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}). \tag{B.3}$$

An extremal action curve is a trajectory of the particle, i.e. $\boldsymbol{\gamma} : [t_0, t_1] \to \mathcal{M}$, that respects the principle of stationary action, meaning that it minimizes an action functional S defined by:

$$\arg\min_{\boldsymbol{\gamma}} S = \arg\min_{\boldsymbol{\gamma}} \int_{t_0}^{t_1} L(\boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}) dt = \arg\min_{\boldsymbol{\alpha}} \int_{t_0}^{t_1} L(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) dt. \tag{B.4}$$

What should be evident from Equation (B.1) to Equation (B.4) is that energies and dynamics of a particle moving on the loss manifold can equivalently be expressed and computed by considering the parameter space trajectory of the particle.

A curve $\boldsymbol{\alpha}(t)$ is a stationary point of S if and only if it satisfies the Euler-Lagrange equations:

$$\frac{\partial L}{\partial \alpha_k} - \frac{d}{dt} \frac{\partial L}{\partial \dot{\alpha}_k} = 0. \tag{B.5}$$

B.1 Solution to the Euler-Lagrange equations

Inserting the definition of the Lagrangian from Equation (B.3) into the Euler-Lagrange equations, we get:

$$\frac{\partial T}{\partial \alpha_k} - \frac{\partial V}{\partial \alpha_k} - \frac{d}{dt} \left(\frac{\partial T}{\partial \dot{\alpha}_k} - \frac{\partial V}{\partial \dot{\alpha}_k} \right) = 0, \tag{B.6}$$

and since the potential is a function of the coordinates only, this simplifies to:

$$\frac{\partial T}{\partial \alpha_k} - \frac{d}{dt} \frac{\partial T}{\partial \dot{\alpha}_k} = \frac{\partial V}{\partial \alpha_k}. \tag{B.7}$$

Remark the coordinate dependency of the kinetic energy in the first term, a result of the fact that the particle is moving on the manifold with metric $\mathbf{G}(\boldsymbol{\theta})$. We insert the kinetic and potential energies and get:

$$\frac{1}{2} \left(\frac{\partial}{\partial \alpha_k} \left(\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha}) \dot{\boldsymbol{\alpha}} \right) - \frac{d}{dt} \frac{\partial}{\partial \dot{\alpha}_k} \left(\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha}) \dot{\boldsymbol{\alpha}} \right) \right) = \kappa \cdot \frac{\partial}{\partial \alpha_k} \mathcal{L}(\boldsymbol{\alpha}). \tag{B.8}$$

We proceed by considering the derivative terms individually. First realize that:

$$\frac{\partial}{\partial \alpha_k} \left(\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha}) \dot{\boldsymbol{\alpha}} \right) = \frac{\partial \mathbf{G}_{ij}(\boldsymbol{\alpha})}{\partial \alpha_k} \dot{\alpha}_i \dot{\alpha}_j. \tag{B.9}$$

Next, we leverage the product and the chain rule to get:

$$\begin{aligned} \frac{d}{dt} \frac{\partial}{\partial \dot{\alpha}_k} \left(\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha}) \dot{\boldsymbol{\alpha}} \right) &= 2 \frac{d}{dt} \mathbf{G}_{kj}(\boldsymbol{\alpha}) \dot{\alpha}_j \\ &= 2 \left(\frac{d}{dt} \mathbf{G}_{kj}(\boldsymbol{\alpha}) \right) \dot{\alpha}_j + 2 \mathbf{G}_{kj}(\boldsymbol{\alpha}) \ddot{\alpha}_j \\ &= 2 \frac{\partial \mathbf{G}_{kj}(\boldsymbol{\alpha})}{\partial \alpha_i} \dot{\alpha}_i \dot{\alpha}_j + 2 \mathbf{G}_{kj}(\boldsymbol{\alpha}) \ddot{\alpha}_j \\ &= \left(\frac{\partial \mathbf{G}_{kj}(\boldsymbol{\alpha})}{\partial \alpha_i} + \frac{\partial \mathbf{G}_{ki}(\boldsymbol{\alpha})}{\partial \alpha_j} \right) \dot{\alpha}_i \dot{\alpha}_j + 2 \mathbf{G}_{kj}(\boldsymbol{\alpha}) \ddot{\alpha}_j, \end{aligned} \tag{B.10}$$

and inserting Equation (B.9) and Equation (B.10) into Equation (B.8) gives:

$$\frac{1}{2}\left(\frac{\partial \mathbf{G}_{ij}(\boldsymbol{\alpha})}{\partial \alpha_k} - \frac{\partial \mathbf{G}_{kj}(\boldsymbol{\alpha})}{\partial \alpha_i} - \frac{\partial \mathbf{G}_{ki}(\boldsymbol{\alpha})}{\partial \alpha_j}\right)\dot{\alpha}_i\dot{\alpha}_j - \mathbf{G}_{kj}(\boldsymbol{\alpha})\ddot{\alpha}_j = \kappa \cdot \frac{\partial}{\partial \alpha_k}\mathcal{L}(\boldsymbol{\alpha}). \tag{B.11}$$

To find the equations of motion, we isolate the acceleration. To do so, realize that:

$$\sum_k \mathbf{G}^{-1}_{lk}(\boldsymbol{\alpha})\mathbf{G}_{kj}(\boldsymbol{\alpha}) = \delta^l_j, \tag{B.12}$$

which allows for rewriting Equation (B.11) using the Christoffel symbols:

$$\frac{1}{2}\mathbf{G}^{-1}_{lk}(\boldsymbol{\alpha})\left(\frac{\partial \mathbf{G}_{kj}(\boldsymbol{\alpha})}{\partial \alpha_i} + \frac{\partial \mathbf{G}_{ki}(\boldsymbol{\alpha})}{\partial \alpha_j} - \frac{\partial \mathbf{G}_{ij}(\boldsymbol{\alpha})}{\partial \alpha_k}\right)\dot{\alpha}_i\dot{\alpha}_j + \ddot{\alpha}_l = -\kappa \cdot \mathbf{G}^{-1}_{lk}(\boldsymbol{\alpha}) \cdot \frac{\partial}{\partial \alpha_k}\mathcal{L}(\boldsymbol{\alpha}) \tag{B.13}$$

$$\Rightarrow \qquad \Gamma^l_{ij}(\boldsymbol{\alpha})\dot{\alpha}_i\dot{\alpha}_j + \ddot{\alpha}_l = -\kappa \cdot \mathbf{G}^{-1}_{lk}(\boldsymbol{\alpha}) \cdot \frac{\partial}{\partial \alpha_k}\mathcal{L}(\boldsymbol{\alpha}), \tag{B.14}$$

from which the acceleration can be isolated and forms a second-order ordinary differential equation (ODE):

$$\ddot{\alpha}_l = -\Gamma^l_{ij}(\boldsymbol{\alpha})\dot{\alpha}_i\dot{\alpha}_j - \kappa \cdot \mathbf{G}^{-1}_{lk}(\boldsymbol{\alpha}) \cdot \frac{\partial}{\partial \alpha_k}\mathcal{L}(\boldsymbol{\alpha}), \tag{B.15}$$

By swapping indices and recalling that the term coming from the potential energy equals the Riemannian gradient, the expression becomes:

$$\ddot{\alpha}_k = -\Gamma^k_{ij}(\boldsymbol{\alpha})\dot{\alpha}_i\dot{\alpha}_j - \kappa \cdot \operatorname{grad}\mathcal{L}(\boldsymbol{\alpha})_k, \tag{B.16}$$

where $\operatorname{grad}\mathcal{L}(\boldsymbol{\alpha})_k$ denotes the k-th index of the Riemannian gradient. Remark that we used no assumptions about the metric and that these dynamics hold for a particle traveling on any Riemannian manifold.

Solution under the pullback metric. By leveraging the Christoffel symbols for the loss manifold provided in Equation (A.5), it is straight-forward to show that the Euler-Lagrange equation from above simplifies to:

$$\ddot{\boldsymbol{\alpha}} = -\frac{\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}}{1 + \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\|^2} \cdot \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}) - \kappa \cdot \operatorname{grad}\mathcal{L}(\boldsymbol{\alpha}) = -\left(\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} + \kappa\right) \cdot \operatorname{grad}\mathcal{L}(\boldsymbol{\alpha}). \tag{B.17}$$

Interpreting the potential term as a gravitational pull. The dynamics derived in the previous section also appear when adding a gravitational force component to the tangential acceleration of the particle traveling on $\mathcal{M}$ along $\boldsymbol{\gamma}(t)$. Leveraging the covariant derivative, see that:

$$\nabla_{\dot{\boldsymbol{\gamma}}}\dot{\boldsymbol{\gamma}} = \boldsymbol{d} - \langle \boldsymbol{d}, \hat{\boldsymbol{n}}(\boldsymbol{\alpha})\rangle \hat{\boldsymbol{n}}(\boldsymbol{\alpha}) = \left[-\kappa \cdot \operatorname{grad}\mathcal{L}(\boldsymbol{\alpha}), \quad \kappa \cdot \tfrac{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\|^2}{1+\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\|^2}\right]^\top, \tag{B.18}$$

where $\boldsymbol{d} = [0, \ldots, 0, -\kappa]^\top \in \mathbb{R}^{K+1}$ is the gravitational pull and the normal vector to the loss surface is given by $\hat{\boldsymbol{n}}(\boldsymbol{\alpha}) = [-\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}), 1]^\top \in \mathbb{R}^{K+1}$. Considering all parameter space dimensions, we see that this corresponds to the term originating from the potential energy when solving the Euler-Lagrange equations.

Gravity-influenced curves are geodesics on Finsler manifolds. As argued by Bucataru & Miron (2007), a Finslerian mechanical system with metric $F(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)$ corresponds to defining the Lagrangian as:

$$L(\boldsymbol{\alpha}, \dot{\alpha}, t) = \frac{1}{2}F(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)^2. \tag{B.19}$$

Inverting the relation and using the definition of the Lagrangian as the energy balance between the kinetic and potential energy, it is easy to see that the proposed dynamics from Equation (B.17) correspond to motion on a Finsler manifold with metric:

$$F(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) = \sqrt{\dot{\boldsymbol{\alpha}}(t)^\top \mathbf{G}(\boldsymbol{\alpha}(t))\dot{\boldsymbol{\alpha}}(t) + 2\kappa \cdot \mathcal{L}(\boldsymbol{\alpha}(t))}. \tag{B.20}$$

B.2 Energy flow equations under the pullback metric

We consider a parameter space curve $\boldsymbol{\alpha}(t)$ induced by the non-dissipative dynamics stated in Equation (B.17). In this section we will show the well-known fact that a non-dissipative system conserves energy by trading of kinetic with potential energy and vice versa and later show the implication of introducing dissipation as a friction force. First, we take the time-derivative of the kinetic energy:

$$\frac{d}{dt}T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) = \frac{1}{2}\frac{d}{dt}\left(\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}\right) = \dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\ddot{\boldsymbol{\alpha}} + \frac{1}{2}\dot{\boldsymbol{\alpha}}^\top \dot{\mathbf{G}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}. \tag{B.21}$$

We study the two terms separately. For the first term, we leverage the definition of the acceleration from Equation (B.17) to get:

$$\begin{aligned}\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\ddot{\boldsymbol{\alpha}} &= -\frac{\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}}{1+\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\|^2}\cdot \dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}) - \kappa\cdot \dot{\boldsymbol{\alpha}}^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}) \\ &= -\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}\cdot \dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\mathbf{G}(\boldsymbol{\alpha})^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}) - \kappa\cdot \dot{\boldsymbol{\alpha}}^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}) \\ &= -\left(\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} + \kappa\right)\cdot \dot{\boldsymbol{\alpha}}^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}).\end{aligned} \tag{B.22}$$

For the second term, we insert the definition of the time-derivative of the metric and get:

$$\frac{1}{2}\dot{\boldsymbol{\alpha}}^\top \dot{\mathbf{G}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} = \dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}\cdot \dot{\boldsymbol{\alpha}}^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}). \tag{B.23}$$

Combining the two terms and applying the chain rule inversely, we obtain the *law of conservation of energy* which is well-known from physics and exactly reveals the direct relation between kinetic and potential energy for the moving particle:

$$\dot{T}(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}}) = \frac{d}{dt}T(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}}) = -\kappa\cdot \dot{\boldsymbol{\alpha}}^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}) = -\kappa\cdot\frac{d}{dt}\mathcal{L}(\boldsymbol{\alpha}) = -\frac{d}{dt}V(\boldsymbol{\alpha}) = -\dot{V}(\boldsymbol{\alpha}). \tag{B.24}$$

Introducing dissipation through friction. The introduction of dissipation as a friction force based on dissipation function $\eta(t) > 0$ results in a modified version of the particle's acceleration:

$$\ddot{\boldsymbol{\alpha}} = -\left(\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} + \kappa\right)\cdot \operatorname{grad}\mathcal{L}(\boldsymbol{\alpha}) - \eta(t)\cdot\dot{\boldsymbol{\alpha}}. \tag{B.25}$$

As a consequence the the time-derivative of the kinetic energy changes. Inserting the dissipative acceleration equation into the previous derivation, it is straight-forward to see that:

$$\begin{aligned}\dot{T}(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}}) &= -\dot{V}(\boldsymbol{\alpha}) - \eta(t)\cdot \dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} \\ &= -\dot{V}(\boldsymbol{\alpha}) - 2\eta(t)\cdot T(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}}).\end{aligned} \tag{B.26}$$

Since $\eta(t)\cdot \dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} \geq 0$ by definition, the kinetic energy gradually decreases even though a drop in potential energy can still locally increase the kinetic energy.

Lastly, we rewrite the speed-dependent dissipation function considered in our sampler DIMS as:

$$\begin{aligned}\eta(t) = \eta_0\sqrt{T(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}})} &= \eta_0\sqrt{\frac{1}{2}\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}}} \\ &= \frac{\eta_0}{\sqrt{2}}\sqrt{\|\dot{\boldsymbol{\alpha}}\|^2 + \|\dot{\boldsymbol{\alpha}}^\top \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\|^2} \\ &= \frac{\eta_0}{\sqrt{2}}\sqrt{\|\dot{\boldsymbol{\alpha}}\|^2 + \left(\|\dot{\boldsymbol{\alpha}}\|\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\|\cos(\omega)\right)^2} \\ &= \frac{\eta_0}{\sqrt{2}}\|\dot{\boldsymbol{\alpha}}\|\sqrt{1+\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha})\|^2\cos^2(\omega)},\end{aligned} \tag{B.27}$$

where $\eta_0 > 0$ is the friction coefficient, treated as a hyperparameter. This clearly shows that the dissipation depends on the angle between the velocity and the gradient, ω, for the position-velocity state at time t.

C PROOFS OF LEMMAS AND THEOREMS

Lemma 1 - energy dissipation and boundedness. Consider a curve $\boldsymbol{\alpha} : [t_0,\infty] \to \Theta$ following the general dissipative dynamics presented in Equation (B.25), and recall that by definition the kinetic and potential from Equation (B.1) and Equation (B.2), the energy terms are bounded below as $T(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}},t), V(\boldsymbol{\alpha},t) \geq 0$. We define the total energy of the particle at time $t \geq t_0$ as the Hamiltonian:

$$H(t) := H(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}},t) = T(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}},t) + V(\boldsymbol{\alpha},t) \geq 0. \tag{C.1}$$

From Equation (B.26) and given that $\eta(t) > 0$ it holds that:

$$\dot{H}(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}},t) = \dot{T}(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}},t) + \dot{V}(\boldsymbol{\alpha},t) = -2\eta(t)T(\boldsymbol{\alpha},\dot{\boldsymbol{\alpha}},t) \leq 0 \tag{C.2}$$

revealing that the total energy $H(t)$ is monotonically non-increasing in t, and hence $0 \leq H(t) \leq H(t_0)$ for all $t \geq t_0$. Assuming $H(t_0) < \infty$, the Hamiltonian is bounded below and monotone, and therefore converges to the limit $H_\infty := \lim_{t\to\infty} H(t) \in \mathbb{R}$ by completeness of $\mathbb{R}$, such that $0 \leq H_\infty \leq H(t) \leq H(t_0)$ for all $t \geq t_0$.

Theorem 1 - convergence guarantee. From Lemma 1 and under the dynamics in Equation (B.25), recall that $0 \leq H_\infty \leq H(t) \leq H(t_0)$ for all $t \geq t_0$, so all trajectories remain in the sublevel set $\Omega = \{(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) : H(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) \leq H(t_0)\}$. If $V(\boldsymbol{\alpha}, t)$ is coercive, i.e. $V(\boldsymbol{\alpha}, t) \to \infty$ as $\|\boldsymbol{\alpha}\| \to \infty$, then since $T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) \to \infty$ as $\|\dot{\boldsymbol{\alpha}}\| \to \infty$, it holds that

$$H(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) = T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) + V(\boldsymbol{\alpha}, t) \to \infty \quad \text{as} \quad \|\boldsymbol{\alpha}\| \to \infty \text{ or } \|\dot{\boldsymbol{\alpha}}\| \to \infty, \tag{C.3}$$

so H is a radially unbounded function in $\boldsymbol{\alpha}$ and $\dot{\boldsymbol{\alpha}}$, and thus Ω is compact in $\boldsymbol{\alpha}$ and $\dot{\boldsymbol{\alpha}}$. Moreover assuming that H is continuously differentiable and since $\dot{H}(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) \leq 0$ on Ω, the conditions of LaSalle's invariance principle (e.g. in the textbook by Khalil & Grizzle (2002)) are satisfied and all trajectories converge to the largest invariant subset contained in

$$\mathcal{S} = \{(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) : \dot{H}(t) = 0\}. \tag{C.4}$$

Recall that $\dot{H}(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) = -2\eta(t)T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) = -\eta(t)\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} \leq 0$, due to positive definiteness of $\mathbf{G}(\boldsymbol{\alpha})$ and $\eta(t) > 0$ for all t, we note that $\dot{H}(t) = 0$ if and only if $\dot{\boldsymbol{\alpha}} = \mathbf{0}$, hence:

$$\mathcal{S} = \{(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) : \dot{\boldsymbol{\alpha}} = \mathbf{0}\}. \tag{C.5}$$

Next, note that a trajectory starting in $\mathcal{S}$ remains in $\mathcal{S}$ only if $\ddot{\boldsymbol{\alpha}} = 0$. Following the dynamics from Equation (B.25) for the particular states $(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) \in \mathcal{S}$, see that:

$$\ddot{\boldsymbol{\alpha}} = -\left(\dot{\boldsymbol{\alpha}}^\top \mathbf{H}_{\mathcal{L}}(\boldsymbol{\alpha})\dot{\boldsymbol{\alpha}} + \kappa\right) \cdot \operatorname{grad} \mathcal{L}(\boldsymbol{\alpha}) - \eta(t) \cdot \dot{\boldsymbol{\alpha}} \overset{(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) \in \mathcal{S}}{=} -\kappa \operatorname{grad} \mathcal{L}(\boldsymbol{\alpha}) = \mathbf{0}. \tag{C.6}$$

The equality is satisfied if and only if $\operatorname{grad} \mathcal{L}(\boldsymbol{\alpha}) = \mathbf{0}$ since $\kappa > 0$ by definition, hence the largest invariant subset $\mathcal{I}$ of $\mathcal{S}$ to which all trajectories converge as $t \to \infty$ is exactly zero-velocity states reaching stationary points of the loss surface:

$$\mathcal{I} = \{(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) : \dot{\boldsymbol{\alpha}} = \mathbf{0}, \operatorname{grad} \mathcal{L}(\boldsymbol{\alpha}) = \mathbf{0}\} \subset \mathcal{S}. \tag{C.7}$$

Assuming that the loss $\mathcal{L}$ satisfies the strict saddle property, saddle points and local maxima in $\mathcal{I}$ are unstable equilibria and trajectories converge almost surely to a local minimum of $\mathcal{L}$. We thereby conclude that the proposed speed-dependent dissipative dynamics converge to a local minimum of $\mathcal{L}$.

Remark 1 (Differentiability and coercivity of V in neural network settings.)**.** *For general neural network loss landscapes, flat or unbounded directions in parameter space may prevent sublevel sets from being compact, as V is not guaranteed to be coercive. In a Bayesian setting with a shifted negative log-joint as the potential, the potential takes the form $V(\boldsymbol{\alpha}) = \kappa \cdot \left(-\log p(\mathcal{D}|\boldsymbol{\alpha}) - \log p(\boldsymbol{\alpha}) - \mathcal{L}(\boldsymbol{\theta}^*_{glob})\right)$ where for a Gaussian prior, $p(\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_K)$, the negative log-prior contributes as an L^2-regularization term by $\frac{1}{2\sigma^2}\|\boldsymbol{\alpha}\|^2$, which grows unboundedly as $\|\boldsymbol{\alpha}\| \to \infty$, and the potential of the log-posterior is coercive, which guarantees compactness of Ω. If moreover the neural network architecture induces a smooth loss function through the activation functions, LaSalle's invariance principle applies.*

Theorem 2 - convergence time. Under the assumptions stated in Lemma 1, the Hamiltonian dissipates energy, hence $0 \leq H_\infty \leq H(t_1) \leq H(t_0)$. We formulate the energy difference equation of the Hamiltonian as the difference in total energy between the start- and endpoint of the curve as:

$$\Delta H = H(t_0) - H(t_1) = -\int_{t_0}^{t_1} \dot{H}(t)dt = 2\int_{t_0}^{t_1} \eta(t)T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)dt = 2\eta_0 \int_{t_0}^{t_1} T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)^{3/2}dt, \tag{C.8}$$

Now, let t_ϵ be a proxy convergence time at which the kinetic energy drops below a threshold ϵ, i.e. $t_\epsilon = \inf\{t \geq t_0 : T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) \leq \epsilon\}$. Due to the infimum condition that defines t_ϵ it holds that:

$$\epsilon^{3/2}(t_\epsilon - t_0) = \int_{t_0}^{t_\epsilon} \epsilon^{3/2}dt \leq \int_{t_0}^{t_\epsilon} T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)^{3/2}dt, \tag{C.9}$$

Now inserting this lower bound into the difference equation along with boundedness of the Hamiltonian gives:

$$2\eta_0\epsilon^{3/2}(t_\epsilon - t_0) \leq 2\eta_0 \int_{t_0}^{t_1} T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t)^{3/2}dt = \Delta H \leq H(t_0). \tag{C.10}$$

By isolating the proxy convergence time, we get an upper bound on the form:

$$t_\epsilon \leq t_0 + \frac{H(t_0)}{2\eta_0\epsilon^{3/2}}. \tag{C.11}$$

Theorem 3 - γ-arc-length. Given the Hamiltonian described in Lemma 1 and the formulation of the Hamiltonian difference equation described in Theorem 2 along with the fact that $T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) \leq H(t) \leq H(t_0)$ for any $t \geq t_0$, it holds that:

$$\Delta H = 2\eta_0 \int_{t_0}^{t_1} T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}})^{3/2} dt = 2\eta_0 \int_{t_0}^{t_1} \sqrt{T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}})} \cdot T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}) dt \leq 2\eta_0 H(t_0) \int_{t_0}^{t_1} \sqrt{T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}})} dt. \quad \text{(C.12)}$$

Note furthermore that the integral of the square-root of the kinetic energy is proportional to the arc-length of the curve $\boldsymbol{\gamma} : [t_0 \to t_1] \in \mathcal{M}$ by:

$$\int_{t_0}^{t_1} \sqrt{T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}})} dt = \frac{1}{\sqrt{2}} \int_{t_0}^{t_1} \sqrt{\dot{\boldsymbol{\alpha}}^\top \mathbf{G}(\boldsymbol{\alpha}) \dot{\boldsymbol{\alpha}}} dt = \frac{1}{\sqrt{2}} \int_{t_0}^{t_1} \|\dot{\boldsymbol{\gamma}}\| dt = \frac{\textsc{Length}(\boldsymbol{\gamma})}{\sqrt{2}}, \quad \text{(C.13)}$$

where $\textsc{Length}(\boldsymbol{\gamma})$ denotes the arc-length of the curve. In combination with the upper bound on the energy difference, we get a lower bound on the arc-length:

$$\frac{\sqrt{2}\Delta H}{2\eta_0 H(t_0)} = \frac{\Delta H}{\sqrt{2}\eta_0 H(t_0)} \leq \textsc{Length}(\boldsymbol{\gamma}). \quad \text{(C.14)}$$

From the convergence properties in Theorem 3, the asymptotic behavior of the Hamiltonian guarantees that:

$$\lim_{t\to\infty} H(t) = \lim_{t\to\infty} T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) + \lim_{t\to\infty} V(\boldsymbol{\alpha}, t) = V(\boldsymbol{\alpha}, \infty) = \kappa \cdot (\mathcal{L}(\boldsymbol{\alpha}_\infty) - \mathcal{L}(\boldsymbol{\alpha}^*)) \geq 0, \quad \text{(C.15)}$$

since $\lim_{t\to\infty} T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t) = 0$. Note that we by $\boldsymbol{\alpha}_\infty$ denote the position to which the trajectory converges where $\operatorname{grad} \mathcal{L}(\boldsymbol{\alpha}_\infty) = \mathbf{0}$ and remark that $\boldsymbol{\alpha}_\infty$ is not necessarily a global minimum, hence $\mathcal{L}(\boldsymbol{\alpha}_\infty) \geq \mathcal{L}(\boldsymbol{\alpha}^*)$. Now, letting the stop time $t_1 \to \infty$, we see that:

$$\lim_{t_1\to\infty} \Delta H = H(t_0) - \lim_{t_1\to\infty} H(t_1) \quad \text{(C.16)}$$

$$= T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t_0) + \kappa \cdot (\mathcal{L}(\boldsymbol{\alpha}(t_0)) - \mathcal{L}(\boldsymbol{\alpha}_\infty)) \quad \text{(C.17)}$$

$$= T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t_0) + \kappa \Delta \mathcal{L}. \quad \text{(C.18)}$$

Collecting the pieces concludes the proof and we get a lower bound on the form:

$$\frac{T(\boldsymbol{\alpha}, \dot{\boldsymbol{\alpha}}, t_0) + \kappa \Delta \mathcal{L}}{\sqrt{2}\eta_0 H(t_0)} \leq \textsc{Length}(\boldsymbol{\gamma}). \quad \text{(C.19)}$$

D Experimental Details

Link to code base: https://anonymous.4open.science/r/dims-iclr-8E66.

Implementation details. All experiments are implemented in PyTorch. Differential equations are solved using the Dormand-Prince 5th-order solver (Dormand & Prince, 1980) from `torchdiffeq` (Chen et al., 2018), with absolute and relative tolerances of `atol=1e-6` and `rtol=1e-7` to ensure stable dynamics. These tolerances are a computational bottleneck for both standard geodesic sampling (RLA) and for DIMS at high exploration rates (low η_0 and large t_1), as the adaptive step-size controller requires many evaluations to satisfy the error bounds. Automatic differentiation is used throughout to compute acceleration terms efficiently via Hessian-vector products. For all experiments, the prior precision and noise variance are optimized by maximizing the log-marginal likelihood following the suggestion by Immer et al. (2021a). The gravitational pull is fixed to $\kappa = 1$ across all runs, leaving the friction coefficient η_0 as the primary hyperparameter controlling exploration. Networks are trained using the Adam optimizer with learning rate 0.001 and weight decay 0.01.

The linearized Laplace (LINLA) uses the same Gaussian samples as the standard, sampled Laplace, i.e. a Gaussian $\boldsymbol{\theta}_s \sim \mathcal{N}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma})$ centered at the MAP $\boldsymbol{\theta}^*$, but it computes predictions via the first-order Taylor expansion $f_{\boldsymbol{\theta}^*}(\boldsymbol{x}) + \mathbf{J}_{f_{\boldsymbol{\theta}^*}}(\boldsymbol{x})(\boldsymbol{\theta}_s - \boldsymbol{\theta}^*)$, which can be evaluated efficiently with forward-mode automatic differentiation. RLA uses the centered Laplace samples $\tilde{\boldsymbol{v}}_s = \boldsymbol{\theta}_s - \boldsymbol{\theta}^*$ as initial velocities and integrates the geodesic equation on the loss manifold from $\boldsymbol{\theta}^*$ for time t_1.

D.1 Toy Regression

A 2-layer fully connected network with 16 hidden units per layer and `GELU` activations was trained for $50{,}000$ epochs using full-batch gradient descent with 20% dropout on the Snelson dataset (Snelson & Ghahramani, 2005). The Laplace approximation was computed using the exact full Hessian over all parameters, with prior precision and variance optimized accordingly. Out-of-distribution data consists of the 52 points in $x \in [1.5, 3]$, and the remaining $N = 148$ points form the training set. We set $\eta_0 = 0.1$ and integrate the dynamics for $t_1 = 50$, drawing $S = 100$ samples from each method.

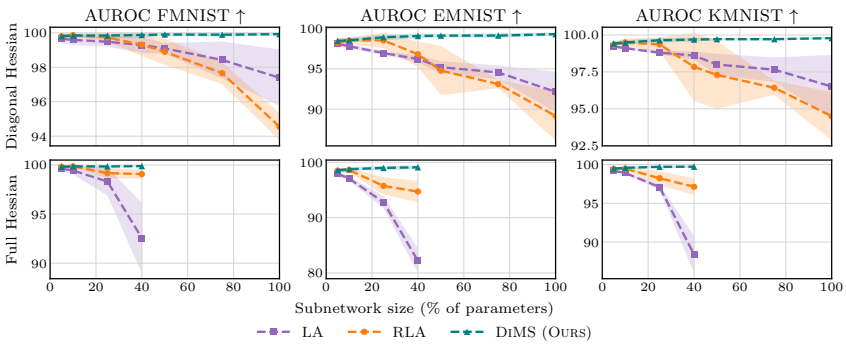


Figure 9: OOD AUROC and subnetwork inference. Trained on MNIST, evaluated on other datasets.

D.2 BINARY CLASSIFICATION

A 4-layer fully connected network with 16 hidden units per layer and `SiLU` activations was trained for 30,000 epochs using full-batch gradient descent, with a learning rate of 10^{-3}, weight decay of 0.01, 50% dropout and label smoothing of 0.05 on the 2D banana-shaped classification problem (Rätsch et al., 2001). Of all the available data, we randomly selected $N = 265$ points as the training set while the test set contains 795 held out observations. We tune the prior precision and draw $S = 100$ samples from the Euclidean and Riemannian Laplace approximations. For DIMS we draw initial velocity samples from an isotropic Gaussian scaled by the largest eigenvalue of the Laplace covariance, before solving the dynamics on a loss surface defined as the mean log-likelihood and without prior. We set $\eta_0 = 0.5$ and integrate for time $t_1 = 200$. We remark that RLA dynamics are unstable on this choice of loss manifold, even for integration time $t_1 = 1$. We provide the in-distribution performance metrics in Table 2.

Table 2: In-distribution performance for the banana classification dataset. We report the mean values over the $S = 100$ samples along with the standard error of the mean as the uncertainty estimate. Calibration metrics are computed with 10 bins.

METHOD	Accuracy ↑	NLL ↓	BRIER ↓	ECE ↓	MCE ↓
MAP	85.41	0.601	0.1279	10.24	50.00
LA	46.42	0.701	0.2539	6.24	**6.24**
LINLA	84.40	0.389	0.1176	8.42	16.94
RLA	**86.54**	**0.312**	**0.0967**	4.90	20.11
DIMS (OURS)	**86.54**	0.355	0.1040	**4.36**	11.41

D.3 IMAGE CLASSIFICATION

We train a LeNet classifier (approx. 44,000 parameters) on MNIST and Fashion MNIST using Adam for 10 epochs using a batch size of 128, with a log-softmax output layer and negative log-likelihood loss. After training, the MAP checkpoint $\boldsymbol{\theta}^*$ serves as the reference point for all posterior approximations. All results reported are computed over 3 repetitions of the experiment; we provide the mean and standard error in the tables and figures.

We compare DIMS to Laplace-based approximations that fit a Gaussian $\mathcal{N}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma})$ where $\Sigma^{-1} = \mathbf{H}_{\tilde{\mathcal{L}}}(\boldsymbol{\theta}^*)$ is the Hessian of the negative log-posterior computed on the full training set. For computational reasons, we estimate the Hessian matrix on a randomly selected subset of $N = 1{,}000$ training samples when solving the dynamical system and since the full Hessian is intractable for a network with a high-dimensional parameter space, we resort to an approximation of the Hessian. For DIMS we set the friction coefficient to $\eta_0 = 0.1$ and consider a low-rank approximation of the Hessian. For the subnetwork inference experiment, we are restricted by the `laplace-torch` library to either use a diagonal Hessian approximation or the full Hessian, which does not scale to the full network size. The experiment reveals the limitation of previous methods when reaching a network scale that requires approximation of the Hessian. In contrast DIMS appears to be robust to this choice, which is a clear benefit of the approach.

In Figure 9 we provide the per-dataset evaluation of OOD AUROC (according to Hendrycks & Gimpel (2016)) scores under subnetwork inference. The previously discussed trend remains on the per-dataset level.