On Equivalences between Weight and Function-Space Langevin Dynamics

Ziyu Wang
Tsinghua University
wzy196@gmail.com

Yuhao Zhou
Tsinghua University
yuhaoz.cs@gmail.com

Ruqi Zhang
Purdue University
ruqiz@purdue.edu

Jun Zhu
Tsinghua University
dcszj@mail.tsinghua.edu.cn

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 "function space", the space of prediction functions. This note 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 $\pi_\theta(d\theta)$. We are interested in the predictive distribution $p(y_* \mid x_*, \mathcal{D}) = \int \pi_{\theta|\mathcal{D}}(d\theta) p(y_* \mid x_*, \theta)$, 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

$$
\begin{aligned}
p(y_* \mid x_*, \mathcal{D}) = \int \pi_{\theta|\mathcal{D}}(d\theta)\, p(y_* \mid f(x_*; \theta)) &= \int (\mathcal{A}_\# \pi_{\theta|\mathbf{X},\mathbf{Y}})(df)\, p(y_* \mid f(x_*)) \\
&= \int \pi_{f|\mathcal{D}}(df)\, p(y_* \mid f(x_*)),
\end{aligned}
$$

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

I Can't Believe It's Not Better Workshop at NeurIPS 2022.

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 and 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 been applied to a broad range of models. Their convergence behaviors are well-understood (Roberts and Tweedie, 1996; Villani, 2009), and intriguingly, their performance often appear to be satisfying on massively overparameterized models (Zhang et al., 2019; Izmailov et al., 2021), even though they are implemented in weight space.

This note bridges the two lines of approach 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 and Risteski (2020) establishes such results for a family of locally overparameterized models, which covers matrix sensing problems. 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. While we are unaware of strict equivalence results as provided in this note, 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 sometimes 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 it provides 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, it complements 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 Equivalences between Weight and Function-Space Langevin Dynamics

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

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

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 lead to a posterior, $\pi_{f|\mathcal{D}}$. When the function space $\mathcal{F} := \operatorname{supp} \pi_f$ can be equipped with a Riemannian manifold structure, it is intuitive that we could sample from $\pi_{f|\mathcal{D}}$ by simulating a Riemannian Langevin dynamics on $\mathcal{F}$. In coordinate form:

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

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

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

$\mu_{H,\mathcal{F}}$ denotes the corresponding Hausdorff / 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 $\pi_\theta = N(0, I)$, 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*

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

(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.*

(ii) 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.

The following claim shows that linearity of $\mathcal{A}$ is not necessary in establishing the equivalence.

Claim 2.1 (nonlinear parameterization, proof in Appendix A.1)**.** *Suppose $\mathcal{A}$ is differentiable, and*

1. *$\mathcal{A}$ is bijective; or*
2. *$\mathcal{A}$ is locally bijective, the priors have constant densities, and the "pushforward" of the weight-space metric can be consistently defined globally; see App. A.1 for precise statements.*

Then the equivalence between (WLD) *and* (FLD) *continues to hold.*

As we will show in Example 2.3, Claim 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. 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 2.2 (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 $\mathrm{supp}\,\pi_\theta$ be contained in $(0, +\infty)^d$, and assume the technical conditions (A1)-(A2) in Appendix A.2.*

By the linear independence of sin functions, for $\mathcal{A}\theta = \mathcal{A}\theta'$, θ' must be a permutation of θ. Thus, the weight space is partitioned into $d!$ polyhedral cones*, based on the ordering of the components $(\theta_1, \ldots, \theta_d)$. Within any cone, the restriction of $\mathcal{A}$ is bijective; thus by Claim 2.1,* (WLD) *will be equivalent to* (FLD) *(with the pushforward metric)* before *θ_t hits any boundary between cones. However, the crossing of boundaries requires extra analyses, since any function with parameters on the boundary cannot be included in the function-space manifold.*

Appendix A.2 proves the equivalence between discretized versions of (WLD) *and a reflected LD algorithm (Sato et al., 2022) in weight space,* where θ is constrained to the initial cone. *Since the restriction of $\mathcal{A}$ to any single cone is a bijection, and the resulted pushforward metric is independent of the choice of cones (see appendix for proof), we may* view[1] *the reflected LD procedure as a reflected LD procedure in function space, in light of the equivalence result in Claim 2.1.*

Discussion. 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 and 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 2.3 (LD and 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 Claim 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, $\tilde{v}_t(\theta)dt$, 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 $\tilde{v}_t dt$ 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 $\nabla_\theta \log p(\mathbf{Y} \mid \mathbf{X}, \theta^*)p(\theta^*) = 0$. 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 work have noted the pathologies of particle-based VI in high dimensions (Zhuo et al., 2018; Ba et al., 2021); this example is interesting as it does not require an increasing dimensionality. Rather, it is global overparameterization that breaks the asymptotic convergence to LD.

[1] It appears unnecessary to introduce the formalism of reflected LD on manifolds.

3 Simulation Study

We now validate our findings on practical BNN models. We 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, we will

1. 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,

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

   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.

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

$$f^{(l)}(h^{(l-1)}) := W^{(l)} h^{(l-1)} + b^{(l)}, \;\; \mathrm{vec}(W^{(l)}) \sim \mathcal{N}\left(0, (\dim h^{(l-1)})^{-1} I\right), \;\; b^{(l)} \sim \mathcal{N}(0, 0.1I).$$

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. 3, 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.

References

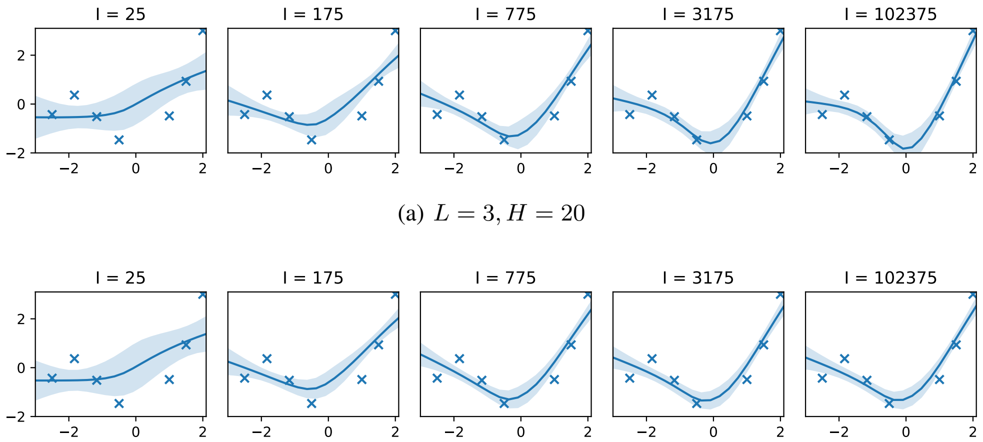
Aitchison, L. (2020). A statistical theory of cold posteriors in deep neural networks. *arXiv preprint arXiv:2008.05912*.

Ba, J., Erdogdu, M. A., Ghassemi, M., Sun, S., Suzuki, T., Wu, D., and Zhang, T. (2021). Understanding the variance collapse of svgd in high dimensions. In *International Conference on Learning Representations*.

Bai, Q., Rosenberg, S., and Xu, W. (2022). Understanding natural gradient in sobolev spaces. *arXiv preprint arXiv:2202.06232*.

Burt, D. R., Ober, S. W., Garriga-Alonso, A., and van der Wilk, M. (2020). Understanding variational inference in function-space. *arXiv preprint arXiv:2011.09421*.

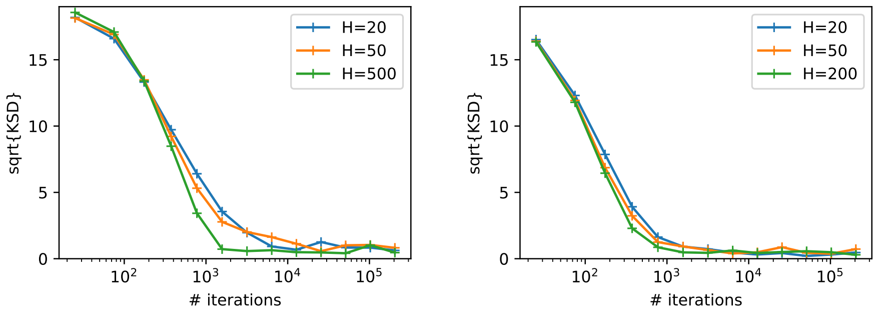
Carrillo, J. A., Craig, K., and Patacchini, F. S. (2019). A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53.



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

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

Figure 1: 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. 3.



(a) $L = 2$

(b) $L = 3$

Figure 2: 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.

Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. (2018). A unified particle-optimization framework for scalable bayesian sampling. *arXiv preprint arXiv:1805.11659*.

Fortuin, V., Garriga-Alonso, A., Wenzel, F., Rätsch, G., Turner, R., van der Wilk, M., and Aitchison, L. (2021). Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. (2019). Limitations of lazy training of two-layers neural networks. *arXiv preprint arXiv:1906.08899*.

He, B., Lakshminarayanan, B., and Teh, Y. W. (2020). Bayesian deep ensembles via the neural tangent kernel. *Advances in neural information processing systems*, 33:1010–1022.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. G. (2021). What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR.

Jacot, A., Gabriel, F., Ged, F., and Hongler, C. (2019). Order and chaos: Ntk views on dnn normalization, checkerboard and boundary artifacts. *arXiv preprint arXiv:1907.05715*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Kurle, R., Herbrich, R., Januschowski, T., Wang, Y., and Gasthaus, J. (2022). On the detrimental effect of invariances in the likelihood for variational inference. arXiv:2209.07157 [cs].

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018). Deep neural networks as gaussian processes. In *International Conference on Learning Representations*.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32.

Luk, K. and Grosse, R. (2018). A coordinate-free construction of scalable natural gradient. *arXiv preprint arXiv:1808.10340*.

Ma, C. and Hernández-Lobato, J. M. (2021). Functional variational inference based on stochastic process generators. *Advances in Neural Information Processing Systems*, 34:21795–21807.

Ma, C., Li, Y., and Hernández-Lobato, J. M. (2019). Variational implicit processes. In *International Conference on Machine Learning*, pages 4222–4233. PMLR.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

Mei, S., Misiakiewicz, T., and Montanari, A. (2019). Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR.

Moitra, A. and Risteski, A. (2020). Fast Convergence for Langevin Diffusion with Manifold Structure. arXiv:2002.05576 [cs, math, stat].

Osband, I., Aslanides, J., and Cassirer, A. (2018). Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Sato, K., Takeda, A., Kawai, R., and Suzuki, T. (2022). Convergence error analysis of reflected gradient langevin dynamics for globally optimizing non-convex constrained problems. *arXiv preprint arXiv:2203.10215*.

Shi, J., Khan, M. E., and Zhu, J. (2019). Scalable training of inference networks for gaussian-process models. In *International Conference on Machine Learning*, pages 5758–5768. PMLR.

Sun, S., Zhang, G., Shi, J., and Grosse, R. (2019). Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, Z., Ren, T., Zhu, J., and Zhang, B. (2019). Function space particle optimization for bayesian neural networks. In *International Conference on Learning Representations*.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32.

Wenzel, F., Roth, K., Veeling, B. S., Świkatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*.

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

Zhang, R., Li, C., Zhang, J., Chen, C., and Wilson, A. G. (2019). Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*.

Zhou, Y., Shi, J., and Zhu, J. (2020). Nonparametric score estimators. In *International Conference on Machine Learning*, pages 11513–11522. PMLR.

Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N., and Zhang, B. (2018). Message passing stein variational gradient descent. In *International Conference on Machine Learning*, pages 6018–6027. PMLR.

A Proofs

A.1 Proof for Claim 2.1

Claim A.1 (Claim 2.1, restated)**.** *Suppose $\mathcal{A}$ is differentiable, and one of the following holds:*

1. $\mathcal{A}$ is bijective;

2. $\mathcal{A}$ is locally bijective, and that

(a) for any $\theta, \theta' \in \operatorname{supp} \pi_\theta$ s.t. $\mathcal{A}\theta = \mathcal{A}\theta' = f$, and their neighborhoods N, N' s.t. the restrictions $\mathcal{A}|_N, \mathcal{A}_{N'}$ are bijective, the pushforward of the local Euclidean metric by $\mathcal{A}|_N$ and $\mathcal{A}|_{N'}$ at f are equivalent;

(b) both $\frac{d\pi_\theta}{d\mu_{Leb}}$ and $\frac{d\pi_f}{d\mu_{H,\mathcal{F}}}$ are constant functions.

Let θ_t follow (WLD), and $f_t^{wld} := \mathcal{A}(\theta_t)$. Then f_t^{wld} follows the same SDE as (FLD), defined with the pushforward metric.

Note that in the locally bijective case, by condition (b) a metric can be consistently defined by considering the pushforward of restrictions of $\mathcal{A}$, to any subset of its preimage on which it becomes bijective. With slight abuse of notion, we refer to this metric as the pushforward of $\mathcal{A}$.

Proof. By definitions, for any $f \in \mathcal{F}$, there exists $\theta \in \operatorname{supp} \pi_\theta$ and one of its neighborhood N, s.t. $f = \mathcal{A}\theta$, 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

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

The above equation only differs from (WLD) in the prior term. In the locally bijective case both prior terms have gradient zero, and thus the proof is complete; for the bijective case, it suffices to show that for all $\theta \in \operatorname{supp} \pi_\theta$, we have

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

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

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

(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

$$\begin{aligned}\int_{\mathcal{F}} g(f)\mu_{H,\mathcal{F}}(df) &= \sum_{i=1}^{h} \int_{N_i} (\rho_i g)(\mathcal{A}(\theta))\sqrt{|G(\theta)|}\mu_{Leb}(d\theta) \\ &= \int_{\mathcal{A}^{-1}(\operatorname{supp} g)} g(\mathcal{A}(\theta))\mu_{Leb}(d\theta).\end{aligned}$$

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

A.2 Proof in Example 2.2

Let $\hat{\theta}_{ld,k}$ be the trajectory of the forward Euler discretization of (WLD) with step-size η:

$$\hat{\theta}_{ld,k+1} := \mathsf{LDUpdate}(\hat{\theta}_{ld,k}, Z_k) := \hat{\theta}_{ld,k} + \eta \nabla_\theta (\log p(\mathbf{Y} \mid \hat{\theta}_{ld,k}) + \log p_\theta(\hat{\theta}_{ld,k})) + \sqrt{2\eta} Z_k. \quad (2)$$

Following Sato et al. (2022), we will use bounded, rotationally invariant independent random variables as Z_k, to avoid the rare large jumps; this is known to not affect weak convergence.

For any permutation $r = \{r_i\}$, define the cone

$$\Theta_r := \{\theta \in (0, +\infty)^d : \text{for all } i \le d, \theta_i \text{ has the } r_i\text{-th largest value}\}.$$

Clearly any $\theta \in \Theta_r$ will have distinct coordinates. Let $\mathcal{P}_r : \mathbb{R}^d \to \mathbb{R}^d$ be the map that permutes the components of its input so that it falls into Θ_r. We restrict to the probability-1 event where for all $k \le K$, there exists some permutation r_k s.t. $\hat{\theta}_{ld,k} \in \Theta_{r_k}$. Then the operator $\mathcal{P}$ introduced in the text equals $\mathcal{P}_{r_0}$. We further impose the assumption that

(A1) The model and likelihood is such that, for any fixed $L > 0$, as $\eta \to 0$, we have, with probability approaching 1, $\forall k \le \lfloor L/\eta \rfloor$, r_k and r_{k+1} differs by at most two elements.

In words, we want all $\hat{\theta}_{ld,k+1}$ to fall into an adjacent cone to that of $\hat{\theta}_{ld,k}$. This holds for continuous time random walk (i.e., LD with no likelihood), and thus is natural to expect; we pose it as an assumption for simplicity. We restrict to this event in the following. We also assume that

(A2) The prior density is symmetric, i.e., $p_\theta = p_\theta \circ \mathcal{P}_r$ for all r; moreover, as $\theta \to \theta_0$ where the vector θ_0 contains at least one zero, $\log p_\theta(\theta) \to -\infty$ at a rate sufficiently fast.

The latter condition ensures that when η is sufficiently small, (2) will never leave the prior support.

$\mathcal{F}$ and its metric structure We define $\mathcal{F}$ as the image of $\mathcal{A}$ restricted to any Θ_r. Clearly, $\mathcal{F}$ is consistently defined as a differentiable manifold. We can equip $\mathcal{F}$ with the pushforward of the Euclidean metric by $\mathcal{A}|_{\Theta_r}$, for any fixed r. Here we show that all such choices are equivalent; thus, similar to Claim 2.1, we can refer to the "pushforward of $\mathcal{A}$" in an unambiguous way.

This is because the metric at $\mathcal{A}\theta$ can be identified through the function $\mathcal{K}_{\mathcal{A}\theta}(x, x') = \sum_{i=1}^d xx' \cos(\theta_i x) \cos(\theta_i x') =: xx' \int \rho_\theta(da) \cos(ax) \cos(ax')$, where $\rho_\theta := \sum_{i=1}^d \delta_{\theta_i}(\cdot)$ is a discrete measure associated with θ. By the linear independence of sinusoid functions (and our restriction $\theta_i > 0$), $\mathcal{A}\theta = \mathcal{A}\theta'$ will imply $\rho_\theta = \rho_{\theta'}$, which in turn implies $\mathcal{K}_{\mathcal{A}\theta} \equiv \mathcal{K}_{\mathcal{A}\theta'}$. This proves the equivalence of choices.

Proof for the equivalence

1. $\{\hat{\theta}_{ld,k}\}$ and $\{\mathcal{P}(\hat{\theta}_{ld,k})\}$ produces equivalent predictions: we have $\mathcal{A}(\mathcal{P}\theta) = \mathcal{A}\theta$ for all θ.
2. Let $\hat{\theta}_{P,k} := \mathcal{P}(\hat{\theta}_{ld,k})$. We now claim that

$$\hat{\theta}_{P,k+1} = \mathcal{P}(\hat{\theta}_{PU,k+1}), \;\; \hat{\theta}_{PU,k+1} := \mathsf{LDUpdate}(\hat{\theta}_{P,k}, \tilde{Z}_k), \quad (3)$$

where $\tilde{Z}_k$ denote a permuted version of Z_k. When $r_{k+1} = r_k$, (3) immediately follows from the fact that, by symmetry, we have

$$\mathcal{P}_r(\mathsf{LDUpdate}(\theta, Z)) = \mathsf{LDUpdate}(\mathcal{P}_r(\theta), \mathcal{P}_r(Z)) \quad (4)$$

for all r and θ. Otherwise, observe that

$$\begin{aligned}\hat{\theta}_{P,k+1} &= \mathcal{P}(\hat{\theta}_{ld,k+1}) = \mathcal{P}(\mathcal{P}_{r_k}(\mathsf{LDUpdate}(\hat{\theta}_{ld,k}, Z_k))) = \mathcal{P}(\mathcal{P}_0(\mathsf{LDUpdate}(\hat{\theta}_{ld,k}, Z_k))) \\ &\overset{(4)}{=} \mathcal{P}(\mathsf{LDUpdate}(\mathcal{P}_0\hat{\theta}_{ld,k}, \mathcal{P}_0 Z_k)) = \mathcal{P}(\mathsf{LDUpdate}(\hat{\theta}_{P,k}, \mathcal{P}_0 Z_k)).\end{aligned}$$

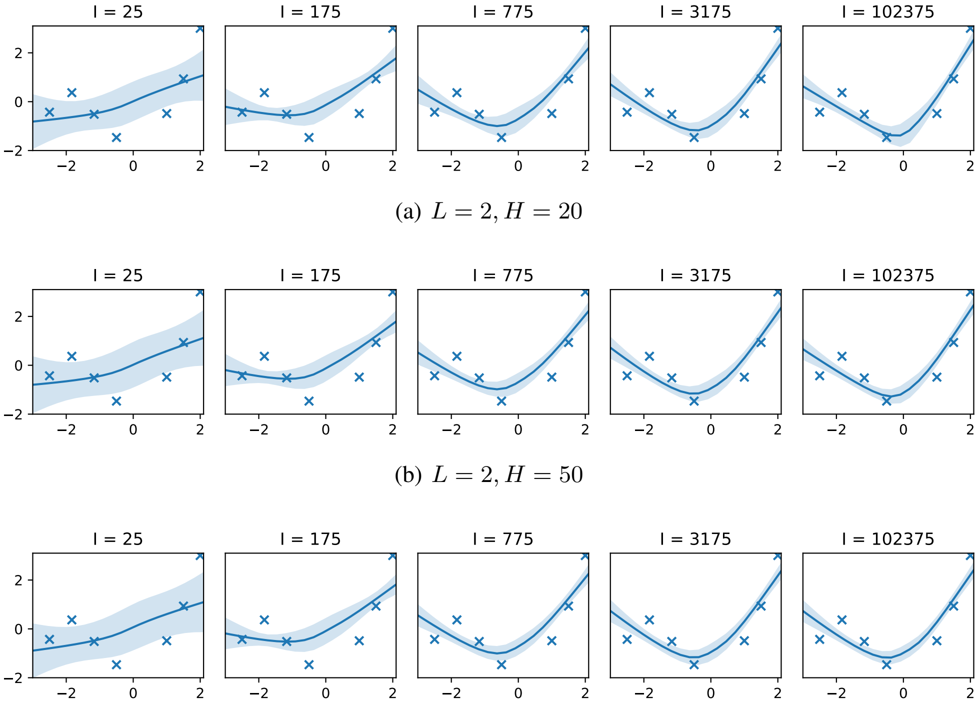
This proves (3). Note that $\tilde{Z}_k$ is still independent of $Z_{<k}$, and identically distributed as Z_k. Thus, $\{\tilde{Z}_k : k \in \mathbb{N}\}$ provide a valid source of driving noise in (2).

3. It remains to show that $\{\hat{\theta}_{P,k}\}$ is equivalent to the particle trajectory of the reflected gradient Langevin dynamics algorithm in Sato et al. (2022), which is defined as

$$\hat{\theta}_{R,k+1} = \mathcal{R}(\hat{\theta}_{RU,k+1}), \;\; \hat{\theta}_{RU,k+1} := \mathsf{LDUpdate}(\hat{\theta}_{R,k}, \tilde{Z}_k)$$

where $\mathcal{R}$ denotes the symmetric reflection operation: $\mathcal{R}(\theta) := 2\bar{\theta} - \theta$, where $\bar{\theta}$ denotes the projection of θ to the constraint set Θ_{r_0}. Thus, to show the equivalence between $\{\hat{\theta}_{R,k}\}$ to $\{\hat{\theta}_{P,k}\}$, we only need to show the equivalence of $\mathcal{P}$ and $\mathcal{R}$. This can be easily verified given (A1).

B Additional Results



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

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

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

Figure 3: Additional visualizations in the setting of Fig. 1.