MULTISCALE INVERTIBLE GENERATIVE NETWORKS FOR HIGH-DIMENSIONAL BAYESIAN INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional Bayesian inference problems cast a long-standing challenge in generating samples, especially when the posterior has multiple modes. For a wide class of Bayesian inference problems equipped with the *multiscale structure* that low-dimensional (coarse-scale) surrogate can approximate the original high-dimensional (fine-scale) problem well, we propose to train a Multiscale Invertible Generative Network (MsIGN) for sample generation. A novel prior conditioning layer is designed to bridge networks at different resolutions, enabling coarse-to-fine multi-stage training. Jeffreys divergence is adopted as the training objective to avoid mode dropping. On two high-dimensional Bayesian inverse problems, MsIGN approximates the posterior accurately and clearly captures multiple modes, showing superior performance compared with previous deep generative network approaches. On the natural image synthesis task, MsIGN achieves the state-of-the-art performance in bits-per-dimension and yields great interpretability of its neurons in intermediate layers.

1 INTRODUCTION

Bayesian inference provides a powerful framework to blend prior knowledge, data generation process and (possibly small) data for statistical inference. With some prior knowledge ρ (distribution) for the quantity of interest $\boldsymbol{x} \in \mathbb{R}^d$, and some (noisy) measurements $\boldsymbol{y}$, it casts on $\boldsymbol{x}$ a posterior

$$q(\boldsymbol{x}|\boldsymbol{y}) \propto \rho(\boldsymbol{x}) L(\boldsymbol{y}|\boldsymbol{x}) \,, \tag{1}$$

where $L(\boldsymbol{y}|\boldsymbol{x})$ is the likelihood. In absence of ambiguity, we short $q(\boldsymbol{x}|\boldsymbol{y})$ as $q(\boldsymbol{x})$ for simplicity.

Sampling or inferring from q is a long-standing challenge, especially for high-dimensional (high-d) cases. An arbitrary high-d posterior can have its importance regions (also called "modes") anywhere in the high-d space, and finding these modes requires computational cost that grows exponentially with the dimension d. This intrinsic difficulty is the consequence of "the curse of dimensionality", which all existing Bayesian inference methods suffer from, e.g., MCMC-based methods (Neal et al., 2011; Welling & Teh, 2011; Cui et al., 2016), SVGD-type methods (Liu & Wang, 2016; Chen et al., 2018; 2019a), and generative modeling (Morzfeld et al., 2012; Parno et al., 2016; Hou et al., 2019).

In this paper, we focus on Bayesian inference problems with multiscale structure and exploit this structure to sample from a high-d posterior. While the original problem has a high spatial resolution (fine-scale), its low resolution (coarse-scale) analogy is computationally attractive because it lies in a low-dimension (low-d) space. A problem has the multiscale structure if such coarse-scale low-d surrogate exists and gives good approximation to the fine-scale high-d problem, see Section 2.1. Such multiscale property is very common in high-d Bayesian inference problems. For example, inferring 3-D permeability field of subsurface at the scale of meters is a reasonable approximation of itself at the scale of centimeters, while the problem dimension is 10^6-times fewer.

We propose a Multiscale Invertible Generative Network (MsIGN) to sample from high-d Bayesian inference problems with multiscale structure. MsIGN consists of multiple scales that recursively lifts up samples to a finer-scale (higher-resolution), except that the coarsest scale directly samples from a low-d (low resolution) distribution. At each scale, a fixed prior conditioning layer mixes coarse-scale samples with some random noise according to the prior to enhance the resolution, and then an invertible flow modifies the samples for better accuracy, see Figure 1. The architecture of MsIGN makes it fully invertible between the final sample and random noise at all scales.

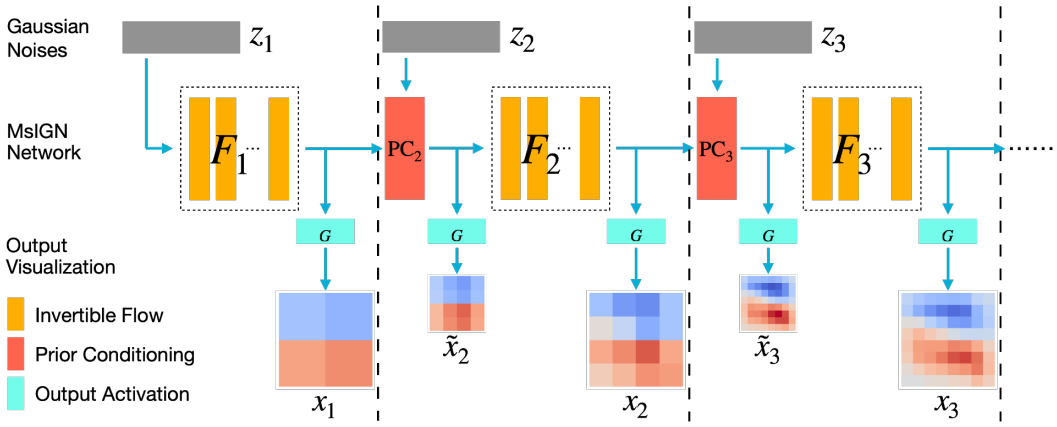

Figure 1: MsIGN generates samples from coarse to fine scale. Each scale, as separated by vertical dash lines, takes in feature from the coarser scale and Gaussian noise, and outputs a sample $(\boldsymbol{x}_2, \boldsymbol{x}_3, \ldots)$ of finer scale. At each scale, the prior conditioning layer lifts up a sample to a finer scale $\tilde{\boldsymbol{x}}_l$, which is the best guess of $\boldsymbol{x}_l$ given its coarse-scale value $\boldsymbol{x}_{l-1}$ and the prior. An invertible flow further modifies $\tilde{\boldsymbol{x}}_l$ to better approximate $\boldsymbol{x}_l$. See Section 2.1 for detailed explanation.

MsIGN undergoes a multi-stage training that learns a hierarchy of distributions with dimensions growing from the lowest to the highest (the target posterior). Each stage gives a good initialization to the next stage thanks to the multiscale property; see Figure 1. To capture multiple modes, we choose Jeffreys divergence as the training objective at each stage:

$$D_{\mathrm{J}}(p\|q) = D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p)\,. \tag{2}$$

Jeffreys divergence removes bad local minima of single-sided Kullback-Leibler (KL) divergence to avoid mode missing. We build an unbiased estimation of it by leveraging prior conditioning layer in importance sampling. Proper loss function and good initialization from multi-stage training solve the non-convex optimization stably and capture multi-modes of the high-d distribution.

In summary, we claim four contributions in this work. First, we propose a Multiscale Invertible deep Generative Network (MsIGN) with a novel prior conditioning layer, which can be trained in a coarse-to-fine scale manner. Second, Jeffreys divergence is used as the objective function to avoid mode collapse. Importance sampling based on the prior conditioning layer is leveraged to estimate Jeffreys divergence. Third, when applied to two Bayesian inverse problems, MsIGN clearly captures multiple modes in the high-d posterior and approximates the posterior accurately, demonstrating its superior performance compared with previous methods via the generative modeling approach. Fourth, we also apply MsIGN to image synthesis tasks, where it achieves SOTA performance in terms of bits-per-dimension and yields great interpret-ability of its neurons in intermediate layers.

2 Methology

Bayesian inference builds its likelihood $L(\boldsymbol{y}|\boldsymbol{x})$ in Equation 1 by comparing data $\boldsymbol{y}$ with the system prediction $\mathcal{F}(\boldsymbol{x})$ on candidate $\boldsymbol{x}$, where $\mathcal{F}$ denotes the forward process. We can use different distributions to model the mismatch ε, and for illustration simplicity, we assume Gaussian here:

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}) + \boldsymbol{\varepsilon}\,, \quad L(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} - \mathcal{F}(\boldsymbol{x}); \boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_{d_y})\,. \tag{3}$$

For example, Bayesian deep learning generates model predicted logits $\mathcal{F}(\boldsymbol{x})$ from model parameters $\boldsymbol{x}$, and compares it with discrete labels $\boldsymbol{y}$ through binomial or multinomial distribution.

In Section 2.1, we discuss the multiscale structure in detail and derive a scale decoupling of the posterior that can be utilized to divide and conquer the high-d challenge of Bayesian inference. Then we presents the architecture of MsIGN in Section 2.2 and the training algorithm in Section 2.3.

2.1 Multiscale Structure and Scale Decoupling

A deterministic pooling operator $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^{d_c}$ connects a fine-scale quantity $\boldsymbol{x} \in \mathbb{R}^d$ to its coarse-scale version $\boldsymbol{x}_c \in \mathbb{R}^{d_c}$ ($d_c < d$): $\boldsymbol{x}_c = \mathcal{A}(\boldsymbol{x})$. We say a Bayesian inference problem has *multiscale structure* if an associated coarse-scale likelihood L_c approximates the original likelihood L well:

$$L(\boldsymbol{y}|\boldsymbol{x}) \approx L_c(\boldsymbol{y}|\boldsymbol{x}_c)\,, \quad \text{where} \quad L_c(\boldsymbol{y}|\boldsymbol{x}_c) := \mathcal{N}(\boldsymbol{y} - \mathcal{F}_c(\boldsymbol{x}_c); \boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_{d_y})\,. \tag{4}$$

Here $\mathcal{F}_c : \mathbb{R}^{d_c} \to \mathbb{R}^{d_y}$ is the coarse-scale forward process. A popular case of Equation 4 is when $\mathcal{F}(\boldsymbol{x}) \approx \mathcal{F}_c(\boldsymbol{x}_c)$. For example, the subsurface flow simulation at the scale of meters approximates well the simulation at the scale of centimeters, where $\mathcal{A}$ is the average pooling operator.

Equation 4 motivates us to define a posterior surrogate $\tilde{q}$ that approximates q well[1]:

$$\tilde{q}(\boldsymbol{x}) := \rho(\boldsymbol{x}) L_c(\boldsymbol{y}|\boldsymbol{x}_c) = \rho(\boldsymbol{x}) L_c(\boldsymbol{y}|\mathcal{A}(\boldsymbol{x})) \approx \rho(\boldsymbol{x}) L(\boldsymbol{y}|\boldsymbol{x}) = q(\boldsymbol{x}) \,. \tag{5}$$

Conditional probability calculation gives $\rho(\boldsymbol{x}) = \rho(\boldsymbol{x}|\boldsymbol{x}_c)\rho(\boldsymbol{x}_c)$, so we decouple the surrogate $\tilde{q}$ as

$$\tilde{q}(\boldsymbol{x}) \propto \rho(\boldsymbol{x}) L_c(\boldsymbol{y}|\boldsymbol{x}_c) = \rho(\boldsymbol{x}|\boldsymbol{x}_c)\rho(\boldsymbol{x}_c) L_c(\boldsymbol{y}|\boldsymbol{x}_c) = \rho(\boldsymbol{x}|\boldsymbol{x}_c) q_c(\boldsymbol{x}_c) \,, \tag{6}$$

where in the last step $q_c(\boldsymbol{x}_c) := \rho(\boldsymbol{x}_c) L(\boldsymbol{y}|\boldsymbol{x}_c)$ is exactly the posterior at the coarse scale. The prior conditional distribution $\rho(\boldsymbol{x}|\boldsymbol{x}_c) = \rho(\boldsymbol{x}|\boldsymbol{x}_c = \mathcal{A}(\boldsymbol{x}))$ bridges the coarse-scale posterior $q_c(\boldsymbol{x}_c)$ and the surrogate $\tilde{q}(\boldsymbol{x})$, which in turn approximates the original fine-scale posterior $q(\boldsymbol{x})$. Parno et al. (2016) proposed Equation 6 similarly, and we leave the discussion and comparison in Appendix A.

Figure 1 shows the integrated sampling strategy. To sample an $\boldsymbol{x}$ from q, we start with an $\boldsymbol{x}_c$ from q_c. The prior conditioning layer performs random upsampling from the prior conditional distribution $\rho(\cdot|\boldsymbol{x}_c)$, and the output will be a sample of the surrogate $\tilde{q}$. We stack multiple invertible blocks for the invertible flow F, which aims to modify $\tilde{q}$ to q, and is initialized as an identity map in training. Finally to obtain the $\boldsymbol{x}_c$ from q_c, we apply the above procedure recursively until the dimension of the coarsest scale is small enough so that q_c can be easily sampled by a standard method.

2.2 Multiscale Invertible Generative Network: Architecture

Our proposed MsIGN has multiple levels to recursively apply the above strategy. We denote L the number of levels, $\boldsymbol{x}_l \in \mathbb{R}^{d_l}$ the sample at level l, and $\mathcal{A}_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_{l-1}}$ the pooling operator from level l to $l-1$: $\boldsymbol{x}_{l-1} = \mathcal{A}_l(\boldsymbol{x}_l)$. Following the idea in Section 2.1, we can define the l-th level target $q_l(\boldsymbol{x}_l)$ and surrogate $\tilde{q}_l(\tilde{\boldsymbol{x}}_l)$, where the coarse scale refers to $(l-1)$-th level. The l-th level of MsIGN tries to capture q_l and contains a prior conditioning layer PC_l and an inverse transform F_l.

Prior conditioning layer. The prior conditioning layer PC_l at level l lifts a coarse-scale sample $\boldsymbol{x}_{l-1} \in \mathbb{R}^{d_{l-1}}$ up to a *random* fine-scale one $\boldsymbol{x}_l \in \mathbb{R}^{d_l}$ following the conditional distribution $\rho(\boldsymbol{x}_l|\boldsymbol{x}_{l-1})$. The difference in dimension is compensated by a Gaussian noise $\boldsymbol{z}_l \in \mathbb{R}^{d_l - d_{l-1}}$, which is the source of randomness: $\boldsymbol{x}_l = \mathrm{PC}_l(\boldsymbol{x}_{l-1}, \boldsymbol{z}_l)$. PC_l depends only on the prior conditional distribution $\rho(\boldsymbol{x}_l|\boldsymbol{x}_{l-1})$, and thus can be pre-computed *independently for different levels* regardless of the likelihood L. When the prior is Gaussian and the pooling operators are linear (e.g., average pooling), the prior conditional distribution is still Gaussian with moments specified as follows.

Lemma 2.1 *Suppose that the prior $\rho(\boldsymbol{x}_l) = \mathcal{N}(\boldsymbol{x}_l; \boldsymbol{0}, \boldsymbol{\Sigma}_l)$, and that the pooling operator is linear, i.e., $\mathcal{A}_l(\boldsymbol{x}_l) = \boldsymbol{A}_l \boldsymbol{x}_l$ for $\boldsymbol{A}_l \in \mathbb{R}^{d_{l-1} \times d_l}$. The prior conditional distribution is also a Gaussian:*

$$\rho(\boldsymbol{x}_l|\boldsymbol{x}_{l-1} = \boldsymbol{A}_l \boldsymbol{x}_l) = \mathcal{N}(\boldsymbol{x}_l; \boldsymbol{U}_{l-1}\boldsymbol{x}_{l-1}, \boldsymbol{\Sigma}_{l|l-1}) \,,$$

where $\boldsymbol{U}_{l-1} = \boldsymbol{\Sigma}_l \boldsymbol{A}_l^T (\boldsymbol{A}_l \boldsymbol{\Sigma}_l \boldsymbol{A}_l^T)^{-1}$ and $\boldsymbol{\Sigma}_{l|l-1} = \boldsymbol{\Sigma}_l - \boldsymbol{\Sigma}_l \boldsymbol{A}_l^T (\boldsymbol{A}_l \boldsymbol{\Sigma}_l \boldsymbol{A}_l^T)^{-1} \boldsymbol{A}_l \boldsymbol{\Sigma}_l$.

With the Cholesky decomposition (or eigen-decomposition) $\boldsymbol{\Sigma}_{l|l-1} = \boldsymbol{B}_l \boldsymbol{B}_l^T$, we design the prior conditioning layer PC_l as below, which is invertible between $\boldsymbol{x}_l$ and $(\boldsymbol{x}_{l-1}, \boldsymbol{z}_l)$ (see Appendix B):

$$\boldsymbol{x}_l = \mathrm{PC}_l(\boldsymbol{x}_{l-1}, \boldsymbol{z}_l) := \boldsymbol{U}_{l-1}\boldsymbol{x}_{l-1} + \boldsymbol{B}_l \boldsymbol{z}_l \,, \quad \boldsymbol{z}_l \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_l - d_{l-1}}) \,. \tag{7}$$

When the prior is non-Gaussian or the pooling operators are nonlinear, there exists a *nonlinear* invertible prior conditioning operator $\boldsymbol{x}_l = \mathrm{PC}_l(\boldsymbol{x}_{l-1}, \boldsymbol{z}_l)$ such that $\boldsymbol{x}_l$ follows the prior conditional distribution $\rho(\boldsymbol{x}_l|\boldsymbol{x}_{l-1})$ given $\boldsymbol{x}_{l-1}$ and $\boldsymbol{z}_l \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_l - d_{l-1}})$. We can pre-train an invertible network to approximate this sampling process, and fix it as the prior conditioning layer.

Invertible flow. The invertible flow F_l at level l modifies the surrogate $\tilde{q}_l$ towards the target q_l. The more accurate the multiscale structure in Equation 4 is, the better $\tilde{q}_l$ approximates q_l, and the closer F_l is to the identity map. Therefore, we parameterize F_l by some flow-based generative model and initialize it as an identity map. In practice, we utilize the invertible block of Glow (Kingma & Dhariwal, 2018), which consists of actnorm, invertible 1×1 convolution, and affine coupling layer, and stack several blocks as the inverse flow F_l in MsIGN.

[1] Here we omit all normalizing constants. Equivalence and approximation are up to normalization.

Output activation. The output activation G is an element-wise invertible map at the end of MsIGN to enforce certain bound of the output. For example, image synthesis task can use the sigmoid function so that pixel values lie in $[0, 1]$. We pick G as the identity map if no bound is required.

After all, MsIGN is precisely formulated as below in Equation 8, and is invertible between the finest-scale output x_L and random noise inputs at different scales $\{z_l\}_{l=1}^{L}$.

$$\begin{aligned} \boldsymbol{h}_1 &= F_1(\boldsymbol{z}_1)\,, \quad \boldsymbol{x}_1 = G(\boldsymbol{h}_1)\,, \\ \boldsymbol{h}_l &= F_l(\mathrm{PC}_l(\boldsymbol{h}_{l-1}, \boldsymbol{z}_l))\,, \quad \boldsymbol{x}_l = G(\boldsymbol{h}_l)\,, \quad 2 \le l \le L\,. \end{aligned} \tag{8}$$

Thanks to the invertibility, we can evaluate density $p(\boldsymbol{x})$ for samples from MsIGN.

2.3 Multiscale Invertible Generative Network: Training

Since the prior conditioning layer PC is pre-computed and the output activation G is fixed, only the inverse flow F contains trainable parameters in MsIGN.

Multi-stage training and interpret-ability. The multiscale strategy (Figure 1) in construction of MsIGN enables a coarse-to-fine multi-stage training. At stage l, we target at capturing q_l, and only train invertible flows $F_{l'}, l' \le l$. We initialize our model as the surrogate $\tilde{q}_l$ by setting $F_{l'}, l' < l$ as the trained model at stage $l-1$ and F_l as the identity map. Our experiments demonstrate such multi-stage strategy significantly stabilizes training and improves final performance.

Figure 1 and Equation 8 imply that intermediate activations, i.e., $\boldsymbol{h}_l$ and $\boldsymbol{x}_l$, who are samples of predefined posterior distributions at the coarse scales (see Equation 6), are semantically meaningful and interpret-able. This is different from Glow (Kingma & Dhariwal, 2018), whose intermediate activations are not interpret-able.

Jeffreys divergence and importance sampling with the surrogate. The KL divergence is easy to compute, and thus is widely used as the training objective. However, its landscape could admit local minima that don't favor the optimization. Nielsen & Nock (2009) suggests that $D_{\mathrm{KL}}(p\|q)$ is zero-forcing, meaning that it enforces p be small whenever q is small. As a consequence, mode missing can still be a local minimum, see Appendix C. Therefore, we turn to the Jeffreys divergence defined in Equation 2 who penalizes mode missing much and can remove such local minima.

Evaluating the Jeffreys divergence requires computing an expectation with respect to the target q, which is normally prohibited. Since MsIGN constructs a good approximation $\tilde{q}$ to q from coarser levels in multi-stage training, we do importance sampling with the surrogate $\tilde{q}$ to estimate:

$$D_{\mathrm{J}}(p\|q) = \mathbb{E}_{\mathrm{x}\sim p}[\log \frac{p(x)}{q(x)}] + \mathbb{E}_{\mathrm{x}\sim q}[\log \frac{q(x)}{p(x)}] = \mathbb{E}_{\mathrm{x}\sim p}[\log \frac{p(x)}{q(x)}] + \mathbb{E}_{\mathrm{x}\sim \tilde{q}}[\frac{q(x)}{\tilde{q}(x)} \log \frac{q(x)}{p(x)}]\,. \tag{9}$$

We optimize it by stochastic gradient descent, with details on gradient calculation in Appendix D.

3 Related Work

Invertible generative models (Deco & Brauer, 1995) are powerful exact likelihood models with efficient sampling and inference. They have achieved great success in natural image synthesis, see, e.g., Dinh et al. (2016); Kingma & Dhariwal (2018); Grathwohl et al. (2018); Ho et al. (2019); Chen et al. (2019b), and variational inferences in providing a tight evidence lower bound (ELBO), see, e.g, Rezende & Mohamed (2015). In this paper, we propose a new multiscale invertible generative network (MsIGN) structure, which utilizes the invertible block in Glow (Kingma & Dhariwal, 2018) as building piece for the invertible flow at each scale. The Glow block can be replaced by any other invertible blocks, without any algorithmic changes. Different from Glow, different scales of MsIGN can be trained separately, and thus features in its intermediate layers can be interpreted as low-resolution approximation of the final high-resolution output. This novel multiscale structure enables better explain-ability of its hidden neurons and makes training much more stable.

Different from the image synthesis task where large amount of samples from target distribution are available, in Bayesian inference problems only an unnormalized density is available and i.i.d. samples from the posterior are the target. This paper's main goal is to train MsIGN to approximate certain high-d Bayesian posteriors. Various kinds of parametric distributions have been proposed to approximate posteriors before, such as polynomials (El Moselhy & Marzouk, 2012), non-invertible generative networks (Feng et al., 2017; Hou et al., 2019), invertible networks (Rezende & Mohamed,

2015; Ardizzone et al., 2018; Kruse et al., 2019) and certain implicit maps (Chorin & Tu, 2009; Morzfeld et al., 2012). Generative modeling approach has the advantage that i.i.d. samples can be efficiently obtained by evaluating the model in the inference stage. However, due to the tricky non-convex optimization problem, this approach for both invertible (Chorin & Tu, 2009; Kruse et al., 2019) and non-invertible (Hou et al., 2019) generative models becomes increasingly challenging as the dimension grows. To overcome this difficulty, we propose (1) to use the Jeffreys divergence as loss function, who has fewer shallow local minima and better landscape compared with the commonly-used KL divergence (see Appendix C for a concrete example), and (2) to train MsIGN in a coarse-to-fine manner with coarse-scale solution serving as an initialization to fine-scale optimization problem. In Kruse et al. (2019), authors list some recent models for low-d inverse problems. We remark that their formulation of posterior assumes no observation or model error in Equation 3, and is different from ours, see Appendix J for detailed discussion and experimental comparison.

Other than the generative modeling, various Markov Chain Monte Carlo (MCMC) methods have been the most popular in Bayesian inference, see, e.g., Beskos et al. (2008); Neal et al. (2011); Welling & Teh (2011); Chen et al. (2014; 2015); Cui et al. (2016). Particle-optimization-based sampling is a recently developed effective sampling technique with Stein variational gradient descent (SVGD) (Liu & Wang, 2016)) and many related works, e.g., Liu (2017); Liu & Zhu (2018); Chen et al. (2018; 2019a). The intrinsic difficulty of Bayesian inference displays itself as highly correlated samples, leading to undesired low sample efficiency, especially in high-d cases. The multiscale structure and multi-stage strategy proposed in this paper can also benefit these particle-based methods, as we can observe that they benefit the amortized-SVGD (Feng et al., 2017; Hou et al., 2019) in Section 4.1.3. We leave a more thorough study of this topic as a future work.

Works in Parno et al. (2016); Matthies et al. (2016) utilize the multiscale structure in Bayesian inference and build generative models with polynomials. They suffer from exponential growth of parameter number for high-d polynomial basis. The Markov property (Spantini et al., 2018) is used to alleviate this exponential growth. Different from these works, we leverage the great capacity of invertible generative networks to parametrize the high-d distribution, and we design novel network architecture to make use of the multiscale structure. The multiscale structure is a more general structure than commonly-used intrinsic low-d structure (Spantini, 2017; Cui et al., 2016; Chen et al., 2019a), which assumes the density of high-d posterior concentrates in a low-d subspace.

4 EXPERIMENT

We study two high-d Bayesian inverse problems (BIPs) known to have at least two equally important modes in Section 4.1 as testbeds for distribution approximation and multi-mode capture: one with true samples available in Section 4.1.1; one without true samples but close to real-world applications in subsurface flow in Section 4.1.2. In Section 4.1.3, we also report the ablation study of MsIGN. Besides, we apply MsIGN to image synthesis task to benchmark with other flow-based generative models and demonstrate its interpret-ability in Section 4.2. We adopt the invertible block in Glow (Kingma & Dhariwal, 2018) as the building piece, and stack several of them to build our invertible flow F. We utilize average pooling with kernel size 2 and stride 2 as our pooling operator $\mathcal{A}$.

4.1 BAYESIAN INVERSE PROBLEMS

Sample $\boldsymbol{x}$ of our target posterior distribution is a vector on a 2-D uniform 64×64 lattice, which means the problem dimension d is 4096. Every $\boldsymbol{x}$ is equivalent to a piece-wise constant function on the unit disk: $x(s)$ for $s \in \Omega = [0,1]^2$, and we don't distinguish between them thereafter. We place a centered Gaussian with a Laplacian-type covariance as the prior: $\mathcal{N}\left(0, \beta^2(-\boldsymbol{\Delta})^{-1-\alpha}\right)$, which is very popular in geophysics and electric tomography. See Appendix E for problem settings in detail.

The key feature to guarantee the multi-modality of our posteriors is the symmetry. Combining good properties of the prior defined above and the likelihood defined afterwards, the posterior is symmetric in the following sense: $q(x(s_1, s_2)) = q(x(s_1, 1 - s_2))$. Based on this symmetry in s_2-axis, we build our problem with at least two modes with equal importance.

As in Figure 1, we plan to learn our 4096-D posteriors at the end of $L = 6$ levels, and set problem dimension at each level as $d_l = 2^l * 2^l = 4^l$. The training follows our multi-stage strategy, and the first stage $l = 1$ is initialized by minimizing the Jeffreys divergence without importance sampling, because samples to q_1 is available since $d_1 = 4$ is relatively small. See Appendix E for details.

We compare MsIGN with representatives of major approaches: amortized-SVGD (short as A-SVGD) (Feng et al., 2017) and Hamilton Monte Carlo (short as HMC) (Neal et al., 2011), for high-d BIPs, see our discussion in Section 3. We measure the computational cost by the number of forward simulations (nFSs), because running the forward simulation $\mathcal{F}$ occupies most training time, especially in Section 4.1.2. We budget a same nFS for all methods for fair comparison.

4.1.1 Synthetic Bayesian Inverse Problems

This problem admits access to groundtruth samples so the comparison is clear and solid. The forward process is defined by $\mathcal{F}(\boldsymbol{x}) = \langle \boldsymbol{f}, \boldsymbol{x} \rangle^2 = (\int_\Omega f(s)x(s)\mathrm{d}s)^2$, where $f(s) = \sin(\pi s_1)\sin(2\pi s_2)$. Together with the prior, our posterior can be factorized into one-dimensional sub-distributions, namely $q(\boldsymbol{x}) = \prod_{k=1}^d q_k(\langle \boldsymbol{w}_k, \boldsymbol{x} \rangle)$ for some orthonormal basis $\{\boldsymbol{w}_k\}_{k=1}^d$. This property gives us access to true samples via inversion cumulative function sampling along each direction $\boldsymbol{w}_k$. Furthermore, these 1-D sub-distributions are all single modal except that there's one, denoted as q_{k^*}, with two symmetric modes. This confirms our construction of two equally important modes. See Appendix E for more details in problem settings. The computation budget is fixed at 8×10^5 nFSs.

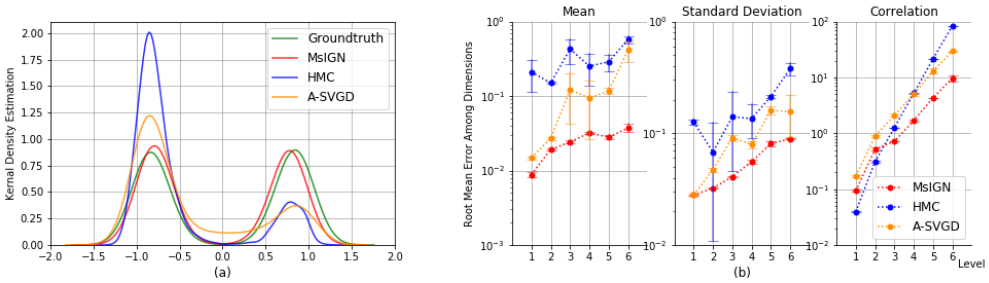

Figure 2: Results of the synthetic BIP. (a): Distribution of 2500 samples along the critical direction $\boldsymbol{w}_{k^*}$. MsIGN is more robust in capturing both modes, its samples are more balanced. (b): Error mean and its 95% confidence interval. MsIGN is more accurate in distribution approximation, especially at finer scale when the problem dimension is high. The margin is statistical significant as shown by the confidence interval. For more experimental results, please refer to Appendix F.

Multi-mode capture. To visualize mode capture, we plot the marginal distribution of generated samples along the critical direction $\boldsymbol{w}_{k^*}$, which by construction is the source of double modality of the posterior. The (visually) worst one in three independent experiments is shown in Figure 2(a).

Distribution approximation. To measure distribution approximation, we report the error of mean, variance and correlation at or between all sub-distributions, as well as the Jeffreys divergence. Thanks to the factorization property, we compare the mean, variance and correlation estimate with theoretical groundtruths, and report the root mean square of error at all dimensions in Figure 2(b). For MsIGN and A-SVGD who gives access to not only samples but also density, we also report the Monte Carlo estimates of the Jeffreys divergence with the target posterior in Table 1. We can see that MsIGN has superior accuracy in approximating the target distribution.

Method	MsIGN	A-SVGD (Feng et al., 2017)
Error	56.77±0.15	3372±21

Table 1: Distribution approximation error by Jeffreys divergence with the target posterior in three independent runs

4.1.2 Elliptic Bayesian Inverse Problems

This problem originates from geophysics and fluid dynamics. The forward model is given by linear measurement of the solution to an elliptic partial differential equation associated with $\boldsymbol{x}$. We define

$$\mathcal{F}(x) = \left[\int_\Omega \varphi_1(s)u(s)\mathrm{d}s \quad \int_\Omega \varphi_2(s)u(s)\mathrm{d}s \quad \dots \quad \int_\Omega \varphi_m(s)u(s)\mathrm{d}s\right]^T ,$$

where φ_k are fixed measurement functions, and u is the solution of

$$-\nabla \cdot \left(e^{x(s)}\nabla u(s)\right) = f(s)\,, s \in \Omega, \text{ with boundary condition } u(s) = 0\,, s \in \partial\Omega\,.$$

This model appears frequently in real applications. For example, x, u can be seen as permeability field and pressure in geophysics. However, there is no known access to true samples of q. Again the trick of symmetry introduced in Section 4.1 and explained in Appendix E guarantees at least two equally important modes in the posterior. We budget our computation cost at 5×10^5 nFSs.

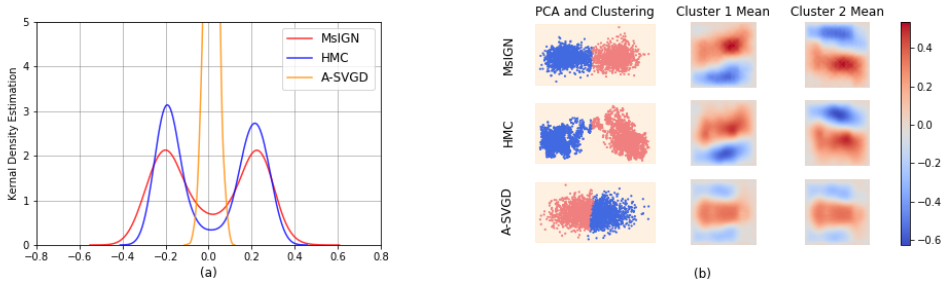

Figure 3: Results of the elliptic BIP. (a): Distribution of 2500 samples along a critical direction. MsIGN and HMC capture two modes in this marginal distribution, but A-SVGD fails. (b): Clustering result of 2500 samples. Samples of MsIGN are more balanced between two modes. The similarity of the cluster means of MsIGN and HMC implies that they both are likely to capture the correct modes. For more experimental results, please refer to Appendix I.

Multi-mode capture. Due to lack of true samples, we check the marginal distribution of the posterior along eigen-vectors of the prior, and pick a particular one that shows the capture of double modes in Figure 3(a). We also confirm the capture of multiple modes by embedding samples by Principle Component Analysis (PCA) to a 2-D space. We report the clustering (by K-means) result and means of each cluster in Figure 3(b), where we can see that A-SVGD failed to capture the two symmetric modes, while MsIGN has a more balanced capture of the symmetric posterior.

4.1.3 Ablation Study of Architecture Design and Training Strategy

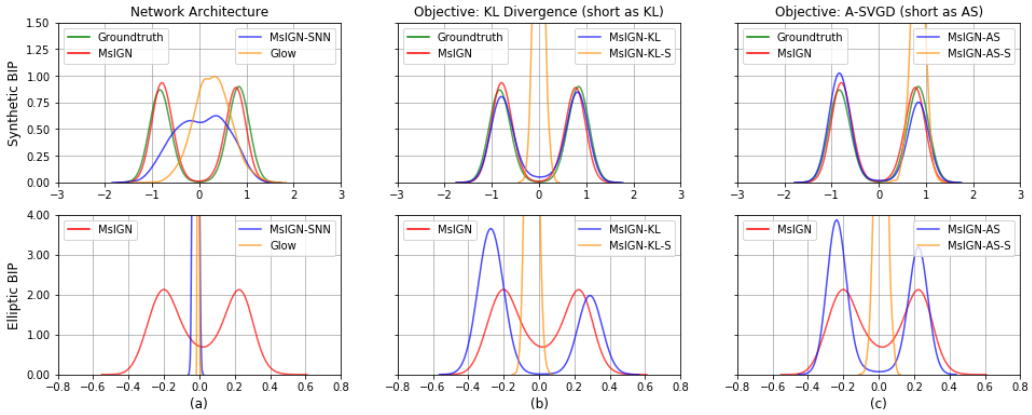

Figure 4: MsIGN here refers to training MsIGN network with Jeffreys divergence and multi-stage strategy. (a): Ablation study of the network architecture of MsIGN. The stochastic nearest neighborhood (SNN) pooling layer can't replace the prior conditioning layer in multi-mode capture, neither can Glow design replace MsIGN. (b) and (c): Ablation study of training objective and multi-stage training. Multi-stage training is crucial for both problems regardless of the objective as compared to single-stage training (experiments end with "-S"). And the Jeffreys divergence leads to more balanced samples for these symmetric problems, especially for the complicated elliptic BIP.

For network architecture, we try replacing the prior conditioning layer by a stochastic nearest-neighbor upsampling layer, or using the Glow design. For training strategy, we try substituting the Jeffreys divergence with the KL divergence (KL) or kernelized Stein discrepancy (which resumes A-SVGD algorithm, short as AS), and switching between multi-stage or single-stage training (short as S). In Figure 4, we show the critical marginal distributions for both problems (Sections 4.1.1 and 4.1.2) for mode capture. The comparison shows that these designs are all crucial to the performance of MsIGN. Besides, our architecture and training strategy can also benefit A-SVGD in multi-mode capture; see Figure 4(c). We refer to Appendix G for details in setting and more results.

4.2 Image Synthesis Task

We train MsIGN architecture with maximum likelihood estimation to benchmark with other flow-based generative models. The prior conditional distribution $\rho(x|x_c)$ is modeled by a simple Gaussian with a scalar matrix as its covariance and is learned from training set. See Appendix H for more experimental details, ad Appendix I for additional results.

We report the bits per dimension value in Table 2. Our MsIGN achieves the state-of-the-art number but is more efficient in parameter size: for example, MsIGN uses 24.4% fewer parameters than Glow for CelebA 64, and uses 37.4% fewer parameters than Residual Flow for ImageNet 64.

Table 2: Bits per dimebsion on test sets. All models in this table do not use "variational dequantization" in Ho et al. (2019), see Appendix I. *: Score obtained by our own reproducing experiment.

Model	MNIST	CIFAR-10	CelebA 64	ImageNet 32	ImageNet 64
Real NVP(Dinh et al., 2016)	1.06	3.49	3.02	4.28	3.98
Glow(Kingma & Dhariwal, 2018)	1.05	3.35	2.20*	4.09	3.81
FFJORD(Grathwohl et al., 2018)	0.99	3.40	–	–	–
Flow++(Ho et al., 2019)	–	3.29	–	–	–
i-ResNet(Behrmann et al., 2019)	1.05	3.45	–	–	–
Residual Flow(Chen et al., 2019b)	0.97	**3.28**	–	**4.01**	3.76
MsIGN (Ours)	**0.93**	**3.28**	**2.15**	4.03	**3.73**

In Figure 5, we show synthesized images of MsIGN from CelebA 64 dataset, and linear interpolation of real images in the latent feature space. In Figure 6, we visualize internal activations at checkpoints of the invertible flow at different scales which demonstrates the interpret-ability of MsIGN.

Figure 5: Left: Synthesized CelebA 64 images with temperature 0.9. Right: Linear interpolation in latent space shows MsIGN's parameterization of natural image manifold is semantically meaningful.

Figure 6: Visualization of internal activation shows the interpret-ability of MsIGN hidden neurons. From left to right, we show how MsIGN progressively generates new samples in high resolution by taking snapshots at internal checkpoints. See Appendix I for details.

5 Conclusion

For high-dimensional Bayesian inference problems with multiscale structure, we propose Multiscale Invertible Generative Networks (MsIGN) and associated training algorithms to approximate the high-dimensional posterior. In this paper, we see its initial success in high-dimensional (up to 4096 dimensions) Bayesian inference problems with spatial multiscale structure, leaving several important directions as future work. The network architecture also achieves the state-of-the-art performance in various image synthesis tasks. We plan to apply this methodology to other Bayesian inference problems, for example, Bayesian deep learning with multiscale structure in model width or depth (e.g., Chang et al. (2017); Haber et al. (2018)) and data assimilation problem with multiscale structure in the temporal variation (e.g., Giles (2008)). We also plan to develop some theoretical guarantee of the posterior approximation performance for MsIGN.

REFERENCES

Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*, 2018.

Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pp. 573–582, 2019.

Alexandros Beskos, Gareth Roberts, Andrew Stuart, and Jochen Voss. Mcmc methods for diffusion bridges. *Stochastics and Dynamics*, 8(03):319–350, 2008.

Bo Chang, Lili Meng, Eldad Haber, Frederick Tung, and David Begert. Multi-level residual networks from dynamical systems view. *arXiv preprint arXiv:1710.10348*, 2017.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient mcmc algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pp. 2278–2286, 2015.

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.

Peng Chen, Keyi Wu, Joshua Chen, Tom O'Leary-Roseberry, and Omar Ghattas. Projected stein variational newton: A fast and scalable bayesian inference method in high dimensions. In *Advances in Neural Information Processing Systems*, pp. 15104–15113, 2019a.

Tian Qi Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, pp. 9913–9923, 2019b.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691, 2014.

Alexandre J Chorin and Xuemin Tu. Implicit sampling for particle filters. *Proceedings of the National Academy of Sciences*, 106(41):17249–17254, 2009.

Tiangang Cui, Kody JH Law, and Youssef M Marzouk. Dimension-independent likelihood-informed mcmc. *Journal of Computational Physics*, 304:109–137, 2016.

Gustavo Deco and Wilfried Brauer. Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures. *Neural Networks*, 8(4):525–535, 1995.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized stein variational gradient descent. *arXiv preprint arXiv:1707.06626*, 2017.

Michael B Giles. Multilevel monte carlo path simulation. *Operations research*, 56(3):607–617, 2008.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Eldad Haber, Lars Ruthotto, Elliot Holtham, and Seong-Hwan Jun. Learning across scales—multiscale methods for convolution neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. *arXiv preprint arXiv:1902.00275*, 2019.

Thomas Y Hou, Ka Chun Lam, Pengchuan Zhang, and Shumao Zhang. Solving bayesian inverse problems from the perspective of deep generative networks. *Computational Mechanics*, 64(2): 395–408, 2019.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Jakob Kruse, Lynton Ardizzone, Carsten Rother, and Ullrich Köthe. Benchmarking invertible architectures on inverse problems. ICML, 2019.

Chang Liu and Jun Zhu. Riemannian stein variational gradient descent for bayesian inference. In *Thirty-second aaai conference on artificial intelligence*, 2018.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pp. 3115–3123, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pp. 2378–2386, 2016.

Hermann G Matthies, Elmar Zander, Bojana V Rosić, Alexander Litvinenko, and Oliver Pajonk. Inverse problems in a bayesian setting. In *Computational Methods for Solids and Fluids*, pp. 245–286. Springer, 2016.

Matthias Morzfeld, Xuemin Tu, Ethan Atkins, and Alexandre J Chorin. A random map implementation of implicit filters. *Journal of Computational Physics*, 231(4):2049–2066, 2012.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Frank Nielsen and Richard Nock. Sided and symmetrized bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

Matthew Parno, Tarek Moselhy, and Youssef Marzouk. A multiscale strategy for bayesian inference using transport maps. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1160–1190, 2016.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2015.

Alessio Spantini. *On the low-dimensional structure of Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2017.

Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.