3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–12

Preconditioned training of normalizing flows for variational inference in inverse problems

Anonymous Authors
Anonymous Institution

Abstract

In the context of inverse problems with computationally expensive forward operators, specially for domains with limited access to high-fidelity training unknown and observed data pairs, we propose a preconditioned scheme for training a conditional normalizing flow (NF) capable of directly sampling the posterior distribution. Our training objective consists of the Kullback-Leibler divergence between the predicted and the desired posterior density. To minimize the costs associated with the forward operator, we initialize the NF via the weights of another pretrained low-fidelity NF, which is trained beforehand on available low-fidelity model and data pairs. Our numerical experiments, including a 2D toy and a seismic image compressed sensing example, demonstrate the improved performance and speed-up of the proposed method compared to training a NF from scratch.

1. Introduction

We attempt to perform approximate Bayesian inference in inverse problems characterized by computationally expensive forward operators, $F : \mathcal{X} \to \mathcal{Y}$, with data likelihood, $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x})$:

$$\boldsymbol{y} = F(\boldsymbol{x}) + \boldsymbol{\epsilon}, \tag{1}$$

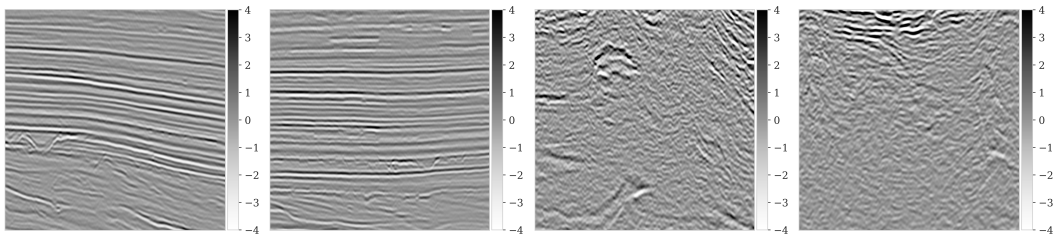
where $\boldsymbol{x} \in \mathcal{X}$ is the unknown model, $\boldsymbol{y} \in \mathcal{Y}$ the observed data, and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ the measurement noise. Given a prior density, $\pi_{\text{prior}}(\boldsymbol{x})$, we can use variational inference (VI, Jordan et al., 1999), based on normalizing flows (NFs, Rezende and Mohamed, 2015), where we minimize the Kullback-Leibler (KL) divergence (Liu and Wang, 2016; Kruse et al., 2019) between the predicted and the high-fidelity posterior density, $\pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z}_x)} \left[\frac{1}{2\sigma^2} \left\| F\big(T_\theta(\boldsymbol{z}_x)\big) - \boldsymbol{y} \right\|_2^2 - \log \pi_{\text{prior}}\big(T_\theta(\boldsymbol{z}_x)\big) - \log \left| \det \nabla_{z_x} T_\theta(\boldsymbol{z}_x) \right| \right]. \tag{2}$$

In the above expression, $T_\theta : \mathcal{Z}_x \to \mathcal{X}$ denotes the NF with parameters $\boldsymbol{\theta}$ and a Gaussian latent variable $\boldsymbol{z}_x \in \mathcal{Z}_x$. After training, we sample the approximated posterior, $\pi_\theta(\boldsymbol{x} \mid \boldsymbol{y}) \approx \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$, by evaluating $T_\theta(\boldsymbol{z}_x)$ for $\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z}_x)$ (Kruse et al., 2019). While the above VI formulation in principle allows us to train a NF to generate samples from the posterior given a single observation $\boldsymbol{y}$, it requires access to a prior density, and the training calls for repeated evaluations of the forward operator and the adjoint of its Jacobian. As in multi-fidelity Markov chain Monte Carlo (MCMC) sampling (Peherstorfer and Marzouk, 2018), the costs associated with the forward model may become prohibitive even though VI-based methods are known to have computational advantages over MCMC (Blei et al., 2017).

Aside from the above computational considerations, reliance on having access to a prior may be problematic especially when dealing with images of the Earth subsurface that are the

© A. Authors.



(a) (b) (c) (d)

Figure 1: Subsurface images estimated from a real seismic survey, indicating strong heterogeneity between (a), (b) shallow and (c), (d) deep parts of the same survey area.

result of complex geological processes that do not lend themselves to be easily captured by hand-crafted priors. Under these circumstances, data-driven priors or even better data-driven posteriors obtained by training over model and data pairs sampled from the joint distribution, $\widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$ are preferable. Following Kruse et al. (2019), Kovachki et al. (2020), and Baptista et al. (2020) this type of training can be formulated in terms of a block triangular conditional NF, $G_\phi : \mathcal{Y} \times \mathcal{X} \to \mathcal{Z}_y \times \mathcal{Z}_x$, with latent space $\mathcal{Z}_y \times \mathcal{Z}_x$:

$$\begin{aligned} &\min_{\phi} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}\sim\widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x})} \left[\frac{1}{2} \left\| G_\phi(\boldsymbol{y}, \boldsymbol{x}) \right\|^2 - \log \left| \det \nabla_{y,x} G_\phi(\boldsymbol{y}, \boldsymbol{x}) \right| \right], \\ &\text{where} \quad G_\phi(\boldsymbol{y}, \boldsymbol{x}) = \begin{bmatrix} G_{\phi_y}(\boldsymbol{y}) \\ G_{\phi_x}(\boldsymbol{y}, \boldsymbol{x}) \end{bmatrix}, \ \boldsymbol{\phi} = \{\boldsymbol{\phi}_y, \boldsymbol{\phi}_x\}. \end{aligned} \tag{3}$$

Due to the block-triangular structure of G_ϕ, we obtain samples from the approximated posterior, $\pi_\phi(\boldsymbol{x} \mid \boldsymbol{y}) \approx \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$, by evaluating $G^{-1}_{\phi_x}(G_{\phi_y}(\boldsymbol{y}), \boldsymbol{z}_x)$ for $\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z}_x)$ (Kruse et al., 2019). Unlike the objective in Equation (2), training G_ϕ does not involve inversion of the forward operator, F. It also does not require specifying a prior density. However, its success during inference heavily relies on having access to high-fidelity training pairs from the joint distribution. Unfortunately, unlike medical imaging where data is abundant and variability among patients is relatively limited, we generally do not have access to high-fidelity samples from the joint distribution in geophysical applications. This lack of access to high-fidelity information, together with the Earth's strong variability, unfortunately limits the scope of the inference approach outlined in Equation (3). To illustrate the challenge imposed by the Earth's heterogeneity, we include two pairs of 256×256 patches sampled from shallow (Figures 1a and 1b) and deep (Figures 1c and 1d) parts of a 3D seismic image obtained from the Parihaka seismic dataset.

To meet the challenges of computational cost and lack of access to training pairs, we propose a preconditioning scheme where the two described VI methods are combined to:

1. maximally make use of information that is available in the form of samples from the *low-fidelity* joint distribution, $\widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$, to pretrain G_ϕ via the objective in Equation (2). We incur these costs one-time and this training can be done beforehand;

2. exploit the invertibility of $G_{\phi_x}(\boldsymbol{y}, \cdot)$, which gives us access to the low-fidelity posterior density, $\pi_\phi(\boldsymbol{x} \mid \boldsymbol{y})$. For a given $\boldsymbol{y}$, this learned prior can be used in Equation (2);

3. initialize T_θ with weights from the pretrained $G_{\phi_x}^{-1}$. This initialization is as an instance of transfer learning (Yosinski et al., 2014), and we expect a considerable speed-up when solving Equation (2), which is important since it involves inverting F.

2. Related work

We base our work on hierarchical conditional NFs described in Kruse et al. (2019), which due to their coupling blocks (for more details refer to Appendix A) provide a dense triangular Jacobian, as opposed to initially introduced coupling blocks (Dinh et al., 2014, 2016; Kingma and Dhariwal, 2018), which are known to require composition of many coupling blocks to achieve a similar representation power as other types of generative models.

In the context of variational inference in inverse problems, Kovachki et al. (2020) use a block triangular map between the joint model and data distribution and their respective latent spaces. By imposing an additional monotonicity constraint, they train a generative adversarial network (GAN, Goodfellow et al., 2014) to directly sample the posterior distribution. We choose to use NFs instead of GANs, since NFs due to their invertibility allow for memory efficient training (Leemput et al., 2019; Putzky and Welling, 2019; Peters et al., 2020; Peters and Haber, 2020) , which is suitable for large-scale problems. Parno and Marzouk (2018) use transport-based maps as non-Gaussian proposal distributions in the context of MCMC sampling, and by adaptively fine-tuning the proposal to the target density, they increase the efficiency of the Metropolis-Hasting sampling. Finally, Peherstorfer and Marzouk (2018) propose a preconditioned MCMC sampling technique via a transport-map based proposal distribution that is trained to sample a low-fidelity posterior distribution. The idea of multifidelity preconditioned MCMC in Peherstorfer and Marzouk (2018) directly inspires this work, however, we will not make use of MCMC. Instead, we setup a VI objective that tends to be faster and easier to scale to large-scale Bayesian inference problems (Blei et al., 2017).

3. Multifidelity preconditioning scheme

For an observation $\boldsymbol{y}$, we define a NF $T_{\phi_x}(\,\cdot\,) : \mathcal{Z}_x \to \mathcal{X}$ as

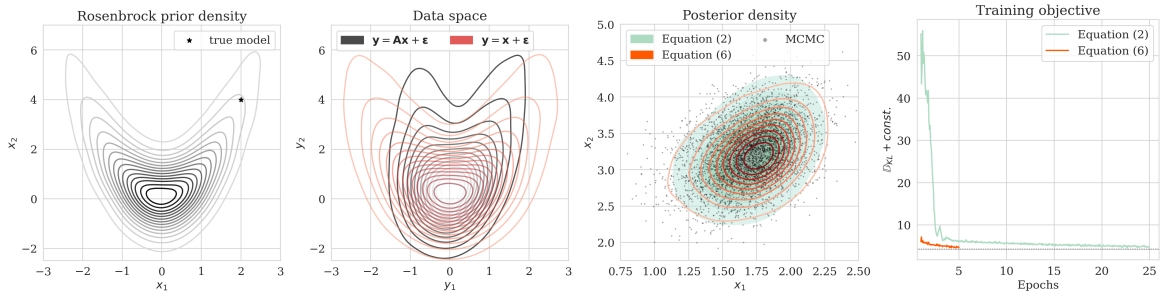
$$T_{\phi_x}(\boldsymbol{z}_x) := G_{\phi_x}^{-1}(G_{\phi_y}(\boldsymbol{y}), \boldsymbol{z}_x), \tag{4}$$

where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_y, \boldsymbol{\phi}_x\}$ is obtained by training G_ϕ via the objective in Equation (3). To train G_ϕ, we use available low-fidelity training pairs $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$. This training phase is performed beforehand, similar to the pretraining phase in the transfer learning (Yosinski et al., 2014) literature. Due to the invertibility of G_ϕ, it provides an expression for the low-fidelity posterior, which we will be using as a (conditional) prior in Equation (2):

$$\pi_{\text{prior}}(\boldsymbol{x}) := \pi_{\boldsymbol{z}_x}\big(G_{\phi_x}(\boldsymbol{x}, \boldsymbol{y})\big) \left|\det \nabla_x G_{\phi_x}(\boldsymbol{x}, \boldsymbol{y})\right|. \tag{5}$$

Finally, we set up the objective where we minimize the KL divergence between the predicted and the high-fidelity posterior density, $\pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$ (Liu and Wang, 2016; Kruse et al., 2019):

$$\min_{\phi_x} \mathbb{E}_{\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z}_x)} \left[\frac{1}{2\sigma^2} \left\| F\big(T_{\phi_x}(\boldsymbol{z}_x)\big) - \boldsymbol{y} \right\|_2^2 - \log \pi_{\text{prior}}\big(T_{\phi_x}(\boldsymbol{z}_x)\big) - \log \left| \det \nabla_{z_x} T_{\phi_x}(\boldsymbol{z}_x) \right| \right], \tag{6}$$



(a) (b) (c) (d)

Figure 2: (a) Prior, (b) low- and high-fidelity data, and (c) approximated posterior densities via MCMC (dark circles), and objectives in Equations (2) in green and (6) in orange. (d) training objective values during training via Equations (2) in green and (6) in orange.

where we use the prior density in Equation (5). We can interpret the problem in Equation (6) as an instance of transfer learning (Yosinski et al., 2014) for conditional NFs. This is particular useful for inverse problems with expensive forward operators where access to high fidelity training samples is limited. It is important to note that Equation (6) trains a NF specific to the observed data $\boldsymbol{y}$. In the next section, we present two numerical experiments intended to show the speed-up and accuracy gained via our proposed method.

4. Numerical experiments

In this section, we present two synthetic examples aimed at verifying the anticipated speed-up and increase in accuracy of the predicted posterior density via our preconditioning scheme. The first example is a two-dimensional problem where the posterior density can be accurately and cheaply sampled via MCMC. The second example demonstrates the effect of the preconditioning scheme in a seismic image compressed sensing (Candes et al., 2006; Donoho, 2006) problem. Details regarding the training can be found in Appendix A.

4.1. 2D toy example

In this example, both the low- and high-fidelity prior densities of $\boldsymbol{x} \in \mathbb{R}^2$ are the 2D Rosenbrock distribution, $\pi_{\text{prior}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}x_1^2 - \left(x_2 - x_1^2\right)^2\right)$, depicted in Figure 2a. High-fidelity data, $\boldsymbol{y} \in \mathbb{R}^2$, are generated via $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, 0.4^2\boldsymbol{I})$ and $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ is the forward operator. $\boldsymbol{A}$ is equal to $\bar{\boldsymbol{A}} \,/\, \rho(\bar{\boldsymbol{A}})$, where $\rho(\cdot)$ is the spectral radius of $\bar{\boldsymbol{A}} = \boldsymbol{\Gamma} + \gamma\boldsymbol{I}$ with $\boldsymbol{\Gamma}_{ij} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\gamma = 3$. We omit the forward operator while simulating low-fidelity data samples. Figure 2b depicts the low- (purple) and high-fidelity (red) data densities. The dark star represents the unknown model. By choosing smaller values for γ, we make $\boldsymbol{A}$ less close to the identity matrix, hence more discrepancy between low- and high-fidelity posterior. For more experiments regarding different values for γ we refer to Appendix B.

To show the speed-up and accuracy of the proposed method, we compare the approximated posterior density via the preconditioned scheme (orange contours) to the approximation obtained via the objective of Equation (2) in Figure (2c). We also include a comparison

with posterior samples (dark circles) obtained via stochastic gradient Langevin dynamics (Welling and Teh, 2011), an MCMC sampling technique. As expected, the estimated posterior densities with and without the preconditioning scheme are in agreement with the MCMC samples. The plot in Figure 2d compares the objective values of Equations (2) and (6) with orange and green, respectively. The comparison clearly shows the computational superiority of the proposed preconditioning approach.

4.2. Seismic image compressed sensing

This experiment intends to show challenges with geophysical inverse problems due to the Earth's strong heterogeneity. Here we invert for $\boldsymbol{x} \in \mathbb{R}^{256\times 256}$ patches sampled from deeper parts of the Parihaka seismic dataset. We generate high-fidelity sensed data via $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, 0.2^2\boldsymbol{I})$. In this example, $\boldsymbol{A} = \bar{\boldsymbol{A}}^T\bar{\boldsymbol{A}}$, where $\bar{\boldsymbol{A}}$ is a compressing sensing matrix with 66.66 % subsampling rate, and $\boldsymbol{y}$ is a pseudo-recovered model, contaminated with noise.

For pretraining, we change both the likelihood model and the prior distribution. We obtain low-fidelity joint training pairs, $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$, by using patches sampled from the shallow part of the Parihaka dataset, which have a different texture compared to patches from the deeper parts. By changing the distribution of noise to a Gaussian distribution with standard deviation 0.01, we create low-fidelity data .

We first show the performance of the pretrained network on low-fidelity posterior density estimation. Figures 3a and 3b correspond to joint samples from the low-fidelity distributions not used in pretraining. Figures 3c and 3d show the estimated low-fidelity posterior mean and pointwise standard deviation (STD), respectively. For more similar results refer to Appendix C. From these results, we observe that the pretrained NF is able to produce a reasonable recovery of the model, especially for the high amplitude events. As expected, the pointwise STD plot indicates more uncertainty in areas with low-amplitude events. Next, we showcase the performance of the preconditioned training scheme on high-fidelity data. Figures 3e and 3f include joint samples from the high-fidelity joint distribution. By comparing the approximated conditional mean estimate via the pretrained NF (Figure 3g) to the approximation via the preconditioned scheme (Figure 3h), we observe that estimates based on the latter result in better recovery of low-amplitude, fine-scale events (4.5 dB improvement compared to Figure 3g). See figures 3i and 3j for the relative reconstruction error associated with the approaches outlined in Equations (2) and (6), respectively. Finally, the pointwise STD obtained via the pretrained network (Figure 3k) is essentially the pointwise second moment of the prior used for unsupervised training. Therefore, as expected, the pointwise STD in Figure 3l has been narrowed down compared to Figure 3k, while still indicating less uncertainty in areas with high-amplitude events.

We did not include results obtained via Equation (2) because the optimization got stuck in undesired local minimum. We believe this is because the learned prior in Equation (5) has many local minima. Starting the optimization with random weights often a remedy to avoid local minima was apparently inadequate. The proposed initialization overcomes this issue.

5. Conclusions

Inverse problems in fields such as seismology are challenging for several reasons. The forward operators are complex and expensive to evaluate numerically while the Earth is highly

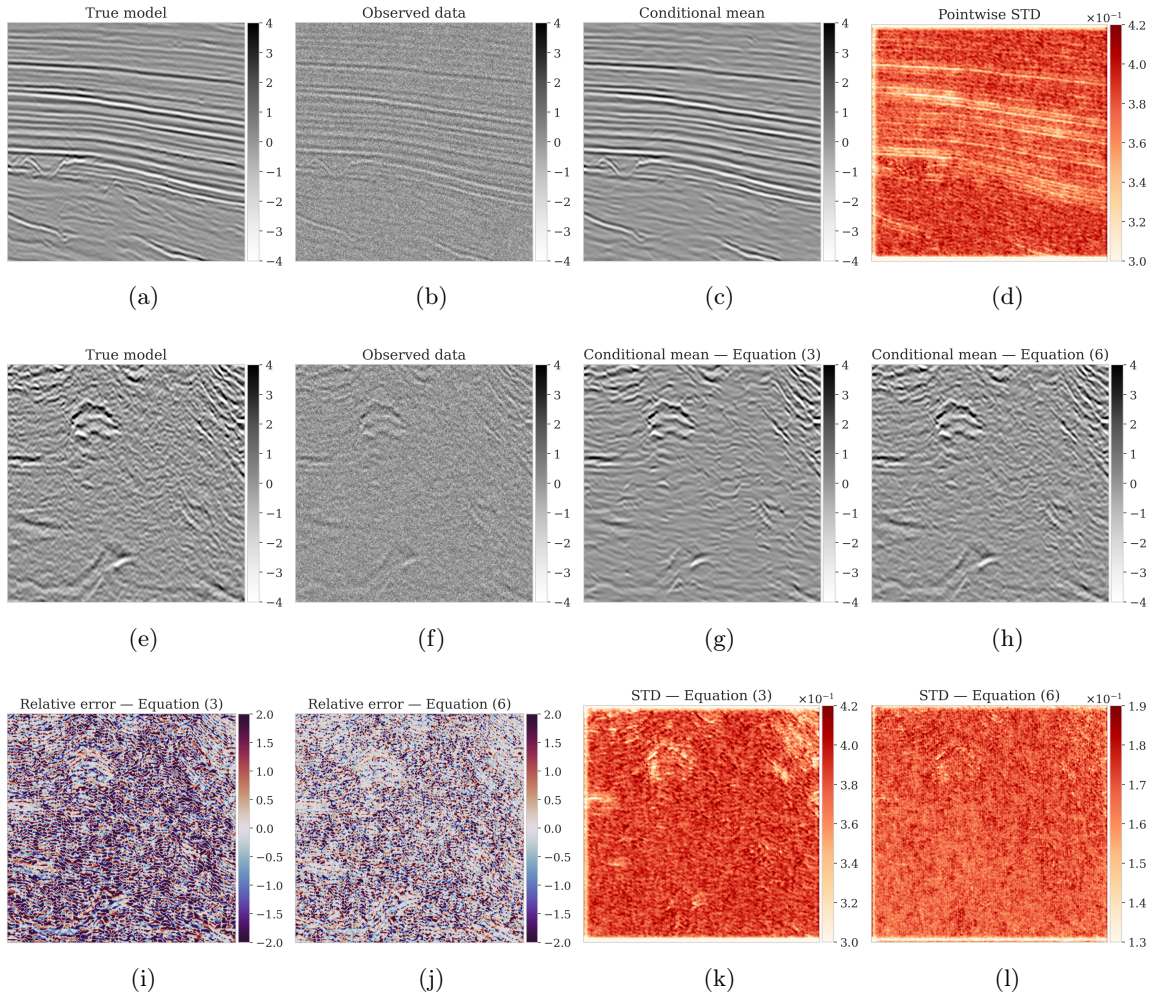


Figure 3: Seismic imaging compressed sensing. First row indicates the performance of the pretrained network on low-fidelity posterior inference. Second row is compares the pretraining-based recovery with the accelerated scheme result. Last row compares the recovery errors and pointwise STDs for the two recoveries.

variable. To handle this situation and to quantify uncertainty, we propose a preconditioned scheme for training normalizing flows for Bayesian inference. The proposed scheme is designed to take full advantage of having access to training pairs drawn from a joint distribution, which for the reasons stated above is close but not equal to the actual joint distribution. We use these samples to train a normalizing flow by minimizing the Kullback-Leibler divergence between the predicted and the desired posterior density. After this pretraining, we deploy this low-fidelity estimate for the posterior as a prior and preconditioner for the actual variational inference on the observed data. By means of a series of examples, we demonstrate that our preconditioned scheme leads to speed-ups compared to solely relying on the low-fidelity posterior.

References

Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. An adaptive transport framework for joint and conditional density estimation. *arXiv preprint arXiv:2009.10303*, 2020.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP, 2016.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2014.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90. URL https://ieeexplore.ieee.org/document/7780459.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999. doi: 10.1023/A:1007665907178.

D. P. Kingma and P. Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980, 2014.

Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. Conditional sampling with monotone gans. *arXiv preprint arXiv:2006.06755*, 2020.

J. Kruse, G. Detommaso, R. Scheichl, and U. Köthe. HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference, 2019.

Sil C. van de Leemput, Jonas Teuwen, Bram van Ginneken, and Rashindra Manniesing. Memcnn: A python/pytorch package for creating memory-efficient invertible neural networks. *Journal of Open Source Software*, 4(39):1576, 7 2019. ISSN 2475-9066. doi: 10.21105/joss.01576. URL http://dx.doi.org/10.21105/joss.01576.

Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, volume 29, pages 2378–2386. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

M. D. Parno and Y. M. Marzouk. Transport Map Accelerated Markov Chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018. doi: 10.1137/17M1134640.

Benjamin Peherstorfer and Youssef Marzouk. A transport-based multifidelity preconditioner for Markov chain Monte Carlo, 2018.

Bas Peters and Eldad Haber. Fully reversible neural networks for large-scale 3d seismic horizon tracking. In *82nd EAGE Annual Conference & Exhibition*, volume 2020, pages 1–5. European Association of Geoscientists & Engineers, 2020.

Bas Peters, Eldad Haber, and Keegan Lensink. Fully reversible neural networks for large-scale surface and sub-surface characterization via remote sensing. *arXiv preprint arXiv:2003.07474*, 2020.

Patrick Putzky and Max Welling. Invert to learn to invert. In *Advances in Neural Information Processing Systems*, pages 446–456, 2019.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538. PMLR, 07–09 Jul 2015. URL http://proceedings.mlr.press/v37/rezende15.html.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, pages 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195. doi: 10.5555/3104482.3104568. URL https://dl.acm.org/doi/abs/10.5555/3104482.3104568.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3320–3328, 2014. URL http://dl.acm.org/citation.cfm?id=2969033.2969197.

Appendix A. Training details and network architectures

We adapt the coupling blocks proposed by Kruse et al. (2019), which use the coupling layers introduced by Dinh et al. (2016) in a hierarchical way. In other words, we recursively divide the incoming states variables and apply a affine coupling layer. The final architecture is obtained by composing several of these hierarchical coupling blocks. The hierarchical structure leads to dense triangular Jacobian, which is essential in representation power of NFs.

For all examples in this paper we use the hierarchical coupling blocks, as described in Kruse et al. (2019), where the affine coupling layers within each hierarchal block has a residual block (He et al., 2016) as a *subnet* with 64 input, 128 hidden, and 64 output channels, except for the first and last coupling layer where we have 4 input and output channels, respectively. We use Wavelet transform and its transpose before feeding seismic images into and after the last layer of the NFs, respectively.

Below we describe network architectures and training details regarding the two numerical experiments described in the paper. Throughout the experiments, we use the Adam optimization algorithm (Kingma and Ba, 2014).

A.1. 2D toy example

We use 8 hierarchal coupling blocks, as described above for both G_{ϕ_x} and G_{ϕ_y}, described in Equation (3). As a result, due to our proposed method in Equation (4), we choose the same architecture for T_{ϕ_x} used in Equation (2).

For pretraining G_θ according to Equation (3), we use 5000 low-fidelity joint training pairs, $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$. We minimize Equation (3) for 25 epochs with batch size 64, starting learning rate 0.001. We decrease the learning rate each epoch by 0.9.

For the preconditioned step—i.e., solving Equation 6, we use 1000 latent training samples. We train for 5 epochs with batch size 64, and learning rate 0.001. Finally, as a comparison, we solve the objective in Equation 6 for a randomlly initialized NF with the same 1000 latent training samples for 25 epochs. We decrease the learning rate each epoch by 0.9.

A.2. Seismic image compressed sensing

We use 12 hierarchal coupling blocks, as described above for both G_{ϕ_x}, G_{ϕ_y}, and we use the same architecture for T_{ϕ_x} as G_{ϕ_x}.

For pretraining G_θ according to Equation (3), we use 5282 low-fidelity joint training pairs, $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$. We minimize Equation (3) for 50 epochs with batch size 4, starting learning rate 0.001. We decrease the learning rate each epoch by 0.9.

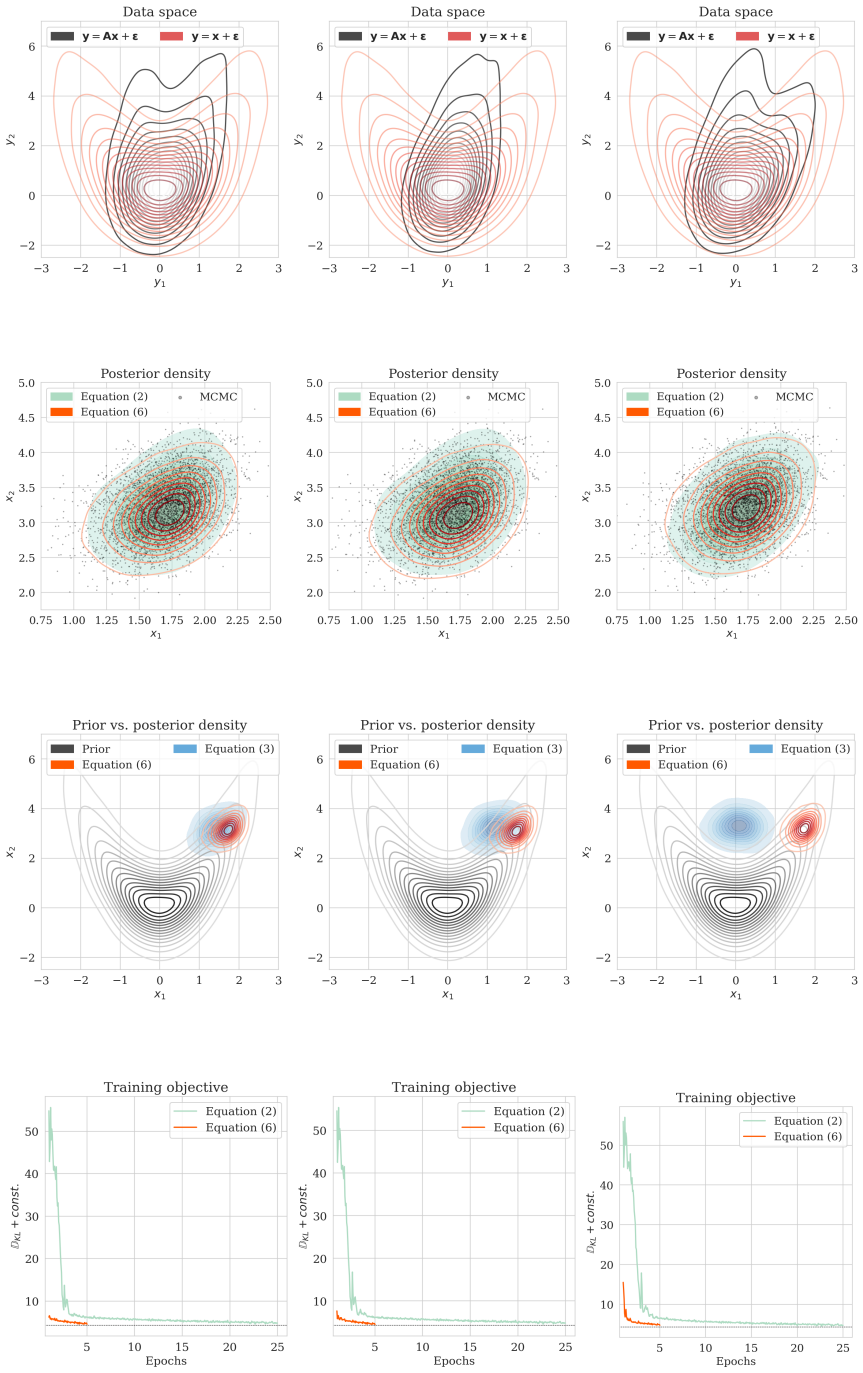
For the preconditioned step—i.e., solving Equation 6, we use 1000 latent training samples. We train for 10 epochs with batch size 4, and learning rate 0.0001, where we decay the step by 0.9 every 5th epoch.

Appendix B. 2D toy example—more results

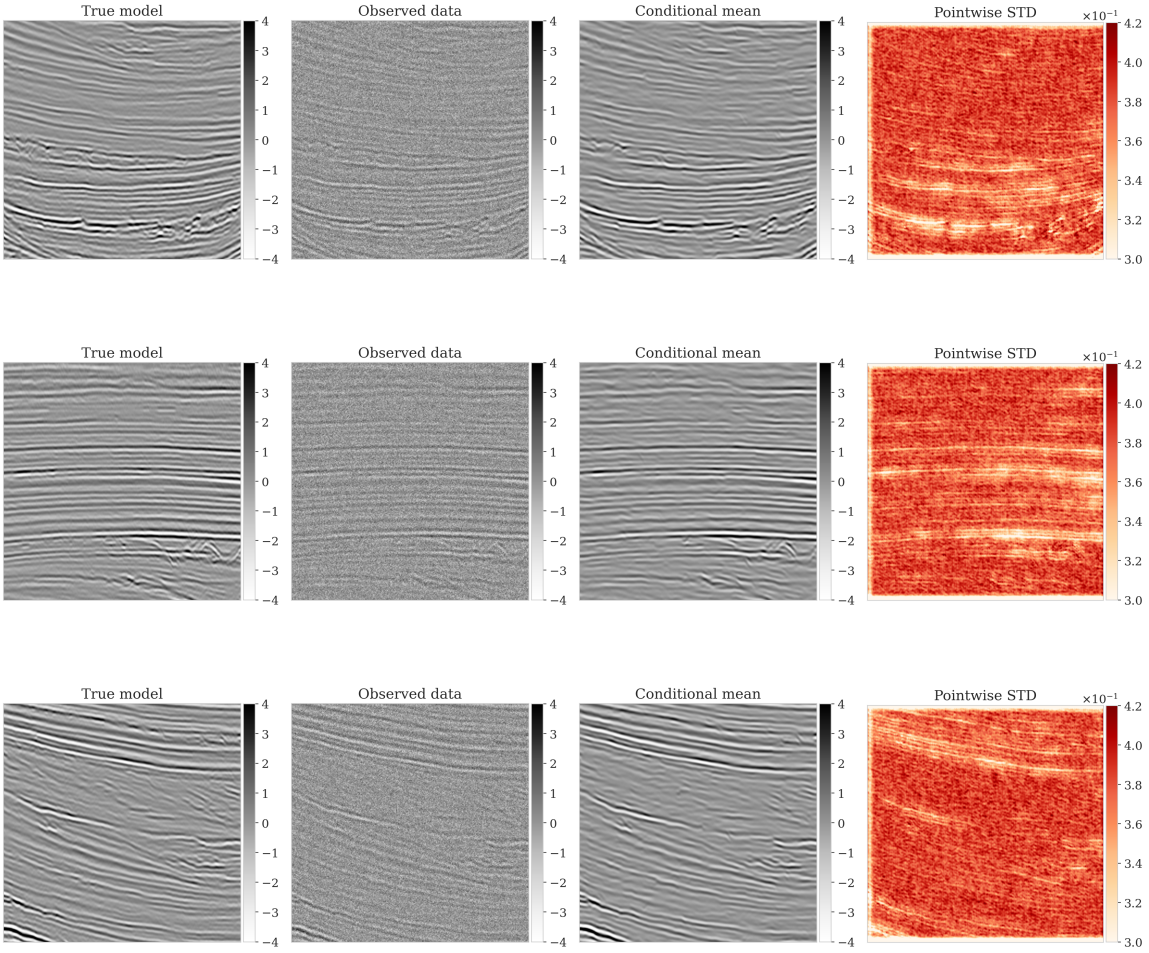
Here we show the effect γ on our proposed method in the 2D toy experiment. By choosing smaller values for γ, we make $\bar{\boldsymbol{A}} \,/\, \rho(\bar{\boldsymbol{A}})$ with $\bar{\boldsymbol{A}} = \boldsymbol{\Gamma} + \gamma \boldsymbol{I}$ less close to the identity matrix, hence more discrepancy between low- and high-fidelity posterior. The first row of the Figure

below shows the low- (purple) and -high-fidelity (red) data densities for decreasing values of γ from 2 down to 0. The second row depicts the predicted posterior densities via the preconditioned (orange contours) and Equation (2) in green along with MCMC samples (dark circles). The third row juxtaposes the preconditioned posterior estimate to the one obtained via the low-fidelity pretrained NF—i.e., Equation (3). Finally, last row shows the objective values during training for with (orange) and without (green) preconditioning.

We observe that by decreasing γ for 2 to 0, the low-fidelity posterior approximations become worse. As a result, the objective function for the preconditioned approach (orange) at the beginning start from a higher value, indicating more mismatch between low- and high-fidelity posterior density. Finally, our preconditioning method consistently improves upon low-fidelity posterior by training for 5 epochs.



2D toy example with decreasing value of $\gamma = 2, 1, 0$ from first to last column, respectively. First row: Low- and high-fidelity data. Second row: approximated posterior densities via MCMC (dark circles), and objectives in Equations (2) and (6). Third row: overlaid prior density with predicted posterior densities via objectives in Equations (3) and (6). Last row: training objective values during training via Equations (2) and (6).



Seismic image compressed sensing for 4 different low-fidelity images. First column: true seismic images. Second column: low-fidelity observed data. Third and last columns: conditional mean and pointwise STD estimates.

Appendix C. Seismic image compressed sensing—more results

Here we show more examples to verify the pretraining phase obtained via solving the objective in Equation (2). Each row in the figure above corresponds to a different testing image (first columns) used to create a low-fidelity sensing data (second column). The third and last column correspond to the conditional mean and pointwise STD estimate, respectively. Clearly, the pretrained network has successfully recovered the original image, and consistently indicates more uncertainty in areas with low-amplitude events.