Multiscale Invertible Generative Networks for High-Dimensional Bayesian Inference

Shumao Zhang [1] Pengchuan Zhang [2] Thomas Y. Hou [1]

Abstract

We propose a Multiscale Invertible Generative Network (MsIGN) and associated training algorithm that leverages multiscale structure to solve high-dimensional Bayesian inference. To address the curse of dimensionality, MsIGN exploits the low-dimensional nature of the posterior, and generates samples from coarse to fine scale (low to high dimension) by iteratively upsampling and refining samples. MsIGN is trained in a multi-stage manner to minimize the Jeffreys divergence, which avoids mode dropping in high-dimensional cases. On two high-dimensional Bayesian inverse problems, we show superior performance of MsIGN over previous approaches in posterior approximation and multiple mode capture. On the natural image synthesis task, MsIGN achieves superior performance in bits-per-dimension over baseline models and yields great interpret-ability of its neurons in intermediate layers.

1. Introduction

To infer about hidden system states $x \in \mathbb{R}^d$ from observed system data $y \in \mathbb{R}^s$, Bayesian inference blends some prior knowledge, given as a distribution ρ, with data y into a powerful posterior. Since direct measurement of x can be inaccessible, the data y is generated through $y = \mathcal{F}(x) + \varepsilon$, where $\mathcal{F}$ is a forward map that can be highly nonlinear and complicated, $\varepsilon \in \mathbb{R}^s$ is random noise modelled by some distribution. For illustration simplicity, we assume a Gaussian $\mathcal{N}(0, \Gamma)$ for ε. The posterior is characterized as

$$q(x|y) = \frac{1}{Z}\rho(x)\mathcal{L}(y|x)\,, \tag{1}$$

where $\mathcal{L}$ is the likelihood given as

$$\mathcal{L}(y|x) = \mathcal{N}(y - \mathcal{F}(x); 0, \Gamma)\,, \tag{2}$$

which is the density of $\varepsilon = y - \mathcal{F}(x)$, and Z is some normalizing constant that is usually intractable in practice. For simplicity reason, in the following context we abbreviate $q(x|y)$ in (1) as $q(x)$, because the data y *only* plays the role of defining the target distribution $q(x)$ in our framework.

A key and long-standing challenge in Bayesian inference is to approximate, or draw samples from the posterior q, especially in high-dimensional (high-d) cases. An arbitrary distribution can concentrate its density anywhere in the space, and these concentrations (also called "modes") become less connected as d increases. As a result, detecting these modes requires computational cost that grows exponentially with d. This intrinsic difficulty of mode collapse is a 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 exploit the multiscale structure to deal with the high-dimensional Bayesian inference problems. The multiscale structure means that the forward map $\mathcal{F}$ depends mostly on some low-d structure of x, referred as coarse scale, instead of the high-d structure, referred as fine scale. For example, the terrain shape x, given as the discretization of 2-D elevation map on a 2-D lattice grid, is a quantity with dimension equal to the number of grid points. Simulating the 2-D precipitation distribution y using terrain shape x at the scale of kilometer is a reasonable approximation to itself at the scale of meter. The former one is a coarse-scale version of the latter, and has 10^6-times fewer problem dimension (grid points). Such multiscale structure is very common in high-d problems, especially when x is some spatial or temporal quantity. The coarse-scale approximation to the original fine-scale problem is low-d and computationally attractive, and can help divide-and-conquer the high-d challenge. The multiscale property is discussed in detail in Section 2.

We approximate the target q by a parametric family of distri-

[1] Department of Computational & Mathematical Sciences, Caltech, Pasadena, California, USA [2] MSR AI Lab, Redmond, Washington, USA. Correspondence to: Shumao Zhang <shumaoz@caltech.edu>.

Proceedings of the 38^th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

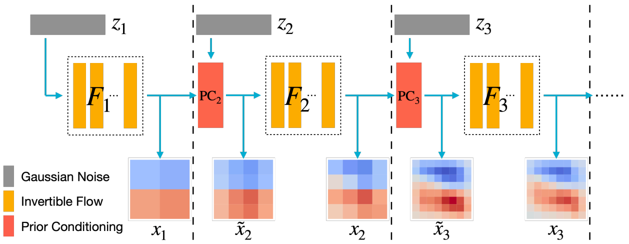

Figure 1. MsIGN generates samples from coarse to fine scale, as depicted by (3). Each scale, separated by dash lines, takes in x_{l-1} from the coarser scale with random seed z_l, and outputs a sample x_l of the finer scale. MsIGN iteratively upsamples (by PC_l) and refines (by F_l) samples to the target scale.

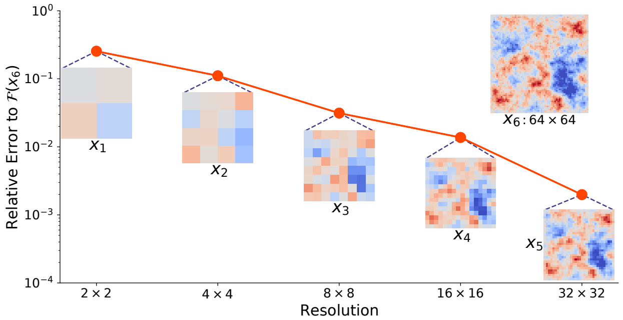

Figure 2. An example of multiscale problem from Section 6.1.2. The coarse-scale $x_1, \ldots, x_5$ are downsampling of the original x_6 from resolution 64×64. As the resolution gets refined, the relative error to $\mathcal{F}(x_6)$ significantly drops. Typically, it suggests a good approximation in (5) when, for example, setting $Ax = x_5$.

bution p_θ, and look for an optimal choice of θ. The working distribution p_θ is the density of $T(z;\theta)$, where z is random seed which we assume to be Gaussian noise here, T is a transport map parameterized by θ that drives z to the sample of p_θ. The optimality of θ is determined by the match of p_θ to q, measured by the Jeffreys divergence $D_{\mathrm{J}}(p_\theta \| q)$.

We propose a Multiscale Invertible Generative Network (MsIGN) as the map T, with a novel training strategy to minimize the Jeffreys divergence. Specifically, T maps z to the sample $x = x_L$ of p_θ in a coarse-to-fine manner:

$$\begin{aligned} x_1 &= F_1(z_1)\,, \\ \tilde{x}_l = PC_l(x_{l-1}, z_l)\,, \quad x_l &= F_l(\tilde{x}_l)\,, \quad 2 \le l \le L\,. \end{aligned} \tag{3}$$

Here we split z into $(z_1, z_2, \ldots, z_L)$. At scale l, the prior conditioning layer PC_l upsamples the coarse-scale $x_{l-1} \in \mathbb{R}^{d_{l-1}}$ to a finer scale $\tilde{x}_l \in \mathbb{R}^{d_l}$, which is the "best guess" of x_l given its coarse scale version x_{l-1} and the prior ρ. The invertible flow F_l then modifies $\tilde{x}_l$ to $x_l \in \mathbb{R}^{d_l}$, which again can be considered as a coarse scale version of x_{l+1}. The final sample $x = x_L$ is constructed iteratively, as the dimension $d_1 < d_2 < \ldots < d_L = d$ grows up, see Figure 1. The overall map T is invertible from z to x.

We train MsIGN by minimizing the Jeffreys divergence $D_{\mathrm{J}}(p_\theta \| q)$, defined by (Jeffreys et al., 1973) as

$$\begin{aligned} D_{\mathrm{J}}(p_\theta \| q) &= D_{\mathrm{KL}}(p_\theta \| q) + D_{\mathrm{KL}}(q \| p_\theta) \\ &= \mathbb{E}_{p_\theta}\left[\log\left(p_\theta / q\right)\right] + \mathbb{E}_q\left[\log\left(q / p_\theta\right)\right]\,. \end{aligned} \tag{4}$$

Jeffreys divergence removes bad local minima of single-sided Kullback-Leibler (KL) divergence to avoid mode missing. We build its unbiased estimation by importance sampling, with the output of the prior conditioning layer as proposal distribution. Furthermore, MsIGN is trained in a multi-stage manner, from coarse to fine scale. At stage l, we train $\{F_{l'} : l' \le l\}$ so that x_l approximates the posterior at its scale, while PC_l are pre-computed and fixed. Each stage provides a good proposal distribution for the importance sampling at the next stage thanks to the multiscale property.

Contribution We claim four contributions in this work. First, we propose a Multiscale Invertible Generative Network (MsIGN) with a novel prior conditioning layer that can generate samples from a coarse-to-fine manner. Second, MsIGN allows multi-stage training to minimize the Jeffreys divergence, which helps avoid mode collapse in high-d problems. 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 over previous methods. Fourth, we also apply MsIGN to image synthesis tasks, where it achieves superior performance in bits-per-dimension among baseline models. MsIGN also yields great interpret-ability of its neurons in intermediate layers.

We introduce the theoretical motivation in Section 2, and give detailed introduction of the network structure of our MsIGN in Section 3, while its training strategy is described in Section 4. Then we review related work, and provide numerical studies in Section 5 and 6 respectively.

2. Theoretical Motivation

Let $A \in \mathbb{R}^{d_c \times d}$ be a linear operator that downsamples x to its coarse-scale low-d version $x_c = Ax \in \mathbb{R}^{d_c}$ with $d_c < d$. For example, A can be the average pooling operator with kernel size 2 and stride 2 which downsamples x to $1/4$ of its original dimensions.

Multiscale structure In many high-d Bayesian inference problems, the observation y relies more on global, coarse-scale structure than local, fine-scale structure of x. This multiscale structure can be described as

$$\mathcal{F}(x) \approx \mathcal{F}(BAx)\,, \quad \forall x \in \mathbb{R}^d\,, \tag{5}$$

where $A \in \mathbb{R}^{d_c \times d}$ is the downsample operator that compress x to a coarse-scale version Ax, and $B \in \mathbb{R}^{d\times d_c}$ transforms the coarse-scale low-d Ax to a valid system input for $\mathcal{F}$. For example, B can be the nearest-neighbor upsample operator such that BAx has the same size as x, but only contains its coarse-scale information. The relation (5) arises frequently when x has some spatial or temporal structure, see an example in Figure 2.

Scale decoupling Let $x_c = Ax$ be the coarse-scale variable. Like in (2), the coarse-scale likelihood is defined as

$$\mathcal{L}_c(y|x_c) = \mathcal{N}(y - \mathcal{F}(Bx_c); 0, \Gamma)\,, \tag{6}$$

and we expect $\mathcal{L}_c(y|Ax) \approx \mathcal{L}(y|x)$ due to (5). On the other hand, let ρ_c be the probability density of $x_c = Ax$ when $x \sim \rho$, which is the coarse-scale prior, the conditional probability rule suggests that $\rho(x|x_c) = \rho(x|Ax = x_c) = \rho(x)/\rho_c(x_c)$, which is equivalent to $\rho(x) = \rho_c(x_c)\rho(x|x_c)$.

With the likelihood approximation and the prior decoupling, the posterior q admits the following scale decoupling:

$$\begin{aligned} q(x) &= \frac{1}{Z}\rho(x)\mathcal{L}(y|x) \approx \frac{1}{Z}\rho(x)\mathcal{L}_c(y|x_c) \\ &= \frac{1}{Z}\rho_c(x_c)\rho(x|x_c)\mathcal{L}_c(y|x_c) \\ &= \frac{Z_c}{Z}\rho(x|x_c)q_c(x_c) \propto \frac{1}{\tilde{Z}}\rho(x|x_c)q_c(x_c) := \tilde{q}(x)\,, \end{aligned} \tag{7}$$

where $q_c(x_c) := \frac{1}{Z_c}\rho_c(x_c)\mathcal{L}_c(y|x_c)$ is the coarse-scale posterior analog to (1), and $\tilde{q}(x) := \frac{1}{\tilde{Z}}\rho(x|x_c)q_c(x_c)$ is a distribution to approximate q, with normalizing constants $Z_c, \tilde{Z}$.

3. Network Architecture

The key observation (7) is essentially

$$\underbrace{q(x) \approx \tilde{q}(x)}_{(iii)} = \underbrace{q_c(x_c)}_{(i)}\underbrace{\rho(x|x_c)}_{(ii)}\,, \tag{8}$$

where $\approx$ and $=$ are up to some multiplicative constant. It suggests a three-step way to sample from q:

(i) generate a sample x_c from q_c;
(ii) sample $\tilde{x}$ from $\rho(\cdot|x_c)$;
(iii) further modify $\tilde{x}$ to x to better approximate q.

We design a prior conditioning layer PC to sample $\tilde{x}$ from $\rho(\cdot|x_c)$ for (ii), and an invertible flow F that modifies $\tilde{x}$ for (iii). To obtain x_c from q_c in (i), 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. As an example of this three-step sampling strategy, in the image synthesis task, a high-resolution image x can be *approximated* by $\tilde{x}$, the upsampled image of its low-resolution version x_c superimposed with random noise according to the prior ρ, which will be specified in Section 6.2 for this task. Needless to say, $\tilde{x}$ needs further modification to achieve good quality in high resolution.

Prior conditioning We feed the coarse-scale sample x_c together with some random seed $z \in \mathbb{R}^{d-d_c}$ to the prior conditioning layer PC to sample from the conditional distribution $\rho(\cdot|x_c)$: $x = PC(x_c, z)$. The conditional sample x should satisfy the constrain $Ax = x_c$. We further require the layer PC to be invertible between x and (x_c, z) to maintain the invertiblity of our overall network. Since PC depends only on the prior distribution ρ and downsampling operator A, it can be pre-computed *regardless* of the likelihood $\mathcal{L}$. In fact, when the prior is a Gaussian, the prior conditional distribution is still a Gaussian and the prior conditioning layer PC admits a closed form:

Theorem 3.1 *Suppose that ρ is a Gaussian with density $\mathcal{N}(x; 0, \Sigma)$ where the covariance Σ is positive definite, then with $U^c := \Sigma A^T(A\Sigma A^T)^{-1} \in \mathbb{R}^{d\times d_c}$ and $\Sigma^c := \Sigma - \Sigma A^T(A\Sigma A^T)^{-1}A\Sigma \in \mathbb{R}^{d\times d}$, we have*

$$\rho(x|Ax = x_c) = \mathcal{N}(x; U^c x_c, \Sigma^c)\,.$$

Furthermore, there exists a matrix $W \in \mathbb{R}^{d\times(d-d_c)}$ such that $\Sigma^c = WW^T$, and the prior conditioning layer PC can be given as, with $z \in \mathbb{R}^{d-d_c}$ being standard Gaussian

$$x = PC(x_c, z) = U^c x_c + Wz\,,$$

and PC is invertible between x and (x_c, z).

We leave the proof in Appendix A. When the prior is non-Gaussian, the prior conditioning layer PC still exists with invertibility guarantee, but it is now *nonlinear*. In this case, we can pre-train an invertible network to approximate the conditional sampling process. Once PC is pre-computed, its parameters are fixed in the training stage.

Invertible flow The invertible flow F is a parametric invertible map that modifies the sample $\tilde{x}$ from the prior conditioning layer to a sample of the target q, in other words, it modifies the distribution $\tilde{q}$ in (8) to the target q. In our experiments in Section 6, 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 such blocks as the inverse flow F in MsIGN. The approximation (8) also suggests that F be initialized as an identity map in training, see Section 4.

Recursive design To initialize our sampling strategy (8) with a sample x_c from the coarse-scale posterior q_c, we recursively apply our strategy until the dimension of the coarsest-scale is small enough. Let L be the number of recursion, also called scales in the following context. Let $x_l \in \mathbb{R}^{d_l}$ be the variable at scale l $(1 \le l \le L)$, whose distribution is the l-th scale posterior q_l analog to the q_c in Section 2 and $q_L = q$. The problem dimension keeps increasing as l goes up: $d_1 < d_2 < \ldots < d_L = d$. Details of constructions at scale l can be found in Appendix D.

Our network structure is shown in (3) and Figure 1, with $z_1 \in \mathbb{R}^{d_1}$ and $z_l \in \mathbb{R}^{d_l - d_{l-1}}(2 \le l \le L)$ be the random seed drawn from standard Gaussian at each scale. At scale l $(2 \leq l \leq L)$, a prior conditioning layer PC_l randomly upsamples $x_{l-1} \in \mathbb{R}^{d_{l-1}}$, taken from q_{l-1} approximately, to $\tilde{x}_l \in \mathbb{R}^{d_l}$, and an invertible flow F_l modifies $\tilde{x}_l$ to x_l to approximate q_l. At scale $l = 1$, we directly learn an invertible flow F_1 that transports $z_1 \sim \mathcal{N}(0, I)$ to $x_1 \sim q_1$ since the problem dimension is small enough to allow efficient application of standard methods.

Write the overall random seed $z \in \mathbb{R}^d$ as a concatenation of $(z_1, z_2, \ldots, z_L)$, and write θ as the parameters in MsIGN. The overall network of MsIGN parameterizes a map $T(\cdot; \theta)$ such that samples are generated by $x = T(z; \theta)$. Let p_z, p_θ be the density of z, x respectively. Our design also allows the invertible mapping $z = T^{-1}(x; \theta)$, so by the change-of-variable formula the density of p_θ is given by

$$p_\theta(x) = p_z(T^{-1}(x; \theta)) |\det \mathrm{J}_x T^{-1}(x; \theta)| \,, \tag{9}$$

where $\mathrm{J}_x T^{-1}$ is the Jacobian of T^{-1} with respect to x.

We also remark that when certain bound needs to be enforced on the output, we can append element-wise output activations at the end of MsIGN. For example, image synthesis can use the sigmoid function so that pixel values lie in $[0, 1]$. Such activations should be bijective to keep the invertible relation between random seed z to the sample x.

4. Training Strategy

We learn network parameter θ by solving the optimization $\min_\theta D_{\mathrm{J}}(p_\theta \| q)$. Since prior conditioning layers PC_l, for $2 \leq l \leq L$, are pre-computed and fixed, trainable parameter θ only comes from the invertible flows F_l, for $1 \leq l \leq L$.

Jeffreys divergence While the KL divergence is widely used as the training objective for its easiness to compute, its landscape could admit local minima that don't favor the optimization. In fact, (Nielsen & Nock, 2009) suggests that $D_{\mathrm{KL}}(p_\theta \| q)$ is zero-forcing, meaning that it enforces p_θ be small whenever q is small. As a consequence, mode missing can still be a local minimum, see Appendix B. Therefore, we turn to the Jeffreys divergence (4) which significantly penalizes mode missing and can remove such local minima.

Estimating 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, we do importance sampling with $\tilde{q}$ as the proposal distribution for the Jeffreys divergence and its derivative:

Theorem 4.1 *The Jeffreys divergence and its derivative to θ admit the following formulation which can be estimated by the Monte Carlo method without samples from q,*

$$D_{\mathrm{J}}(p_\theta \| q) = \mathbb{E}_{p_\theta}\left[\log \frac{p_\theta}{q}\right] + \mathbb{E}_{\tilde{q}}\left[\frac{q}{\tilde{q}} \log \frac{q}{p_\theta}\right] \,. \tag{10}$$

$$\begin{aligned} \frac{\partial}{\partial \theta} D_{\mathrm{J}}(p_\theta \| q) = & \; \mathbb{E}_{p_\theta}\left[\left(1 + \log \frac{p_\theta}{q}\right) \frac{\partial \log p_\theta}{\partial \theta}\right] \\ & - \mathbb{E}_{\tilde{q}}\left[\frac{q}{\tilde{q}} \frac{\partial \log p_\theta}{\partial \theta}\right] \,. \end{aligned} \tag{11}$$

Furthermore, the Monte Carlo estimation doesn't need the normalizing constant Z in (1) as it can cancel itself.

Detailed derivation is left in Appendix C. With the derivative given above, we optimize the Jeffreys divergence by stochastic gradient descent. We remark that $\partial \log p_\theta / \partial \theta$ is available by the backward propagation of MsIGN, and $\tilde{q}$ comes from coarser scale model in multi-stage training.

Multi-stage training The multiscale design of MsIGN enables a coarse-to-fine multi-stage training. At stage l, we target at capturing the posterior q_l at scale l, and only train invertible flows before or at this scale: $F_{l'}$, with $l' \leq l$.

Additionally, at stage l, we initialize F_l as the identity map, and $F_{l'}$, with $l' < l$, as the trained model at stage $l - 1$. The reason is implied by (8), where now q, q_c represents q_l, q_{l-1} respectively. The stage $l - 1$ model provides good approximation to q_{l-1}, and together with PC_l it provides a good approximation $\tilde{q}_l$ to q_l. Thus, setting F_l as the identity map will give a good initialization to MsIGN in training. Our experiment shows such multi-stage strategy significantly stabilizes training and improves final performance.

We conclude the training of MsIGN in Algorithm 1.

Algorithm 1 Train MsIGN by optimizing the Jeffreys divergence in a multi-stage manner

```
Output: $\theta = (\theta_1, \ldots, \theta_L)$, $\theta_l$ are parameters in $F_l$.
Pre-compute and fix all prior conditioning layers $PC_l$.
Learn $\theta_1$ by standard methods such that sample $x_1 = F_1(z_1)$ approximates $q_1$.
for $l = 2$ to $L$ do
    Initialize $\theta_l$ so that $F_l$ is an identity map.
    Concatenate last-stage model with $PC_l$ as $\tilde{q}_l$.
    With $q_l$ as the target $q$, $\tilde{q}_l$ as the proposal $\tilde{q}$ in (11), compute the gradient using Monte Carlo.
    Learn $\theta_l$ by stochastic gradient descent.
end for
```

5. 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), and variational inference in providing a tight evidence lower bound, see, e.g, (Rezende & Mohamed, 2015). In this paper, our proposed MsIGN 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, MsIGN adopts a novel multiscale structure such that different scales can be trained separately, making training much more stable. Besides, the multiscale idea enables better explain-ability of its hidden neurons. Invertible generative models like (Dinh et al., 2016; Kingma & Dhariwal, 2018; Ardizzone et al., 2019) adopted a similar multiscale idea, but their multiscale strategy is not in a "spatial" sense: the intermediate neurons are not semantically interpret-able as shown in Figure 7. The multiscale idea is also used in generative adversarial networks (GANs), as in (Denton et al., 2015; Odena et al., 2017; Karras et al., 2017; Xu et al., 2018). But the lack of invertibility in these models makes it difficult for them to apply to Bayesian inference problems.

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 main goal of this paper 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; Parno et al., 2016; Matthies et al., 2016; Spantini et al., 2018), 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 to minimize the Jeffreys divergence, which has fewer local minima and better landscape compared with the commonly-used KL divergence, and to train MsIGN in a coarse-to-fine manner.

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; Chen & Ghattas, 2020). 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 6.1.3. We leave more discussion about the related work in Appendix E.

6. Experiment

We study two high-d Bayesian inverse problems (BIPs) in Section 6.1 as test beds for distribution approximation and multi-mode capture. We also apply MsIGN to the image synthesis task to benchmark with flow-based generative models and demonstrate its interpret-ability in Section 6.2.

In both experiments, we utilize average pooling with kernel size 2 and stride 2 as the operator A, and stack several of the invertible block in Glow (Kingma & Dhariwal, 2018) to build our invertible flow F, as mentioned in Section 3.

6.1. Bayesian Inverse Problems

We study two nonlinear and high-d BIPs known to have at least two equally important modes in this section: one with true samples available as reference in Section 6.1.1; one without true samples but close to real-world applications of subsurface flow in fluid dynamics in Section 6.1.2. In both problems, sample x of the target posterior q is a vector on a 2-D uniform 64×64 lattice, which means the problem dimension d is 4096. Every 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 equip x with a Gaussian prior $\mathcal{N}(0, \Sigma)$ with Σ as the discretization of $\beta^2 (-\Delta)^{-1-\alpha}$, where α, β are parameters.

To make the high-d inference more challenging, the target q is built to be multi-modal by leveraging spatial symmetry. Combining properties of the prior defined above and the likelihood defined afterwards, the posterior is innately mirror-symmetric: $q(x) = q(x')$ if $x(s_1, s_2) = x'(s_1, 1 - s_2)$ for any $s = (s_1, s_2) \in \Omega$. Furthermore, we carefully select the prior and the likelihood so that q has at least two modes. They are mirror-symmetric to each other and possess equal importance, see discussion in Appendix F.

We train MsIGN following Algorithm 1 with $L = 6$ scales. The problem dimension at scale l is $d_l = 2^l * 2^l = 4^l$. We compare MsIGN with representative approaches for high-d BIPs: Hamiltonian Monte Carlo (short as HMC) (Neal et al., 2011), SVGD (Liu & Wang, 2016), amortized-SVGD (short as A-SVGD) (Feng et al., 2017), and projected SVGD (short as pSVGD) (Chen & Ghattas, 2020). Since simulating the forward map $\mathcal{F}$ dominates the training time cost, especially in Section 6.1.2 (more than 75% of the wall clock time), we set a budget for the **n**umber of **f**orward **s**imulations (nFSs) for all methods for fair comparison in computational cost. For both problems, we aim at generating 2500 samples from the target. More details of experimental setting and additional numerical results can be found in Appendix F.

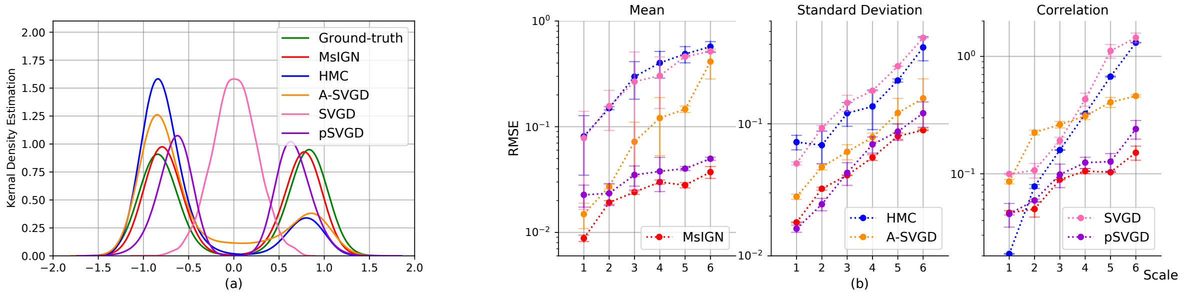

Figure 3. Results in the synthetic BIP. (a): Sample marginal distribution along the critical direction w_{k^*}. MsIGN is more robust in capturing both modes and close to ground-truth. (b): Root mean square error (RMSE) and its 95% confidence interval of three independent experiments. MsIGN is more accurate in distribution approximation, especially at finer scale when the problem dimension is high. The margin is statistical significant as shown by the confidence interval.

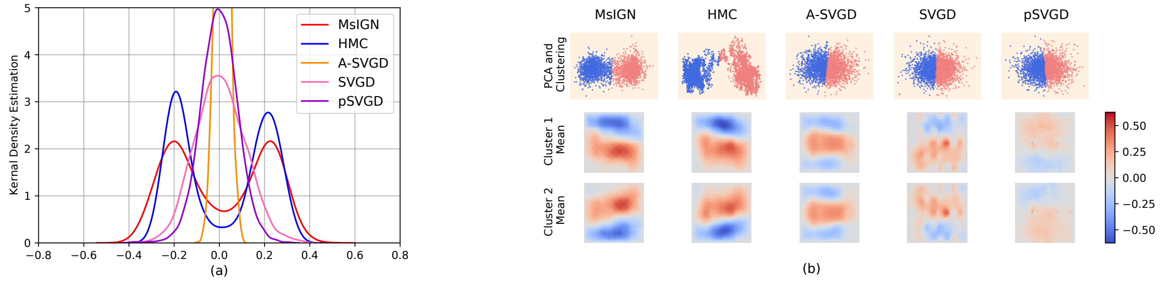

Figure 4. Results in the elliptic BIP. (a): Sample marginal distribution along the critical direction. MsIGN and HMC capture two modes in this marginal distribution, but the others fail. (b): Clustering result of samples. Samples of MsIGN are more balanced between two modes. The similarity of the cluster means of MsIGN and HMC implies that they both are likely to capture the correct modes.

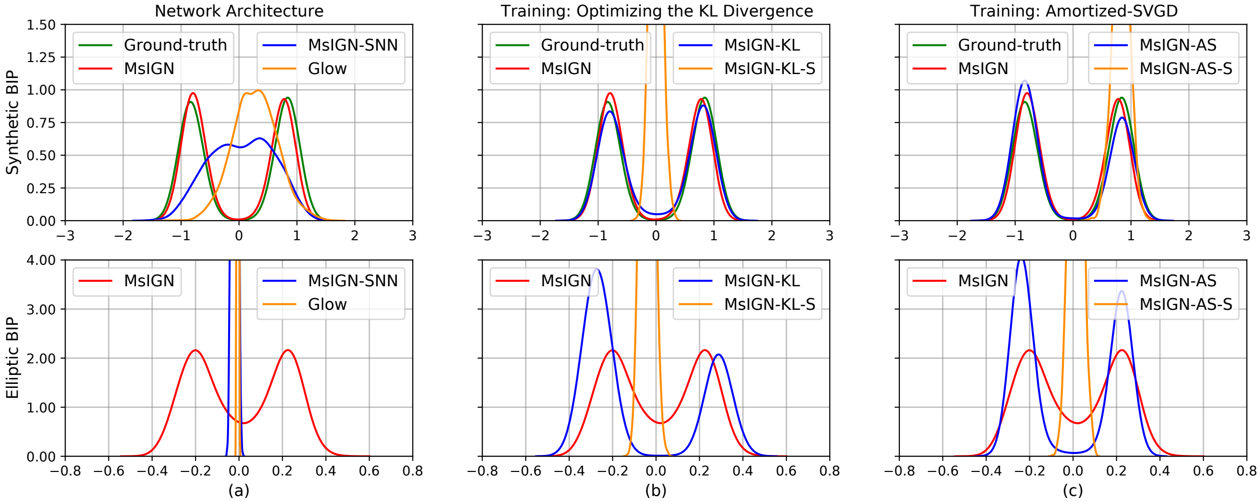

Figure 5. Ablation study of the network architecture and training strategy. "MsIGN" means our default setting: training MsIGN with Jeffreys divergence and multi-stage strategy. Other models are named by a base model (MsIGN or Glow), followed by strings indicating its difference from the default setting. For example, "MsIGN-KL" refers to training MsIGN by optimizing the KL divergence in a multi-stage way, while "MsIGN-AS-S" means training MsIGN using the amortizied-SVGD algorithm in a single-stage way. See Section 6.1.3 and Appendix F for thorough discussion.

6.1.1. Synthetic Bayesian Inverse Problems

This problem allows access to ground-truth samples so the comparison is clear and solid. We set $\mathcal{F}(x) = \langle \varphi, x \rangle^2 = (\int_\Omega \varphi(s)x(s)\mathrm{d}s)^2$, where $\varphi(s) = \sin(\pi s_1)\sin(2\pi s_2)$. Together with the prior, our posterior can be factorized into 1-D sub-distributions, namely $q(x) = \prod_{k=1}^d q_k(\langle w_k, x \rangle)$ for some orthonormal basis $\{w_k\}_{k=1}^d$ of $\mathbb{R}^d$. This property gives us access to true samples via inversion cumulative function sampling along each direction w_k. Furthermore, these 1-D sub-distributions are all single modal except that there's one, which is the marginal distribution q_{k^*} along direction w_{k^*}, with two symmetric modes. This confirms our construction of two equally important modes. The computation budget is fixed at 8×10^5 nFSs.

Multi-mode capture To visualize mode capture, we plot the marginal distribution of generated samples along the critical direction w_{k^*}, which is the source of double-modality. From Figure 3(a), MsIGN gives the best mode capture among our baselines in this $d = 4096$ problem.

Distribution approximation We use the root mean square errors (RMSE) of sample mean, standard deviation, and correlation, with the Jeffreys divergence to measure distribution approximation. We compare the sample mean, variance and correlation with theoretical ground-truths, and report the averaged RMSE of all sub-distributions at all scales in Figure 3(b). Additionally, since MsIGN and A-SVGD also gives density estimation, we report the Monte Carlo estimates of the Jeffreys divergence (4) with the target posterior in Table 1. We can see that MsIGN has superior accuracy in approximating the target distribution.

Table 1. Jeffreys divergence $D_{\mathrm{J}}(p_\theta|q)$ in three independent runs.

Model	MsIGN	A-SVGD
Error	56.77±0.15	3372±21

6.1.2. Elliptic Bayesian Inverse Problems

This problem is a benchmark problem for high-d inference from geophysics and fluid dynamics (Iglesias et al., 2014; Cui et al., 2016). The forward map $\mathcal{F}(x) = \mathcal{O} \circ \mathcal{S}(x)$, where $u = \mathcal{S}(x)$ is the solution to an elliptic partial differential equation with zero Dirichlet boundary condition:

$$-\nabla \cdot \left(e^{x(s)} \nabla u(s)\right) = f(s)\,, \quad s \in \Omega\,,$$

And $\mathcal{O}$ is linear measurements of the field function u:

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

where f and φ_k are given and fixed. The map $\mathcal{S}$ is solved by the finite element method with mesh size $1/64$. Unfortunately, there is no known access to true samples of q. But the trick of symmetry introduced in Section 6.1 guarantees at least two equally important modes in the posterior. We put a 5×10^5-nFS budget on our computation cost.

Multi-mode capture Due to the lack of true samples, we check the marginal distribution of the posterior along eigenvectors of the prior, and pick a particular one to show if we capture double modes in Figure 4(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 4(b), where we can see that MsIGN has a more balanced capture of the symmetric posterior than HMC, while others fail to detect two modes. We refer readers to Appendix F for more comprehensive study of mode capture ability of different methods.

6.1.3. Ablation Study

We run extensive experiments to study the effectiveness of the network architecture and training strategy of MsIGN. Detailed setting and extra results are left in Appendix F.

Network architecture We replace the prior conditioning layer PC by two direct alternatives: a stochastic nearest-neighbor upsample layer independent of the prior (model named "MsIGN-SNN"), or the split and squeeze layer in Glow design (it resumes Glow model, so we call it "Glow"). Figure 5(a) shows that the prior conditioning layer design is crucial to the performance of MsIGN on both problems, because neither alternatives has a successful mode capture.

Training strategy We study the effectiveness of the Jeffreys divergence objective and multi-stage training. We substitute our optimizing of the Jeffreys divergence in Algorithm 1 by optimizing the KL divergence (marked with a suffix "-KL"), or using the amortized-SVGD (A-SVGD) method (marked with a suffix "-AS"). We also switch between multi-stage (the default, no extra suffix) or single-stage training (marked with an additional suffix "-S"). We remark that single-stage training using Jeffreys divergence is infeasible because of the difficulty to estimate $D_{\mathrm{KL}}(q\|p_\theta)$. Figure 5(b) and (c) show that, all models trained in the single-stage manner ("MsIGN-KL-S", "MsIGN-AS-S") will face mode collapse. We observe that our multi-stage training strategy can benefit training with other objectives, see "MsIGN-KL" and "MsIGN-AS". We also notice that the Jeffreys divergence leads to a more balanced samples for these symmetric problems, especially for the complicated elliptic BIP.

6.2. Image Synthesis

The transport map approach to Bayesian inference has two critical difficulties: the model capacity and the training effectiveness. Since the distribution of images is complicated

Table 2. Bits-per-dimension value comparison with baseline models of flow-based generative networks. All models in this table do not use the "variational dequantization" technique in (Ho et al., 2019). *: 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**

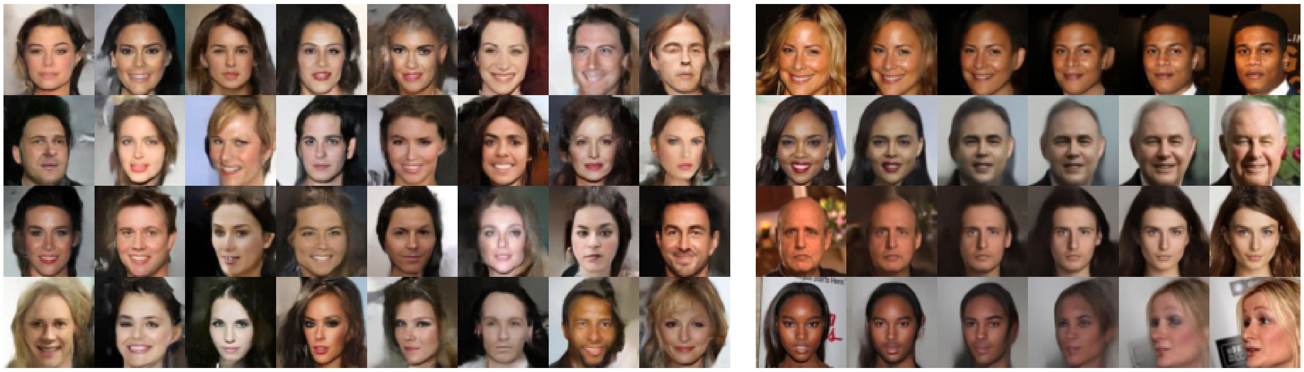

Figure 6. Left: Synthesized CelebA images of resolution 64×64 with temperature 0.9. Right: Linear interpolation in latent space shows MsIGN's parameterization of natural image manifold is semantically meaningful. For images x_1, x_2 at the left and right ends, we retrieve their latent feature by $z_i = T^{-1}(x_i; \theta), i = 1, 2$, and then interpolate between them by $T((1-\lambda)z_1 + \lambda z_2; \theta)$ for $\lambda = 0.2, 0.4, 0.6, 0.8$.

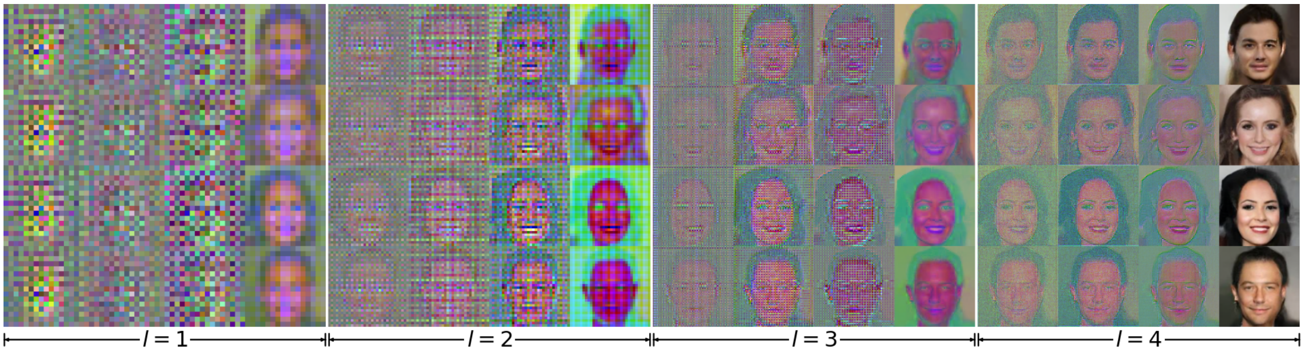

Figure 7. Visualization of internal activation shows the interpret-ability of MsIGN hidden neurons. This MsIGN model has $L = 4$ scales. From left to right, we take 4 snapshots (head, two trisection points, and tail) in each invertible flow F_l for $l = 1, 2, 3, 4$, to show how MsIGN progressively generates new samples from low to high resolution.

and multi-modal, we present the image synthesis task result to show case the model capacity of the MsIGN. It also provides a good test bed for our MsIGN to benchmark with other flow-based generative networks.

We train MsIGN by maximum likelihood estimation. We assume a simple Gaussian prior ρ for natural images, whose covariance is a scalar matrix learned from the data. See Appendix G for experimental details and additional results.

We report the bits-per-dimension value comparison with baseline models in Table 2. Our MsIGN is superior in number and also 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.

Figure 6 shows synthesized images of MsIGN from CelebA data set, and linear interpolation of real images in the latent feature space. In Figure 7, we visualize internal activations at checkpoints in the invertible flow at different scales which demonstrate the interpret-ability of MsIGN.

7. Conclusion

For high-dimensional Bayesian inference problems with multiscale structure, we propose Multiscale Invertible Generative Networks (MsIGN) and associated training algorithms to approximate the posterior. We demonstrate the potential of this approach in high-dimensional (up to 4096) Bayesian inference problems, leaving several important directions as future work. The network architecture also achieves superior performance over benchmarks in various image synthesis tasks. We plan to apply this methodology to other Bayesian inference problems, e.g., 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)).

Acknowledgements

We appreciate valuable comments and discussion from the reviewers. The work of Thomas Y. Hou was in part supported by NSF Grants DMS-1912654 and DMS-1907977, and Shumao Zhang was in part supported by NSF Grant DMS-1912654. The authors would also like to thank Ka Chun Lam and Xiaodong He for their helpful discussion and thank Lin Xiao, Qiang Liu and Bo Dai for their helpful suggestions.

References

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., and Köthe, U. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*, 2018.

Ardizzone, L., Lüth, C., Kruse, J., Rother, C., and Köthe, U. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., and Jacobsen, J.-H. Invertible residual networks. In *International Conference on Machine Learning*, pp. 573–582, 2019.

Beskos, A., Roberts, G., Stuart, A., and Voss, J. Mcmc methods for diffusion bridges. *Stochastics and Dynamics*, 8(03):319–350, 2008.

Chang, B., Meng, L., Haber, E., Tung, F., and Begert, D. Multi-level residual networks from dynamical systems view. *arXiv preprint arXiv:1710.10348*, 2017.

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

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

Chen, P. and Ghattas, O. Projected stein variational gradient descent. *arXiv preprint arXiv:2002.03469*, 2020.

Chen, P., Wu, K., Chen, J., O'Leary-Roseberry, T., and Ghattas, O. 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.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691, 2014.

Chen, T. Q., Behrmann, J., Duvenaud, D. K., and Jacobsen, J.-H. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, pp. 9913–9923, 2019b.

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

Cui, T., Law, K. J., and Marzouk, Y. M. Dimension-independent likelihood-informed mcmc. *Journal of Computational Physics*, 304:109–137, 2016.

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

Denton, E. L., Chintala, S., szlam, a., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1486–1494, 2015.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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

Feng, Y., Wang, D., and Liu, Q. Learning to draw samples with amortized stein variational gradient descent. *arXiv preprint arXiv:1707.06626*, 2017.

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

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Haber, E., Ruthotto, L., Holtham, E., and Jun, S.-H. Learning across scales—multiscale methods for convolution neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pp. 2722–2730. PMLR, 2019.

Hou, T. Y., Lam, K. C., Zhang, P., and Zhang, S. Solving bayesian inverse problems from the perspective of deep generative networks. *Computational Mechanics*, 64(2): 395–408, 2019.

Iglesias, M. A., Lin, K., and Stuart, A. M. Well-posed bayesian geometric inverse problems arising in subsurface flow. *Inverse Problems*, 30(11):114001, 2014.

Jeffreys, H. et al. *Scientific inference*. Cambridge University Press, 1973.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Kruse, J., Ardizzone, L., Rother, C., and Köthe, U. Benchmarking invertible architectures on inverse problems. In *Thirty-sixth International Conference on Machine Learning*, 2019.

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

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

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

Matthies, H. G., Zander, E., Rosić, B. V., Litvinenko, A., and Pajonk, O. Inverse problems in a bayesian setting. In *Computational Methods for Solids and Fluids*, pp. 245–286. Springer, 2016.

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

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

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

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pp. 2642–2651, 2017.

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

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows, 2015.

Spantini, A., Solonen, A., Cui, T., Martin, J., Tenorio, L., and Marzouk, Y. Optimal low-rank approximations of bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

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

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

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.