On Equivalences between Weight and Function-Space Langevin Dynamics

Anonymous authors
Paper under double-blind review

Abstract

Approximate inference for overparameterized Bayesian models appears challenging, due to the complex structure of the posterior. To address this issue, a recent line of work has investigated the possibility of directly conducting approximate inference in the "function space", the space of prediction functions. This paper provides an alternative perspective to this problem, by showing that for many models – including a simplified neural network model – Langevin dynamics in the overparameterized "weight space" induces equivalent function-space trajectories to certain Langevin dynamics procedures in function space. Thus, the former can already be viewed as a function-space inference algorithm, with its convergence unaffected by overparameterization. We provide simulations on Bayesian neural network models and discuss the implication of the results.

1 Introduction

Consider a common Bayesian predictive modeling setting, where we are provided with i.i.d. observations $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$, a likelihood model $p(\{y_i\} \mid \{x_i\}, \theta) = \prod_{i=1}^n p(y_i \mid f(x_i; \theta))$ determined by a prediction function $f(\cdot; \theta)$, and a prior $\pi_\theta(\mathrm{d}\theta)$. We are interested in the predictive distribution $p(y_* \mid x_*, \mathcal{D}) = \int \pi_{\theta|\mathcal{D}}(\mathrm{d}\theta) p(y_* \mid x_*, \theta)$, induced by the posterior $\pi_{\theta|\mathcal{D}}$.

Modern machine learning models are often overparameterized, meaning that multiple parameters may define the same likelihood. For example, in Bayesian neural network (BNN) models where $\theta \in \mathbb{R}^d$ denote the network *weights*, we can obtain a combinatorial number of equivalent parameters by reordering the neurons, after which $f(\cdot; \theta)$, and thus the likelihood, remain unchanged. Consequently, the posterior measure exhibits complex structures and becomes hard to approximate; for example, its Lebesgue density may contain a large number of global maxima.

Starting from Sun et al. (2019); Wang et al. (2019); Ma et al. (2019), a recent literature investigates the possibility of simplifying inference by approximating a *function-space posterior*. Concretely, let $\mathcal{A} : \mathbb{R}^d \to \mathcal{F} \subset \mathbb{R}^{|\mathcal{X}|}, \theta \mapsto f(\cdot; \theta)$ denote a "parameterization map". Then

$$p(y_* \mid x_*, \mathcal{D}) = \int \pi_{\theta|\mathcal{D}}(\mathrm{d}\theta)\, p(y_* \mid f(x_*; \theta)) = \int (\mathcal{A}_\# \pi_{\theta|\mathcal{D}})(\mathrm{d}f)\, p(y_* \mid f(x_*)) = \int \pi_{f|\mathcal{D}}(\mathrm{d}f)\, p(y_* \mid f(x_*)),$$

where $\mathcal{A}_\#(\cdot)$ refers to the pushforward, and $\pi_{f|\mathcal{D}}$ denotes the function-space posterior defined by the prior $\mathcal{A}_\# \pi_\theta =: \pi_f$ and likelihood $p(y \mid x, f) = p(y \mid f(x))$. As shown above, $\pi_{f|\mathcal{D}}$ is sufficient for prediction. Moreover, it often has simpler structures: for example, for ultrawide BNN models with a Gaussian π_θ, π_f may converge to a Gaussian process (GP) prior (Lee et al., 2018; Matthews et al., 2018; Yang, 2019), in which case $\pi_{f|\mathcal{D}}$ will also converge to a GP posterior. Thus, it is natural to expect approximate inference to be easier in function space.

While the intuition has been appealing, existing works on function-space inference tend to be limited by theoretical issues: principled applications may require full-batch training (Sun et al., 2019), Gaussian likelihood (Shi et al., 2019), or specifically constructed models (Ma et al., 2019; Ma & Hernández-Lobato, 2021). Many approaches rely on approximations to the function-space prior, which can make the functional KL divergence

unbounded (Burt et al., 2020). Additionally, there is a lack of understanding about optimization convergence, or the expressivity of the variational families used. In contrast, gradient-based MCMC methods, such as Hamiltonian Monte Carlo (HMC) or Langevin dynamics (LD)-based algorithms, can be applied to a broad range of models. Their convergence behaviors are well-understood (Roberts & Tweedie, 1996; Villani, 2009), and intriguingly, their performance often appears to be satisfying on massively overparameterized models (Zhang et al., 2019; Izmailov et al., 2021), even though they are implemented in weight space.

This paper bridges the two lines of approaches by showing that

- In various overparameterized models, including a simplified BNN model, weight-space Langevin dynamics (LD) is equivalent to a reflected / Riemannian LD procedure in function space, defined by the pushforward metric.
- For practical feed-forward network models, the equivalence still appears to hold in simulations: weight-space LD still produces predictive distributions that appears to approach the functional posterior, at a rate that does not depend on the degree of overparameterization.

The equivalence has important implications: it means that principled function-space inference has always been possible and in use. Thus, explicit consideration of function-space posteriors *alone* will not be sufficient to guarantee improvement over existing approaches, and more careful analyses are necessary to justify possible improvement.

It should be noted that in several scenarios, it has been established that overparameterization does not necessarily hinder the convergence of LD. Moitra & Risteski (2020) proves that polynomial convergence can be possible for a family of *locally* overparameterized models, despite the non-convexity introduced by the overparameterization.[1] Dimensionality-independent convergence has also been established for infinite-width NNs in the mean-field regime (e.g., Mei et al., 2019), even though its implication for practical, finite-width models is less clear. We are unaware of strict equivalence results as provided in this paper, but we should also emphasize that it is not their technical sophistication that makes them interesting; it is rather *their implications for BNN inference, which appear underappreciated*: the results justify the use of LD as an effective function space inference procedure, in settings that match or generalize previous work. For example, Example 2.1 covers overparameterized linear models, and many popular approaches (e.g., Osband et al., 2018; He et al., 2020) are only justified in this setting.

Our results contribute to the understanding of the real-world performance of BNN models, as they provide a theoretical support for the hypothesis that inference may be good enough in many applications, and is not necessarily the limiting factor in a predictive modeling workflow. In this aspect, our results complement a long line of existing work which examined the influence of likelihood, prior and data augmentation in BNN applications, with an emphasis on classification tasks with clean labels; see Aitchison (2020); Wenzel et al. (2020); Fortuin et al. (2021), to name a few.

2 Equivalence between Weight and Function-Space Langevin Dynamics

2.1 A Warm-up Example

Suppose the prior measure π_θ is supported on an open subset of $\mathbb{R}^d$ and has Lebesgue density p_θ. The weight-space posterior $\pi_{\theta|\mathcal{D}}$ can be recovered as the stationary measure of the (weight-space) Langevin dynamics

$$\mathrm{d}\theta_t = \nabla_\theta(\log p(\mathbf{Y} \mid \theta_t, \mathbf{X}) + \log p_\theta(\theta_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{WLD}$$

where we write $\mathbf{X} := \{x_i\}_{i=1}^n, \mathbf{Y} := \{y_i\}_{i=1}^n$ for brevity.

The pushforward measure $\mathcal{A}_\# \pi_\theta =: \pi_f$ provides a prior in function space. Combining π_f and the likelihood leads to a posterior, $\pi_{f|\mathcal{D}}$. When the function space $\mathcal{F} := \operatorname{supp} \pi_f$ can be equipped with a Riemannian

[1] This result is still not fully unimpeded by overparameterization, as it quantifies convergence to the weight-space posterior, which necessarily requires traversal through all symmetric regions.

manifold structure of dimensionality $k \leq d$, it is intuitive that we could sample from $\pi_{f|\mathcal{D}}$ by simulating a Riemannian Langevin dynamics on $\mathcal{F}$ (Girolami & Calderhead, 2011). In coordinate form:

$$\mathrm{d}\tilde{f}_t = V(\tilde{f}_t)\mathrm{d}t + \sqrt{2G^{-1}(\tilde{f})}\mathrm{d}B_t, \tag{FLD}$$

where $\tilde{f}_t \in \mathbb{R}^k$ is the coordinate of $f_t \in \mathcal{F}$, $G^{-1}(\tilde{f}) = (g^{ij})_{i,j}$ is the inverse of the coordinate matrix of the metric, $\mathrm{d}B_t$ is the standard Brownian motion, and

$$V^i(\tilde{f}) = g^{ij}\partial_j\Big(\log p(\mathbf{Y} \mid f(\tilde{f}), \mathbf{X}) + \log \frac{d\pi_f}{d\mu_{\mathcal{F}}}(f) - \frac{\log|G|}{2}\Big) + \partial_j g^{ij}.$$

$\mu_{\mathcal{F}}$ denotes the corresponding Riemannian measure.

We are interested in possible equivalences between the induced function-space trajectory of (WLD), $\{\mathcal{A}\theta_t\}$, and the trajectory of possibly generalized versions of (FLD), with metric defined as the pushforward of the Euclidean metric by $\mathcal{A}$ or its generalization. The easiest example is the following:

Example 2.1 (equivalence in linear models)**.** *Suppose the map $\mathcal{A}$ is linear. For expository simplicity, further assume that $\pi_\theta = \mathcal{N}(0, I)$, and that the input space $\mathcal{X} = \{\mathtt{x}_1, \ldots, \mathtt{x}_{|\mathcal{X}|}\}$ has finite cardinality, so that any function can be identified as a $|\mathcal{X}|$ dimensional vector $(f(\mathtt{x}_1), \ldots, f(\mathtt{x}_{|\mathcal{X}|}))$.*

(i) If $\mathcal{A}$ is a bijection, the above vector representation will provide a coordinate for $\mathcal{F}$. In this coordinate, the pushforward metric has coordinate $(AA^\top)^{-1}$ (see e.g., Bai et al., 2022), where A denote the coordinate matrix of $\mathcal{A}$. (FLD) *with this metric reduces to*

$$\mathrm{d}\tilde{f}_t = (AA^\top)\nabla_{\tilde{f}}\left(\log p(\mathbf{Y} \mid \tilde{f}, \mathbf{X}) - \frac{1}{2}\|A^{-1}\tilde{f}\|_2^2\right)\mathrm{d}t + \sqrt{2AA^\top}\mathrm{d}B_t.$$

(Derivation for the prior term may be found in Appendix A.1.) By Ito's lemma, the above SDE also describes the evolution of $A\theta_t$, for θ_t following (WLD)*.*

(ii) The equivalence continue to hold in the overparameterized case (e.g., when $d > |\mathcal{X}|$): consider the decomposition $\mathbb{R}^d = \mathrm{Ran}(\mathcal{A}^\top) \oplus \mathrm{Ker}(\mathcal{A})$. Then the evolution of θ_t in (WLD) *"factorizes" along the decomposition: the likelihood gradient is fully contained in $\mathrm{Ran}(\mathcal{A}^\top)$ and thus only influences $\mathrm{Proj}_{\mathrm{Ran}(\mathcal{A}^\top)}\theta_t$, whereas $\mathrm{Proj}_{\mathrm{Ker}(\mathcal{A})}\theta_t$ has no influence on $\mathcal{A}\theta_t$. Therefore, we can describe the evolution of the former independently, thereby reducing to the exactly parameterized case.*

The second case above provides the first intuition on why (WLD) is not necessarily influenced by overparameterization. While technically simple, it is relevant as it covers random feature models, which (formally) include infinitely wide DNNs in the "kernel regime" (Jacot et al., 2018), where the pushforward metric converges to a constant value.

2.2 Overparameterization via Group Actions

It is often the case that overparameterization can be characterized by group actions; in other words, there exists some group H on $\mathbb{R}^d$ s.t. any two parameters $\theta, \theta' \in \mathbb{R}^d$ induce the same function $\mathcal{A}\theta = \mathcal{A}\theta'$ if and only if they belong to the same orbit. In such cases, we can identify $\mathcal{F}$ as the quotient space $\mathbb{R}^d/H$ and the map $\mathcal{A} : \mathbb{R}^d \to \mathcal{F}$ as the quotient map, and it is desirable to connect (WLD) to possibly generalized versions of (FLD) on $\mathcal{F}$. This subsection presents such results.

To introduce our results, we first recall some basic notions in group theory. Let H be a Lie group, i.e., H is both a group and a smooth manifold. An *action* of H on $\mathbb{R}^d$ is a map $(\varphi, p) \mapsto \varphi \cdot p$ with $\varphi \in H$ and $p \in \mathbb{R}^d$, s.t. for all $\varphi_1, \varphi_2 \in H$ and $p \in \mathbb{R}^d$, we have $e \cdot p = p, \varphi_1 \cdot (\varphi_2 \cdot p) = (\varphi_1\varphi_2) \cdot p$ where $e \in H$ denotes the identity. For any $\varphi \in H$, introduce the map $\Gamma_\varphi : \mathbb{R}^d \to \mathbb{R}^d, p \mapsto \varphi \cdot p$. Then the action is *free* if Γ_φ has no fixed point for all $\varphi \neq e$, *proper* if the preimage of any compact set of the map $(\varphi, p) \mapsto \varphi \cdot p$ is also compact, and *smooth* if Γ_φ is smooth for each $\varphi \in H$. An *orbit* is defined as $H \cdot p := \{\varphi \cdot p : \varphi \in H\}$ where $p \in \mathbb{R}^d$.

Analysis of free group actions The quotient manifold theorem (Lee, 2012, Theorem 21.10) guarantees that $\mathbb{R}^d/H$ is a smooth manifold if the action is smooth, proper and free. To define the pushforward metric on $\mathcal{F}$, we further assume that the action is isometric, i.e., Γ_φ is an isometry for every $\varphi \in H$. Under this condition, a metric on $\mathcal{F}$ can be defined as[2]

$$\langle \mathrm{d}\mathcal{A}|_p u, \mathrm{d}\mathcal{A}|_p v\rangle_{T_{\mathcal{A}p}\mathcal{F}} := \langle u, v\rangle_{\mathbb{R}^d}, \quad \forall p \in \mathcal{F},\ u, v \in T_p(H \cdot p)^\perp \subset \mathbb{R}^d.$$

The following proposition establishes the equivalence under discrete group action.

Proposition 2.1 (proof in Appendix A.1)**.** *Suppose H is a discrete group acting smoothly, freely, properly on $\mathbb{R}^d$, and $\mathcal{A}$ is such that $\mathcal{A}\theta = \mathcal{A}\theta'$ if and only if $\theta' \in H \cdot \theta$. If either (a) the prior p_θ is constant and the group action is isometric; or (b) $H = \{e\}$ is trivial, then the equivalence between* (WLD) *and* (FLD) *will continue to hold.*

Remark 2.1. For continuous groups that act freely, the situation is more complicated, and depends on how the orbits are embedded in the ambient space $\mathbb{R}^d$. For example, a drift term depending on the mean curvature vector of the orbit may be introduced when pushing a Brownian motion using the quotient map (JE, 1990), and when the mean curvature vanishes, the equivalence will continue to hold, as shown in our Example 2.1 (ii). Analysis for non-free group actions is primarily complicated by the fact that the quotient space is no longer a manifold in general (Satake, 1956). Still, as we show in Example 2.3, similar equivalence results can be established under the action of symmetric groups.

It is intuitive that simulation of (FLD) should constitute an efficient function-space inference algorithm, in light of the established guarantees of (Riemannian) LD. Thus, the established equivalences provide strong justifications for the use of (WLD) in practice.

The pushforward metric used to define the equivalent (FLD) is often believed to encode a desirable inductive bias, and has been used to characterize or design first-order optimization methods (e.g., Luk & Grosse, 2018; Lee et al., 2019). However, there are also models on which it may be unsuitable, such as very deep feed-forward networks, for which the pushforward metric may degenerate (Jacot et al., 2019). It should also be noted that VI and MCMC methods can have different behavior on overparameterized models: for VI methods, it may still be necessary to explicitly account for overparameterization. While recent works have made similar observations (e.g., Sun et al., 2019), and provided some examples (Wang et al., 2019; Kurle et al., 2022), the following example may provide additional insight:

Example 2.2 (LD vs. particle-based VI on torus)**.** *Let $\mathcal{A}\theta := ([\theta_1], \ldots, [\theta_d])$, where $[a] := a - \lfloor a \rfloor \in [0, 1)$. Let π_θ, π_f have constant densities, and the negative log likelihood be unimodal and locally strongly convex. Then we have $\mathcal{F} = \mathbb{T}^d$, the d-dimensional torus, and by Proposition 2.1,* (WLD) *is equivalent to Riemannian LD on $\mathcal{F}$. As $\mathbb{T}^d$ is a compact manifold,* (FLD) *enjoys exponential convergence (Villani, 2009), and so does the induced function-space measure of* (WLD)*.*

Particle-based VI methods approximate the weight-space posterior with an empirical distribution of particles $\{\theta^{(i)}\}_{i=1}^M$, and update the particles iteratively. Consider the W-SGLD method in Chen et al. (2018): its update rule resembles (WLD)*, but with the diffusion term replaced by a deterministic "repulsive force" term, $\tilde{v}_t(\theta)\mathrm{d}t$, where*

$$\tilde{v}_t(\theta) := \sum_{j=1}^M \frac{\nabla_{\theta^{(j)}} k_h(\theta, \theta^{(j)})}{\sum_{k=1}^M k_h(\theta^{(j)}, \theta^{(k)})} + \frac{\sum_{j=1}^M \nabla_{\theta^{(j)}} k_h(\theta, \theta^{(j)})}{\sum_{k=1}^M k_h(\theta, \theta^{(k)})},$$

and k_h is a radial kernel with bandwidth h. Formally, in the infinite-particle, continuous time limit, as $h \to 0$, both $\tilde{v}_t\mathrm{d}t$ and the diffusion term implements the Wasserstein gradient of an entropy functional (Carrillo et al., 2019), and W-SGLD and LD are formally equivalent (Chen et al., 2018).

The asymptotic equivalence between (WLD) *and W-SGLD breaks down in this example: whereas* (WLD) *induces a function-space measure that quickly converges to $\pi_{f|\mathcal{D}}$, this is not necessarily true for W-SGLD. Indeed, its induced function-space measure may well collapse to a point mass around the MAP, regardless of the number of particles. To see this, let $\theta^* \in [0, 1)^d$ be any MAP solution so that $\nabla_\theta \log p(\mathbf{Y} \mid \mathbf{X}, \theta^*)p(\theta^*) = 0$.*

[2]It is well-defined since $\mathrm{d}\mathcal{A}|_p$ is an isomorphism between $T_p(H \cdot p)^\perp$ and $T_{\mathcal{A}p}\mathcal{F}$, and the isometry assumption ensures that the definition is independent of the choice of p in the orbit (Lee, 2018).

Then for any fixed $h = O(1)$, as $M \to \infty$, the configuration $\{\theta^{(i,M)} = (10^{10^{Mi}}, 0, \ldots, 0) + \theta^*\}_{i=1}^M$ will constitute an approximate stationary point for the W-SGLD update. This is because the posterior gradient term is always zero, but the repulsive force term vanishes due to the very large distances between particles in weight space.

Past works have noted the pathologies of particle-based VI in high dimensions (Zhuo et al., 2018; Ba et al., 2021), but this example is interesting as it does not require an increasing dimensionality. Rather, it is global overparameterization that breaks the asymptotic convergence to LD.

Analysis of non-free group actions As we have shown in Example 2.2, Proposition 2.1 already demonstrates some equivalence between (WLD) and (FLD), in the presence of global overparameterization. It can also be combined with Example 2.1 (ii) to construct models exhibiting both local and global overparameterization. Still, we present a more relevant example below, which is a simplified BNN model exhibiting permutational symmetry. We note that this model allows for a non-constant neural tangent kernel, which is an important feature of realistic NN models (see e.g., Ghorbani et al., 2019; Wei et al., 2019).

Example 2.3 (simplified BNN model)**.** *Consider the model $f(x;\theta) := \sum_{i=1}^d \sin(\theta_i x)$, which is a two-layer BNN with the second layer frozen at initialization.*

Let the prior support $\operatorname{supp} \pi_\theta$ *be contained in $(0, +\infty)^d$. Then by the linear independence of sine functions, for $\mathcal{A}\theta = \mathcal{A}\theta'$ to hold, θ' must be a permutation of θ, and thus the symmetry in this model can be described by the symmetric group S_d consisting of all permutations on the set $\{1, \ldots, d\}$. The action of S_n on the weight space $\mathbb{R}^d$ is non-free, and the function space is a manifold with boundary, namely a polyhedral cone $C_n := \{\theta \in \mathbb{R}^d : \theta_1 \le \theta_2 \le \cdots \le \theta_d\}$.*

Let p_t denote the distribution of θ_t. Appendix A.2 proves that the pushforward distribution $\tilde{p}_t := \mathcal{A}_\# p_t$ follows the Fokker-Planck equation with the Neumann boundary condition:

$$\begin{cases} \partial_t \tilde{p}_t(\theta) = -\nabla \cdot (\tilde{p}_t(\theta) \nabla_\theta (\log p(\mathbf{Y} \mid \theta, \mathbf{X}) + \log p_\theta(\theta))) + \Delta \tilde{p}_t(\theta), & \theta \in \mathcal{F}^\circ \\ \partial_\theta \tilde{p}_t(\theta) / \partial v = 0, & v \in N_\theta, \theta \in \partial \mathcal{F}, \end{cases}$$

where $\partial\mathcal{F}$ and $\mathcal{F}^\circ$ are the boundary and the interior of $\mathcal{F}$, respectively, and N_θ is the set of inward normal vectors of $\mathcal{F}$ at θ. The evolution of $\tilde{p}_t$ is closely related to the reflected Langevin dynamics *in $\mathcal{F}$ (Sato et al., 2022), which keeps its trajectory in $\mathcal{F}$ by reflecting it at $\partial\mathcal{F}$. When the posterior is strongly log-concave in C_n, the equivalence implies that the function-space measure $\tilde{p}_t$ enjoys a fast convergence. In contrast,* convergence of (WLD) to the weight-space posterior can be much slower*, as it needs to visit an exponential number of equivalence classes.*

3 Numerical Study

While our theoretical results have covered two simplified BNN models, the models are still different from those employed in practice. In this section we validate our findings on practical BNN models on a toy 1D regression dataset, as well as a collection of semi-synthetic datasets adapted from the UCI regression datasets commonly used in previous work (e.g., Sun et al., 2019; Wang et al., 2019; Ma et al., 2019).

3.1 1D Regression Dataset

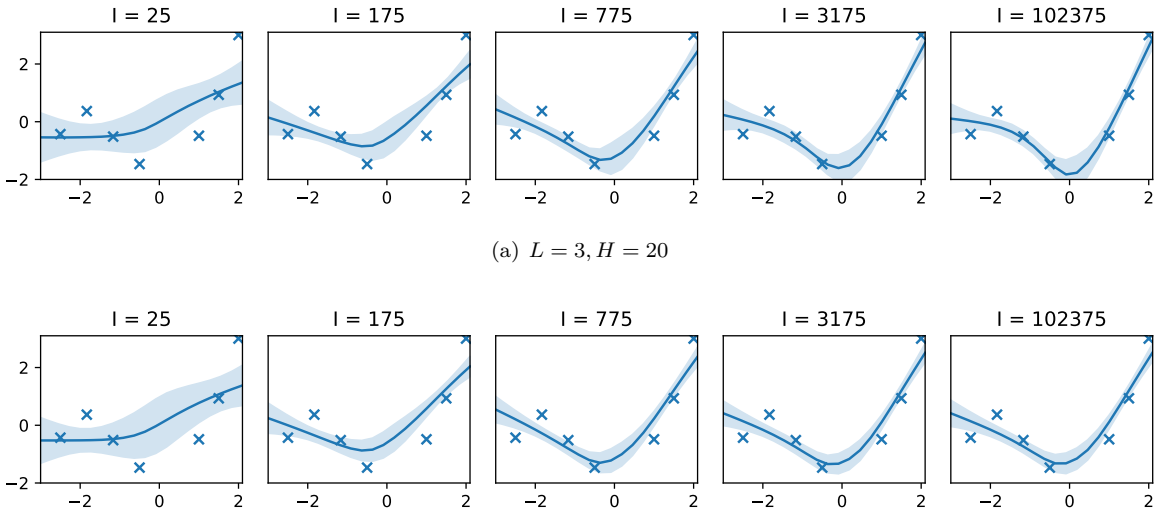
We first consider BNN inference on a toy 1D regression dataset, and check if the function-space measure induced by (WLD) appears to converge at a similar rate, across models with increasing degree of overparameterization. Concretely,

1. we will visualize the pointwise credible intervals, which are informative about one-dimensional marginal distributions of the function-space measure;
2. when the training sample size n is small, we approximately evaluate the approximation quality of $(n+1)$-dimensional marginal distributions of $f(\mathbf{X}_e) := (f(x_1), \ldots, f(x_n), f(x_*))$, by estimating the kernelized

Stein discrepancy (KSD) between the marginal distribution q induced by (WLD), and the approximate ground truth p.
The KSD can be estimated because it only accesses p through its score function,

$$\begin{aligned}\nabla_{f(\mathbf{X}_e)} \log p &= \nabla_{f(\mathbf{X}_e)} \Big(\log \frac{d\pi_{f(\mathbf{X}_e)}}{d\mu_{Leb}} + \log p(\mathbf{Y} \mid f(\mathbf{X}_e)) \Big) \\ &= \nabla_{f(\mathbf{X}_e)} \Big(\log \frac{d\pi_{f(\mathbf{X}_e)}}{d\mu_{Leb}} + \log p(\mathbf{Y} \mid f(\mathbf{X})) \Big), \qquad \text{(since } \mathbf{X} \subset \mathbf{X}_e\text{)} \end{aligned} \tag{1}$$

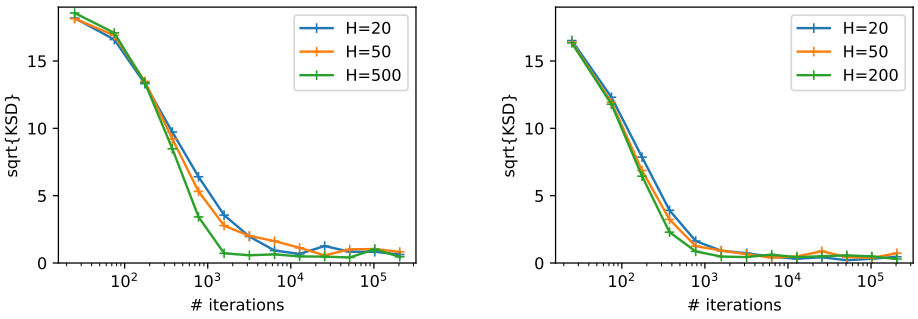
where $\pi_{f(\mathbf{X}_e)}$ denotes the respective marginal distribution of π_f, and μ_{Leb} denotes the Lebesgue measure. We estimate the first term by fitting nonparametric score estimators (Zhou et al., 2020) on prior samples. The second term can be evaluated in closed form.



(a) $L = 3, H = 20$

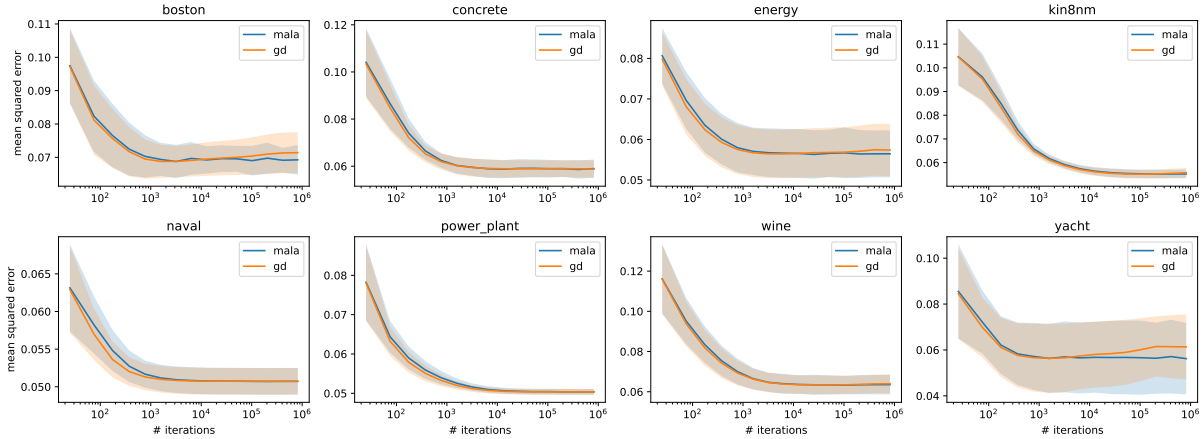
(b) $L = 3, H = 500$

Figure 1: 1D regression: visualization of the induced function-space measure of MALA after I iterations. We plot the pointwise 80% credible intervals. The results for $L = 2$ are deferred to Fig. 4.

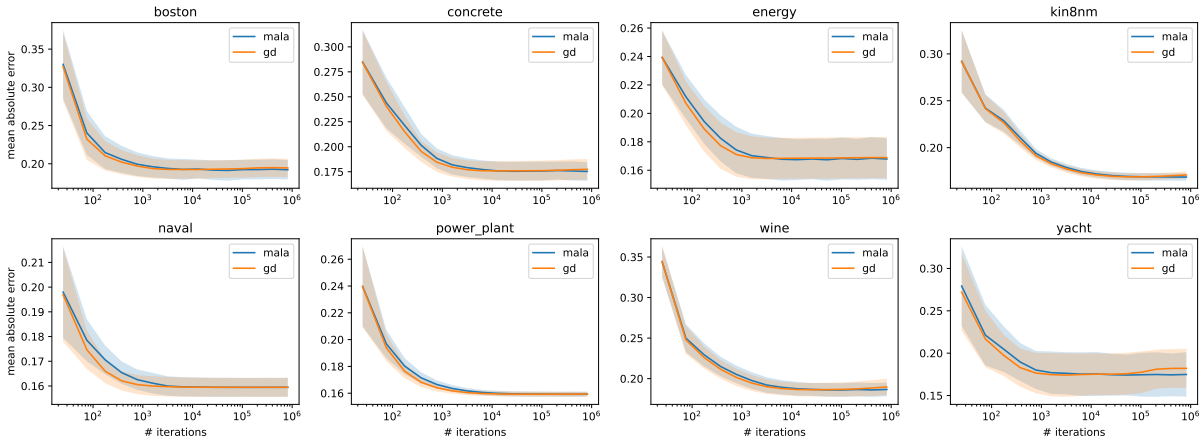


(a) $L = 2$ (b) $L = 3$

Figure 2: 1D regression: estimated $\sqrt{\text{KSD}}$ between the LD predictive distribution and the approximate function-space posterior. We simulate 1000 LD chains. For the approximate posterior, we estimate the prior score term in (1) using 5×10^6 samples.



(a) Gaussian likelihood / mean square error



(b) Laplace likelihood / mean absolute error

Figure 3: Semi-synthetic experiment: estimated average-case loss (4) under different choices of likelihood, for $H = 2, L = 200$. Shade indicates standard deviation across 8 independent replications.

We use feed-forward networks with factorized Gaussian priors, and the standard initialization scaling: $f(x;\theta) := f^{(L)}(f^{(L-1)}(\dots f^{(0)}(x)))$, where

$$f^{(l)}(h^{(l-1)}) := \sigma^{(l)}\left(W^{(l)}h^{(l-1)} + b^{(l)}\right), \quad \mathrm{vec}(W^{(l)}) \sim \mathcal{N}\left(0, (\dim h^{(l-1)})^{-1} I\right), \quad b^{(l)} \sim \mathcal{N}(0, 0.2I), \tag{2}$$

and the activation functions $\sigma^{(l)}$ are SELU (Klambauer et al., 2017) for hidden layers ($l < L$) and the identity map for the output layer ($l = L$). We vary the network depth $L \in \{2, 3\}$, and the width of all hidden layers $H \in [20, 500]$.

The training data is generated as follows: the inputs consist of $\lfloor 2n/3 \rfloor$ evenly spaced points on $[-2.5, -0.5]$, and the remaining points are evenly placed on $[1, 2]$. The output is sampled from $p(y \mid x) = \mathcal{N}(x \sin(1.5x) + 0.125x^3, 0.01)$. We use $n = 7$ for visualization, and $n = 3$ for KSD evaluation. The difference is due to challenges in approximating the KSD: we need the score estimator to generalize to out-of-distribution inputs (approximate posterior as opposed to prior samples), which is challenging in high dimensions.

We implement (WLD) with the Metropolis-adjusted Langevin algorithm (MALA), and evaluate the induced function-space samples for varying number of iterations. The step size is set to $0.025/nH$, so that the function-space updates have a similar scale.

We visualize the posterior approximations in Fig. 1 and Fig. 4, and report the approximate KSD in Fig. 2. As we can see, the convergence appears to happen at a similar rate, which supports the equivalence results.

3.2 Semi-Synthetic Experiments

We now investigate the behavior of (WLD) on datasets that better reflect real-world applications. The previous experiments cannot scale to larger datasets due to challenges in evaluating the KSD. Thus, here we turn to less direct evaluations, using semi-synthetic datasets adapted from the UCI machine learning repository. Specifically, we modify the UCI datasets by keeping the input data and replacing the output with samples from the model likelihood $p(y \mid x, f_0)$, where $f_0 = f(\cdot; \theta_0)$ is sampled from the BNN prior:

$$\theta_0 \sim \pi_\theta, \quad y \mid x \sim p(y = \cdot \mid f(x; \theta_0)). \tag{3}$$

We will check whether an approximate posterior mean estimator, constructed from MALA samples, has a competitive *average-case* performance across randomly sampled θ_0. This will happen if weight-space MALA provides a reasonably accurate approximation to the function-space posterior, since the *exact* posterior mean estimator will minimize the average-case risk

$$\hat{f} \mapsto \mathbb{E}_{f_0 \sim \pi_f} \mathbb{E}_{\mathbf{Y} \sim p(\cdot \mid f_0(\mathbf{X}))} \mathbb{E}_{x_* \sim p_x, y_* \sim p(\cdot \mid f_0(x_*))} \ell(\hat{f}(x_*), y_*), \tag{4}$$

where ℓ denotes the loss function derived from the model likelihood. Therefore, competitive predictive performance of the MALA-approximated predictor will provide indirect evidence on the quality of posterior approximation.

We consider Gaussian and Laplacian likelihoods, which correspond to the square loss and the absolute error loss, respectively, and estimate (4) using 8 independently sampled θ_0. We use the feed-forward network architecture in Section 3.1 and vary $L \in \{2, 3\}, H \in \{50, 200\}$. We construct the approximate posterior mean predictor using 50 independent MALA chains, and compare its performance with an ensemble of 50 NN models trained with gradient descent (GD) using the MAP objective. For both MALA and GD, the step size is selected from $\{\eta/2nH : \eta \in \{1, 0.5, 0.1, 0.05, 0.01, 0.005\}\}$ such that the average acceptance rate of the first 200 MALA iterations is closest to 0.7, where n denotes the size of training set. We use 80% samples for training and 20% for testing.

We plot the estimated average-case loss in Fig. 3 and Fig. 5-6 in appendix, and report the best loss in Table 1-3. As we can see, across all settings, MALA leads to a similar predictive performance to the GD ensemble. As it is well known that GD methods perform well on DNN models, these results provide further evidence on the efficacy of the weight-space Langevin algorithm.

4 Conclusion

In this work we have investigated the function space behavior of weight-space Langevin-type algorithms on overparameterization models. Across multiple settings that encompass simplified BNN models, we have established the equivalence of the function-space pushforward of weight-space Langevin dynamics to its various function-space counterparts. Numerical studies on more realistic models provide further evidence of the possible equivalence.

References

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. *arXiv preprint arXiv:2008.05912*, 2020.

Jimmy Ba, Murat A Erdogdu, Marzyeh Ghassemi, Shengyang Sun, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Understanding the variance collapse of SVGD in high dimensions. In *International Conference on Learning Representations*, 2021.

Qinxun Bai, Steven Rosenberg, and Wei Xu. Understanding natural gradient in Sobolev spaces. *arXiv preprint arXiv:2202.06232*, 2022.

David R Burt, Sebastian W Ober, Adrià Garriga-Alonso, and Mark van der Wilk. Understanding variational inference in function-space. *arXiv preprint arXiv:2011.09421*, 2020.

José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53, 2019.

Changyou Chen, Ruiyi Zhang, Wenlin Wang, Bai Li, and Liqun Chen. A unified particle-optimization framework for scalable Bayesian sampling. *arXiv preprint arXiv:1805.11659*, 2018.

David Steven Dummit and Richard M Foote. *Abstract algebra*, volume 3. Wiley Hoboken, 2004.

Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural networks. *arXiv preprint arXiv:1906.08899*, 2019.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011.

Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:1010–1022, 2020.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning*, pp. 4629–4640. PMLR, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Arthur Jacot, Franck Gabriel, François Ged, and Clément Hongler. Order and chaos: NTK views on DNN normalization, checkerboard and boundary artifacts. *arXiv preprint arXiv:1907.05715*, 2019.

Pauwels JE. Riemannian submersions of Brownian motions. *Stochastics: An International Journal of Probability and Stochastic Processes*, 29(4):425–436, 1990.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Richard Kurle, Ralf Herbrich, Tim Januschowski, Yuyang Bernie Wang, and Jan Gasthaus. On the detrimental effect of invariances in the likelihood for variational inference. *Advances in Neural Information Processing Systems*, 35:4531–4542, 2022.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2012.

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

Kevin Luk and Roger Grosse. A coordinate-free construction of scalable natural gradient. *arXiv preprint arXiv:1808.10340*, 2018.

Chao Ma and José Miguel Hernández-Lobato. Functional variational inference based on stochastic process generators. *Advances in Neural Information Processing Systems*, 34:21795–21807, 2021.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In *International Conference on Machine Learning*, pp. 4222–4233. PMLR, 2019.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464. PMLR, 2019.

Ankur Moitra and Andrej Risteski. Fast convergence for Langevin diffusion with manifold structure, September 2020. arXiv:2002.05576 [cs, math, stat].

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Ichirô Satake. On a generalization of the notion of manifold. *Proceedings of the National Academy of Sciences*, 42(6):359–363, 1956.

Kanji Sato, Akiko Takeda, Reiichiro Kawai, and Taiji Suzuki. Convergence error analysis of reflected gradient Langevin dynamics for globally optimizing non-convex constrained problems. *arXiv preprint arXiv:2203.10215*, 2022.

Jiaxin Shi, Mohammad Emtiyaz Khan, and Jun Zhu. Scalable training of inference networks for Gaussian-process models. In *International Conference on Machine Learning*, pp. 5758–5768. PMLR, 2019.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Ziyu Wang, Tongzheng Ren, Jun Zhu, and Bo Zhang. Function space particle optimization for Bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świkatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for Bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019.

Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *International Conference on Machine Learning*, pp. 11513–11522. PMLR, 2020.

Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing Stein variational gradient descent. In *International Conference on Machine Learning*, pp. 6018–6027. PMLR, 2018.

A Proofs

A.1 Proof of Proposition 2.1

Proof of Proposition 2.1. By definitions, for any $f \in \mathcal{F}$, there exists some $\theta \in \mathbb{R}^d$ and one of its neighborhood N such that $f = \mathcal{A}\theta$, and that for $U = \mathcal{A}(N)$, $(U, \mathcal{A}|_N)$ forms a coordinate chart. On this chart, the coordinate matrix of the pushforward metric tensor equals identity, by its definition. Thus, the coordinate representation (FLD) reduces to

$$\mathrm{d}\theta_t = \nabla_\theta \left(\log p(\mathbf{Y} \mid \theta_t, \mathbf{X}) + \log \frac{d\pi_f}{d\mu_{\mathcal{F}}} \right) \mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

and it differs from (WLD) only on the prior term. When condition (a) in the proposition holds, the prior is uniform so the gradient vanishes. When condition (b) holds, the group is trivial and the quotient map $\mathcal{A}$ is a bijection. Thus, it suffices to show that for all $\theta \in \operatorname{supp} \pi_\theta$, we have

$$\frac{d\pi_f}{d\mu_{\mathcal{F}}}(\mathcal{A}\theta) = \frac{d\pi_\theta}{d\mu_{Leb}}(\theta) = p_\theta(\theta),$$

where μ_{Leb} denotes the Lebesgue measure. By the change of measure formula, the above will be implied by

$$\pi_f \overset{(i)}{=} \mathcal{A}_\# \pi_\theta, \quad \mu_{\mathcal{F}} \overset{(ii)}{=} \mathcal{A}_\# \mu_{Leb}.$$

(i) is the definition of π_f. For (ii), let $g : \mathcal{F} \to \mathbb{R}$ be any measurable function with a compact support, $\{(U_i = \mathcal{A}(N_i), \mathcal{A}|_{N_i}) : i \in [h]\}$ be a finite chart covering of $\operatorname{supp} g$, and $\{\rho_i\}$ be a corresponding partition of unity. Then

$$\int_{\mathcal{F}} g(f)\mu_{\mathcal{F}}(\mathrm{d}f) = \sum_{i=1}^{h} \int_{N_i} (\rho_i g)(\mathcal{A}(\theta))\sqrt{|G(\theta)|}\mu_{Leb}(\mathrm{d}\theta) = \int_{\mathcal{A}^{-1}(\operatorname{supp}\ g)} g(\mathcal{A}(\theta))\mu_{Leb}(\mathrm{d}\theta).$$

This establishes (ii), and thus completes the proof. □

A.2 Details in Example 2.3

Recall the definition of the cone $C_d := \{x \in \mathbb{R}^d : x_1 \le x_2 \le ... \le x_d\}$, and the group S_d that consists of all permutations of length d. An action of S_d on $\mathbb{R}^d$ can be naturally defined, under which we have $C_d = \mathbb{R}^d / S_d$.

We introduce a few additional notations. For $x \in \mathbb{R}^d$, the *stabilizer subgroup* is defined as $\operatorname{Stab}_{S_d} x := \{\varphi \in S_d : \varphi \cdot x = x\}$, and the orbit is $S_d \cdot x := \{\varphi \cdot x : \varphi \in S_d\}$. A vector $n_x \in \mathbb{R}^d$ is an *inward normal vector* of

C_d at x if $\langle n_x, y - x\rangle \geq 0$ holds for all $y \in C_d$. Denote by N_x the set of all inward normal vector of C_d at x. For any $f : \mathbb{R}^d \to \mathbb{R}$, define the function

$$\tilde{f} : C_d \to \mathbb{R}, \quad \tilde{f}(x) := \frac{1}{|S_d|} \sum_{\varphi \in S_d} f(\varphi \cdot x). \tag{5}$$

When f is the density function of a measure π on $\mathbb{R}^d$, the pushforward measure under the quotient map $\mathbb{R}^d \to C_d$ has the density function $\tilde{f}$. The following lemma shows that the directional derivative of $\tilde{f}$ along the normal direction vanishes.

Lemma A.1. *Let $x \in C_d$ and assume f is differentiable at every $y \in S_d \cdot x$. Then*

$$D_v \tilde{f}(x) = \frac{1}{|S_d|} \sum_{y := \psi \cdot x \in S_d \cdot x} D_{\psi \cdot W_x(v)} f(y), \qquad \text{where } \; W_x(v) := \sum_{\varphi \in \mathrm{Stab}\, x} \varphi \cdot v,$$

where D_v denotes the directional derivative along v. Moreover, $W_x(v) = v$ for $x \in C_n^\circ$ and $v \in \mathbb{R}^d$, and $W_x(v) = 0$ for $x \in \partial C_d$ and $v \in N_x$.

We postpone the proof of the above lemma to the end of this section, and first present the following lemma, which implies the invariance of the Fokker-Planck equation under orthogonal transformations.

Lemma A.2. *Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be two functions and $Q \in \mathbb{R}^{d\times d}$ be an orthogonal matrix, then $[\nabla(f \circ Q)]^T \nabla(g \circ Q) = [(\nabla f)^T \nabla g] \circ Q$ and $\Delta(f \circ Q) = \Delta f \circ Q$, in which Q is also regarded as a linear map $Q : \mathbb{R}^d \to \mathbb{R}^d$.*

Proof. Note that $\nabla(f \circ Q) = Q^T(\nabla f \circ Q)$. Let Q_i be the i-th column of Q, then

$$[\nabla(f \circ Q)]^T \nabla(g \circ Q) = \sum_{i=1}^d (\nabla f \circ Q)^T Q_i Q_i^T (\nabla g \circ Q) = (\nabla f \circ Q)^T (\nabla g \circ Q).$$

A similar result also holds for the Laplacian:

$$\Delta(f \circ Q) = \sum_{i=1}^d \partial_i \partial_i (f \circ Q) = \sum_{i,j=1}^d \partial_i (\partial_j f \circ Q) q_{ji} = \sum_{i,j,k=1}^d (\partial_k \partial_j f \circ Q) q_{ji} q_{ki}.$$

As Q is orthogonal, we know $\sum_{i=1}^d q_{ji} q_{ki} = \delta_{jk}$, which completes the proof. □

As the pushforward measure $\mathcal{A}_\# p$ has density $\tilde{p}$, the following proposition establishes the equivalence result claimed in the text.

Proposition A.1. *Let $p : \mathbb{R}^d \to \mathbb{R}$ be any function that is invariant under the action of S_d, and X_t follow the Langevin dynamics on $\mathbb{R}^d$,*

$$\mathrm{d}X_t = \nabla \log p(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

Then, the pushforward density $\tilde{p}_t$ of X_t will evolve as

$$\begin{cases} \partial_t \tilde{p}_t = -\nabla \cdot (\tilde{p}_t \nabla \log p) + \Delta \tilde{p}_t, & \text{in } C_n^\circ, \\ \frac{\partial \tilde{p}_t}{\partial v}(x) = 0, & \forall v \in N_x, x \in \partial C_d. \end{cases}$$

Proof. Let p_t be the density of the distribution of X_t, then it follows the Fokker-Planck equation

$$\partial_t p_t = -\nabla \cdot (p_t \nabla \log p) + \Delta p_t = -(\nabla p_t)^T \nabla \log p - p_t \Delta \log p + \Delta p_t.$$

For $\varphi \in S_d$, we denote $P_\varphi \in \mathbb{R}^{d\times d}$ by the corresponding matrix such that $\varphi \cdot x = P_\varphi x$ for every $x \in \mathbb{R}^d$. Then, P_φ is an orthogonal matrix, and by Lemma A.2

$$\begin{aligned} \partial_t (p_t \circ P_\varphi) &= (\partial_t p_t) \circ P_\varphi = -(\nabla p_t \cdot \nabla \log p) \circ P_\varphi - (p_t \Delta \log p) \circ P_\varphi + (\Delta p_t) \circ P_\varphi \\ &= -[\nabla(p_t \circ P_\varphi)]^T \nabla \log p - (p_t \circ P_\varphi) \Delta \log p + \Delta(p_t \circ P_\varphi), \end{aligned}$$

where the first equation is because P_φ is independent to t, and the last equation follows from Lemma A.2 and $\log p \circ P_\varphi = \log p$.

Therefore, we obtain the equation for $\tilde{p}_t$:

$$\begin{aligned}\partial_t \tilde{p}_t = \frac{1}{|S_d|}\sum_{\varphi\in S_d}\partial(p_t\circ P_\varphi) &= \frac{1}{|S_d|}\sum_{\varphi\in S_d}\left(-\nabla\cdot((p_t\circ P_\varphi)\nabla\log p) + \Delta(p_t\circ P_\varphi)\right)\\ &= -\nabla\cdot(\tilde{p}_t\nabla\log p) + \Delta\tilde{p}_t. \end{aligned} \tag{6}$$

Combining with Lemma A.1 yields the boundary condition

$$\frac{\partial\tilde{p}_t}{\partial v}(x) = 0, \quad \forall v\in N_x, x\in\partial C_d. \tag{7}$$

□

Proof of Lemma A.1. Since the group action is linear (i.e., $\varphi\cdot(x+y) = \varphi\cdot x + \varphi\cdot y$ and $\varphi\cdot(tx) = t\varphi\cdot x$), we have

$$D_v\tilde{f}(x) = \lim_{t\to 0+}\frac{1}{t}\left(\tilde{f}(x+tv) - \tilde{f}(x)\right) = \frac{1}{|S_d|}\sum_{\varphi\in S_d} D_{\varphi\cdot v}f(\varphi\cdot x).$$

To simplify the above summation, we introduce the coset $\varphi\operatorname{Stab}x := \{\varphi\psi : \psi\in\operatorname{Stab}x\}$ for each $\varphi\in S_d$, and the set of cosets $S_d/\operatorname{Stab}x := \{\varphi\operatorname{Stab}x : \varphi\in S_d\}$. Clearly, any two cosets are either equal or disjoint, and the group S_d is partitioned by $S_d/\operatorname{Stab}x$. The orbit-stabilizer theorem (Dummit & Foote, 2004, p. 114) states that the map $\varphi\operatorname{Stab}x\mapsto\varphi\cdot x$ is a bijection between cosets $S_d/\operatorname{Stab}x$ and the orbit $S_d\cdot x$, and thus[3]

$$\begin{aligned}
D_v\tilde{f}(x) &= \frac{1}{|S_d|}\sum_{\varphi\in S_d} D_{\varphi\cdot v}f(\varphi\cdot x)\\
&= \frac{1}{|S_d|}\sum_{\substack{\varphi\in C\\ C=\psi\operatorname{Stab}x\in S_d/\operatorname{Stab}x}} D_{\varphi\cdot v}f(\varphi\cdot x) && \text{(partition)}\\
&= \frac{1}{|S_d|}\sum_{\substack{\varphi\in\psi\operatorname{Stab}x\\ y:=\psi\cdot x\in S_d\cdot x}} D_{\varphi\cdot v}f(\varphi\cdot x) && (\psi\operatorname{Stab}x\mapsto\psi\cdot x \text{ bijective})\\
&= \frac{1}{|S_d|}\sum_{y:=\psi\cdot x\in S_d\cdot x}\sum_{\varphi'\in\operatorname{Stab}x} D_{\psi\cdot(\varphi'\cdot v)}f(y) && (\varphi' := \psi^{-1}\varphi)\\
&= \frac{1}{|S_d|}\sum_{y:=\psi\cdot x\in S_d\cdot x} D_{\psi\cdot W_x(v)}f(y). && (\text{linearity of } D_{(\cdot)}f)
\end{aligned}$$

This proves the first claim.

For any interior point $x\in C_d^\circ$, we have $\operatorname{Stab}x = \{e\}$ and thus $W_x(v) = v$. For any boundary point $x\in\partial C_d$, the stabilizer subgroup is non-trivial, and it remains to show that $W_x(v) = 0$ for normal vectors.

An element $\varphi\in S_d$ can be identified as a permutation matrix $P_\varphi\in\mathbb{R}^{d\times d}$ s.t. the group action is the matrix-vector multiplication $\varphi\cdot v = P_\varphi v$, and clearly, the stabilizer of $x\in\partial C_d$ always has the form of a Cartesian product, $\prod_{j=1}^{m_x} S_{c_j}$, where $\{c_j\}$ is s.t. $\sum_{j=1}^{m_x} c_j = d$.[4] Therefore, we have

$$W_x(v) = \sum_{\varphi\in\operatorname{Stab}x}\varphi\cdot v = \left(\sum_{\varphi\in\prod_{j=1}^{m_x} S_{c_j}} P_\varphi\right)v.$$

[3]It can be verified that the proof is independent on the choice of ψ.

[4]For example, for $x\in C_5$ with $x_1 = x_2 < x_3 = x_4 < x_5$, the stabilizer is $S_2\times S_2\times S_1$.

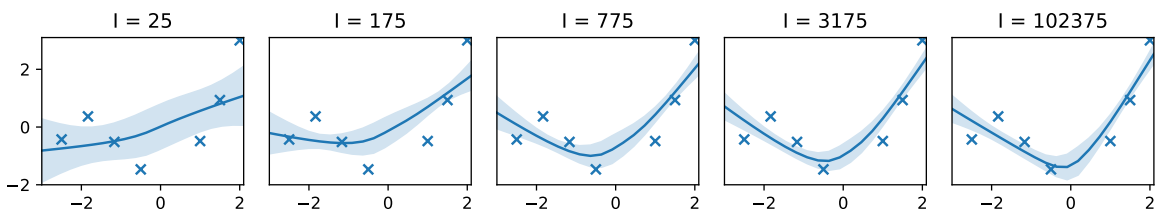
Note that $P_\varphi = \text{blkdiag}(P_1, P_2, \ldots, P_{m_x})$, with each $P_j \in \mathbb{R}^{c_j \times c_j}$ being a permutation matrix, and the sum of all size c_j permutation matrices is $(c_j - 1)!\mathbf{1}_{c_j \times c_j}$, where $\mathbf{1}$ denotes the all-ones matrix. Thus, by decomposing $W_x(v) \in \mathbb{R}^d$ into $\mathbb{R}^{c_1} \times \mathbb{R}^{c_2} \times \cdots \times \mathbb{R}^{c_{m_x}}$ we have

$$W_x(v) \;=\; \left(\frac{A_0}{c_1} \sum_{i=s_0+1}^{s_1} v_i \mathbf{1}_{c_1}, \;\; \frac{A_0}{c_2} \sum_{i=s_1+1}^{s_2} v_i \mathbf{1}_{c_2}, \;\; \ldots, \;\; \frac{A_0}{c_{m_x}} \sum_{i=s_{m_x-1}+1}^{s_{m_x}} v_i \mathbf{1}_{c_{m_x}} \right),$$

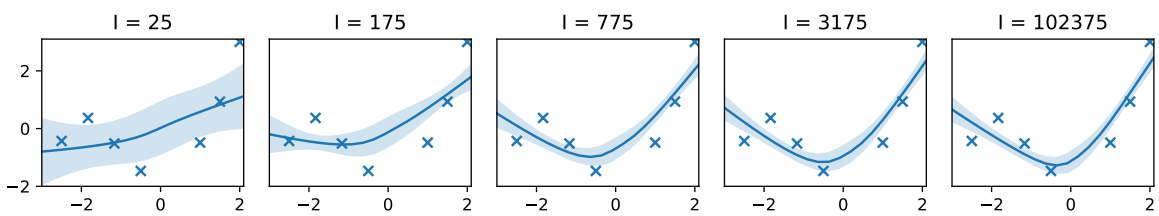
where $A_0 = \prod_{j=1}^{m_x} c_j!$ and $s_j = \sum_{l \le j} c_l$.

Let $e^{(j)} \in \mathbb{R}^d$ be such that $e_k^{(j)} = 1$ if $s_{j-1} < k \le s_j$, and $e_k^{(j)} = 0$ otherwise. Then a sufficient condition for $W_x(v) = 0$ is that $\langle v, e^{(j)} \rangle = 0$ for all $j \in [m_x]$. Let $n_x \in N_x$ be an inward normal vector and fix $j \in [m_x]$. Since $x \pm \alpha_j e^{(j)} \in C_d$ for $\alpha_j = \min(x_{s_j} - x_{s_{j-1}}, x_{s_{j+1}} - x_{s_j}) > 0$, we conclude that $\langle n_x, \pm e^{(j)} \rangle \ge 0$ and hence $\langle n_x, e^{(j)} \rangle = 0$. Thus, $W_x(n_x) = 0$. □

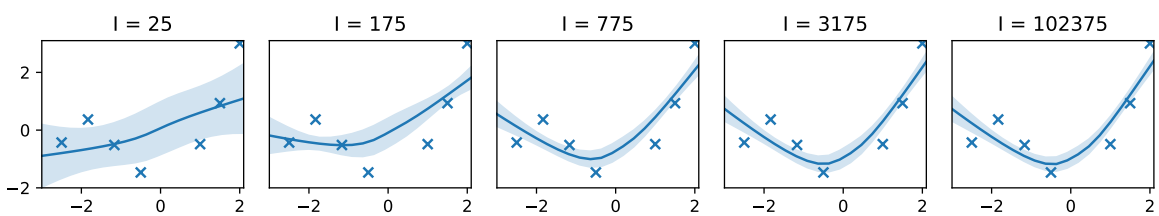
B Additional Results



(a) $L = 2, H = 20$

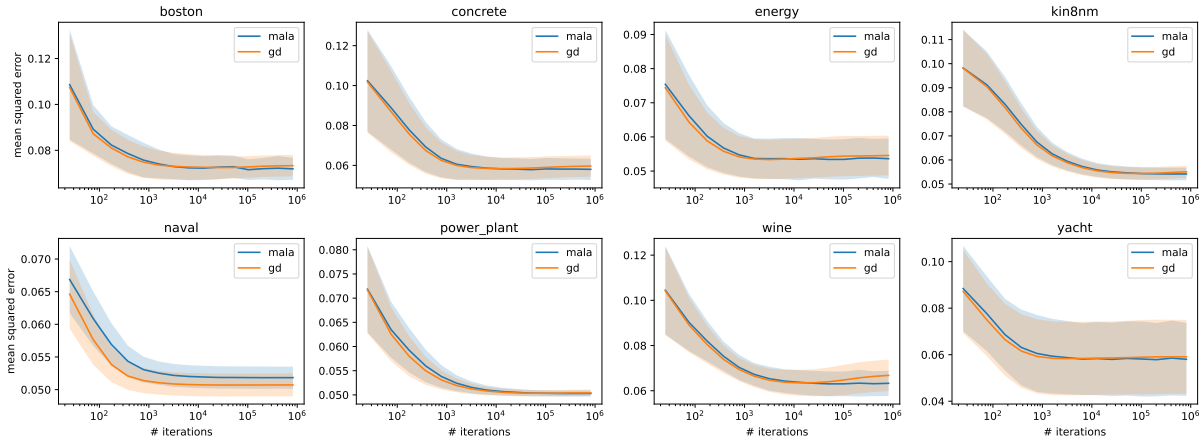


(b) $L = 2, H = 50$

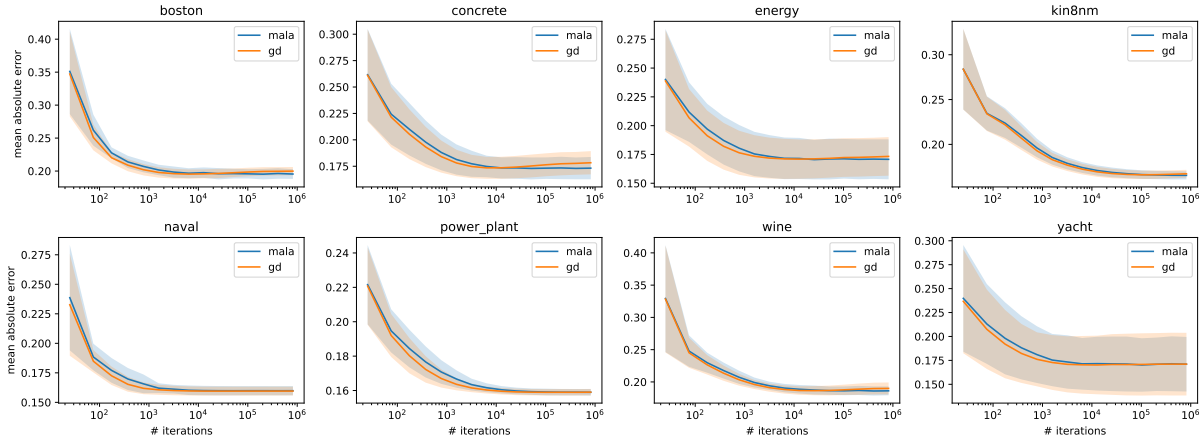


(c) $L = 2, H = 500$

Figure 4: Additional visualizations in the setting of Fig. 1.



(a) Gaussian likelihood / mean square error



(b) Laplace likelihood / mean absolute error

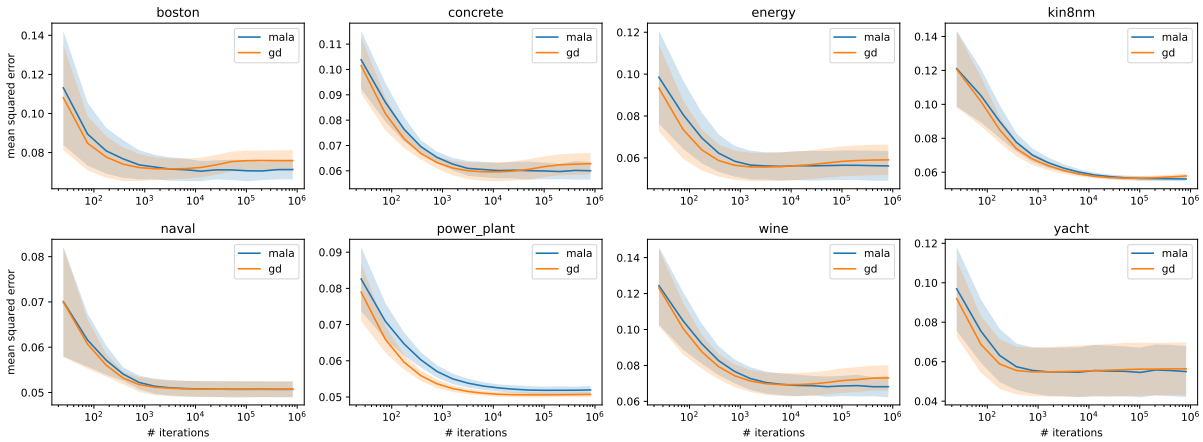
Figure 5: Semi-synthetic experiment: estimated loss (4) under different likelihoods, for $H = 2, L = 50$.

Table 1: Semi-synthetic experiment: average-case test risk for the best stopping iteration, for $H = 2, L = 200$.

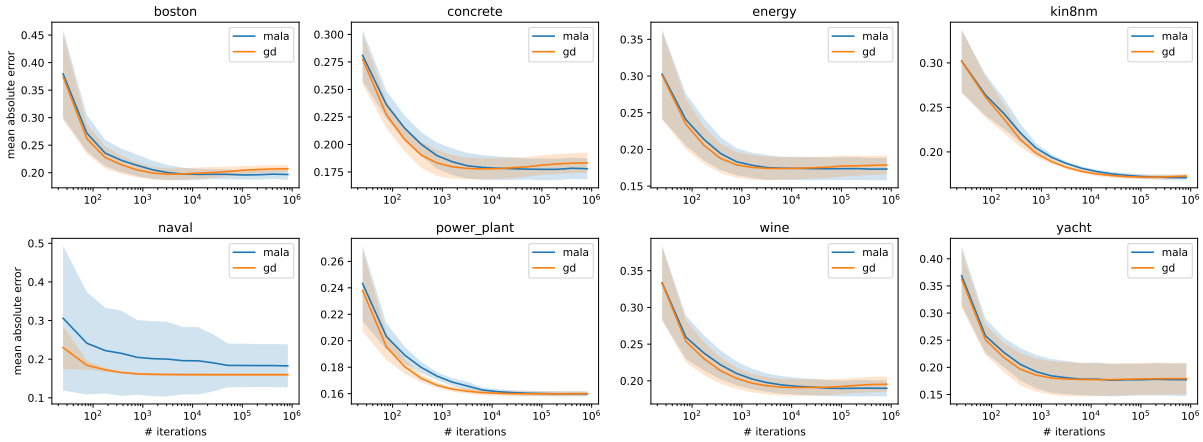
Likelihood	**Algorithm**	**boston**	**concrete**	**energy**	**kin8nm**	**naval**	**power plant**	**wine**	**yacht**
Gaussian	MALA	0.067	0.058	0.056	0.055	0.051	0.050	0.063	0.055
	GD	0.068	0.059	0.056	0.055	0.051	0.050	0.063	0.056
Laplacian	MALA	0.190	0.175	0.167	0.168	0.159	0.159	0.185	0.172
	GD	0.191	0.176	0.168	0.169	0.159	0.159	0.185	0.173

Table 2: Semi-synthetic experiment: average-case test risk for the best stopping iteration, for $H = 2, L = 50$.

Likelihood	**Algorithm**	**boston**	**concrete**	**energy**	**kin8nm**	**naval**	**power plant**	**wine**	**yacht**
Gaussian	MALA	0.071	0.058	0.053	0.054	0.052	0.050	0.063	0.057
	GD	0.071	0.058	0.053	0.054	0.051	0.050	0.063	0.057
Laplacian	MALA	0.193	0.172	0.170	0.165	0.160	0.159	0.185	0.168
	GD	0.195	0.173	0.170	0.166	0.159	0.159	0.186	0.169



(a) Gaussian likelihood / mean square error



(b) Laplace likelihood / mean absolute error

Figure 6: Semi-synthetic experiment: estimated loss (4) under different likelihoods, for $H = 3, L = 50$.

Table 3: Semi-synthetic experiment: average-case test risk for the best stopping iteration, for $H = 3, L = 50$.

Likelihood	**Algorithm**	**boston**	**concrete**	**energy**	**kin8nm**	**naval**	**power plant**	**wine**	**yacht**
Gaussian	MALA	0.069	0.059	0.055	0.056	0.051	0.052	0.068	0.053
	GD	0.070	0.059	0.056	0.056	0.051	0.051	0.069	0.054
Laplacian	MALA	0.194	0.176	0.172	0.171	0.183	0.160	0.189	0.175
	GD	0.197	0.177	0.173	0.172	0.160	0.160	0.190	0.175