Multifidelity conditional normalizing flows for physics-guided Bayesian inference

Ali Siahkoohi ¹  Rafael Orozco ¹  Gabrio Rizzuti ²  Philipp A. Witte ³
Mathias Louboutin ¹  Felix J. Herrmann ¹

¹Georgia Institute of Technology
²Utrecht University
³Microsoft
Problem setup

Find $x \in X$ such that

$$y = F(x) + \epsilon$$

- expensive (nonlinear) forward operator $F$
- unknown quantity $x$
- observed data $y \in Y$
- (non-Gaussian) noise $\epsilon$

Bayesian inference

$$p_{post}(x \mid y) \propto p_{like}(y \mid x) \cdot p_{prior}(x)$$

Conditional mean estimate

Pointwise standard deviation, normalized by the envelope of conditional mean

Uncertainty in horizon tracking due to uncertainties in imaging


Xinming Wu and Sergey Fomel. xinwucwp/mhe. 2018. URL: https://github.com/xinwucwp/mhe.

Markov chain Monte Carlo (MCMC) sampling

\[ x_{k+1} = x_k + \frac{\alpha_k}{2} \nabla_x \left[ \log p_{\text{like}}(y | x_k) + \log p_{\text{prior}}(x_k) \right] + \xi_k, \quad \xi_k \sim N(0, \alpha_k I) \]

- asymptotically exact samples
- high-dimensional integration/sampling
- costs of forward operator
- requires choosing a prior distribution


Prior art: MCMC


Variational Inference

Approximate target density $p_{\text{post}}(x \mid y)$ via parametric density $p_\theta(x \mid y)$

$$\forall x, y : \quad p_\theta(x \mid y) \approx p_{\text{post}}(x \mid y)$$

- admits stochastic optimization
- cheap sampling after optimization
- known to scale better than MCMC


Posterior inference with Normalizing Flows (NFs)

sampling the posterior distribution directly via NFs
Purely data-driven approach

\[
\min_{\theta} \mathbb{E}_{y,x \sim p(y,x)} \left[ \frac{1}{2} \| G_\theta(y, x) \|^2 - \log \left| \det \nabla_{y,x} G_\theta(y, x) \right| \right]
\]

\[
G_\theta(y, x) = \begin{bmatrix} G_{\theta_y}(y) \\ G_{\theta_x}(y, x) \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_y \\ \theta_x \end{bmatrix}
\]

conditional NF, \( G_\theta : \mathcal{Y} \times \mathcal{X} \rightarrow \mathcal{Z}_y \times \mathcal{Z}_x \)

expectation estimated via training pairs, \( \{ y_i, x_i \}_{i=1}^n \sim p_{y,x}(y, x) \)

---


Inference with the data-driven approach

Sampling from $p(x \mid y)$

$$G_{\theta_x}^{-1}(G_{\theta_y}(y), z) \sim p(x \mid y), \quad z \sim N(0, I)$$

Posterior density estimation

$$p_{\theta}(x \mid y) = p_z(G_{\theta_x}(y, x)) \left| \det \nabla_x G_{\theta_x}(y, x) \right|$$
Training dataset for the data-driven approach

training pairs

\[
\{y_i, x_i\}_{i=1}^{n} \sim p(y, x)
\]

where

\[
(y_i, x_i) = (J^\top (J \delta m_i + \epsilon_i), \delta m_i)
\]
Low-fidelity reverse-time migrated image, $\mathbf{y}$

Depth (km)
0.0 0.5 1.0 1.5 2.0 2.5 3.0

Horizontal distance (km)
0 1 2 3 4 5

$\times 10^6$

-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
Samples from the posterior

\[ G_{\theta_x}^{-1}(G_{\theta_y}(y), z) \sim p(x \mid y), \quad z \sim \mathcal{N}(0, I) \]
Conditional mean estimate

$$
E[x \mid y] = \int x p(x \mid y) \, dx
$$
Conditional mean, $\delta m_{CM}$
High-fidelity image, $\mathbf{x}$

Depth (km)

Horizontal distance (km)

$\times 10^3$
Normalized pointwise standard deviation, $\sigma/\mu$
Purely data-driven approach

- learns the prior distribution from training data
- samples from the posterior virtually for free in test time
- not specific to one observation $y$
- needs supervised pairs of model and data
- heavily relies on the training data to generalize
- not directly tied to physics/data
Purely physics-based approach

\[
\min_{\phi} \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)} \left[ \frac{1}{2\sigma^2} \left\| F(T_{\phi}(\mathbf{z})) - \mathbf{y} \right\|_2^2 - \log p(T_{\phi}(\mathbf{z})) - \log |\det \nabla_{\mathbf{z}} T_{\phi}(\mathbf{z})| \right]
\]

NF, \( T_{\phi} : \mathcal{Z} \rightarrow \mathcal{X} \)

prior distribution, \( p(\mathbf{x}) \)

specific to one observation, \( \mathbf{y} \)

---


Inference with the physics-based approach

sampling from $p(x | y)$

$$T_\phi(z) \sim p(x | y), \quad z \sim N(0, I),$$

posterior density estimation,

$$p_\phi(x | y) = p_z(T_\phi^{-1}(x)) \left| \det \nabla_x T_\phi^{-1}(x) \right|$$
Purely physics-based approach

tied to the physics and data

no training data needed

requires choosing a prior distribution $p(x)$

repeated evaluations of $F$ and $\nabla F^T$

specific to one observation, $y$
Multi-fidelity preconditioned scheme

exploit the information in the pretrained conditional NF (data-driven approach)

tie the results to physics and data
Multi-fidelity preconditioned scheme

Use the density encoded by the pretrained conditional NF as a prior

\[ p_{\text{prior}}(\mathbf{x}) := p_z(G_{\theta_x}(\mathbf{y}, \mathbf{x})) \left| \det \nabla_x G_{\theta_x}(\mathbf{y}, \mathbf{x}) \right| \]

allows for using a data-driven prior

removes the bias introduced by hand-crafted priors

can be used as a prior density in inverse problems
Multi-fidelity preconditioned scheme

Use transfer learning and initialize $T : \mathcal{Z} \rightarrow \mathcal{X}$ by

$$T_{\theta_x}(z) := G_{\theta_x}^{-1}(G_{\theta_y}(y), z)$$

can significantly reduce ($5\times$) the cost of the purely physics-based approach

can be used as an implicit prior by reparameterizing the unknown with $T$
Preconditioned MAP estimation

\[
\min_z \frac{1}{2\sigma^2} \left\| F(T_{\theta_x}(z)) - y \right\|^2_2 + \frac{1}{2} \left\| z \right\|^2_2
\]

initializing \( z = 0 \) acts as an implicit prior

since \( T \) is invertible, in principle it can represent any unknown \( x \in \mathcal{X} \)

can be used if \( y \) is out-of-distribution

final estimate is tied to the physics/data
Signal-to-noise ratio comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTM</td>
<td>-54.83</td>
</tr>
<tr>
<td>MLE</td>
<td>7.40</td>
</tr>
<tr>
<td>Preconditioned MAP</td>
<td>11.29</td>
</tr>
<tr>
<td>Data-driven CM</td>
<td>4.34</td>
</tr>
</tbody>
</table>

SNR comparison
Preconditioned physics-based approach

\[
\min_{\theta_x} \mathbb{E}_{z \sim N(0, I)} \left[ \frac{1}{2\sigma^2} \left\| F(T_{\theta_x}(z)) - y \right\|^2_2 - \log p_{prior}(T_{\theta_x}(z)) - \log \left| \det \nabla z T_{\theta_x}(z) \right| \right]
\]

transfer learning \(T_{\theta_x}(z) := G_{\theta_x}^{-1}(G_{\theta_y}(y), z)\)

- corrects for errors due to out-of-distribution data
- less evaluations of \(F\) and \(\nabla F^\top\)

learned prior density, \(p_{prior}(x) := p_z(G_{\theta_x}(y, x)) \left| \det \nabla_x G_{\theta_x}(y, x) \right|\)

fast to adapt to new observation

---

Preconditioned conditional mean

Depth (km)

Horizontal distance (km)
# Signal-to-noise ratio comparison

<table>
<thead>
<tr>
<th></th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data-driven CM</td>
<td>4.34</td>
</tr>
<tr>
<td>Preconditioned CM</td>
<td>4.76</td>
</tr>
</tbody>
</table>

SNR comparison
## Number of PDE solves

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of PDE solves</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
</tr>
<tr>
<td>RTM</td>
<td>102</td>
</tr>
<tr>
<td>MLE</td>
<td>204</td>
</tr>
<tr>
<td>Preconditioned MAP</td>
<td>510</td>
</tr>
<tr>
<td>Preconditioned VI approach</td>
<td>2,000</td>
</tr>
<tr>
<td>MCMC w/ deep prior</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Comparison of number of PDE solves in the seismic imaging example
Related work

Purely physics-based approach NF-based Bayesian inference

show orders of magnitude speed up compared to traditional MCMC methods

need to train the NF from scratch for a new observation $y$

use handcrafted priors, often negatively bias the inversion


Related work

Data-driven approaches for directly sampling from the posterior distribution

- fast Bayesian inference given new observation $y$
- sample the posterior virtually for free
- not directly tied to the physics/data
- not reliable when applied to out-of-distribution data
- not trivial to scale GAN-based approaches to large-scale problems

---


Related work

Injective network for inverse problems and uncertainty quantification

use an injective map to map data to a low-dimensional space

can be used as a prior (projection operator) in inverse problems

heavily relies on availability of training data to generalize

unclear how the learned projection operator behaves when applied to out-of-distribution data

for new observation $y$, Bayesian inference requires training an additional NF, involving the forward operator

Contributions

Take full advantage of existing training data to provide

- low-fidelity but fast conditional mean estimate
- a first assessment of the image’s reliability
- preconditioned physics-based high-fidelity MAP estimate via the learned prior
- preconditioned, scalable, and physics-based Bayesian inference
Conclusions

Obtaining UQ information is rendered impractical when

▶ the forward operators are expensive to evaluate
▶ the problem is high dimensional

There are strong indications that

▶ Bayesian inference with normalizing flows can lead to orders of magnitude computational improvements compared to MCMC methods
▶ preconditioning with a pretrained conditional normalizing flow can lead to another order of magnitude speed up

Future work

▶ characterize uncertainties due to modeling errors
Code

https://github.com.slimgroup/InvertibleNetworks.jl

https://github.com.slimgroup/FastApproximateInference.jl

https://github.com.slimgroup/Software.SEG2021
Acknowledgment

This research was carried out with the support of Georgia Research Alliance and partners of the ML4Seismic Center
Variational Inference

Approximate distribution $p$ by minimizing the Kullback-Leibner (KL) divergence

$$\mathbb{D}_{KL}(p \parallel p_\theta) = \int p(x) \log \frac{p(x)}{p_\theta(x)} \, dx$$

$$= \mathbb{E}_{x \sim p(x)} \left[ - \log p_\theta(x) + \log p(x) \right],$$

parametric distribution $p_\theta$

---

Variational Inference

Objective of variational inference is to solve

\[ \theta^* = \operatorname{arg\,min}_{\theta} \mathbb{D}_{KL}(p \parallel p_{\theta}) \]

\(p_{\theta}\) designed to be easily sampled
Generative models

probabilistic models to characterize an unknown distribution
Generative models

\[ p_\theta(x) \approx p(x), \quad x \in \mathcal{X} \]

high dimensional unknown density, \( p(x) \)

generative model, \( p_\theta(x) \), parameterized by \( \theta \)

\( \theta \) estimated via available samples, \( \{x^{(i)}\}_{i=1}^N \sim p(x) \)
Variational Auto Encoders (VAEs)

\[
\arg\max_{\theta, \phi} \prod_{i=1}^{N} p_\theta(x^{(i)}) = \arg\max_{\theta, \phi} \prod_{i=1}^{N} \text{ELBO}_{p,q}(x^{(i)}) \\
= \arg\max_{\theta, \phi} \prod_{i=1}^{N} \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right],
\]

\[
p_\theta(x \mid z) = \mathcal{N}(x \mid \mu_\theta(z), \sigma_\theta^2(z)), \quad p_z(z) = \mathcal{N}(z \mid 0, I)
\]

variational approximation with a encoder network, \( q_\phi(z \mid x) \approx p_\theta(z \mid x) \)

VAEs—pros vs cons

- fast joint and conditional sampling

- intractable density evaluation, $p_\theta(x) = \sum_z p_\theta(x, z)$

- high-variance training gradients due to approximating $\text{ELBO}_{p,q}(x)$
Generative adversarial networks (GANs)

\[
\min_{\theta} \max_{\phi} \mathbb{E}_{x \sim p(x)} \left[ \log D_\phi(x) \right] + \mathbb{E}_{z \sim p_z(z)} \left[ \log (1 - D_\phi(G_\theta(z))) \right]
\]

\[z \sim N(0, I), \quad G_\theta(z) \sim p(x)\]

discriminator network, \(D_\phi : \mathcal{X} \rightarrow [0, 1]\)

\(D_\phi(x)\) classifies input based on its distribution—i.e. \(x \sim p(x)\) or \(x \sim p_\theta(x)\)

GANs—pros vs cons

- fast joint and conditional sampling
- do not provide a density model
- often unstable training due to min-max optimization
Normalizing flows (NFs)

Invertible neural network $G_\theta : \mathcal{X} \rightarrow \mathcal{Z}$, $\mathcal{X}, \mathcal{Z} \in \mathbb{R}^d$

- flexible function for variational inference
- Gaussian latent space $p_z(z) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2} \|z\|_2^2}$
- closed-from and exact inverse (up to numerical precision)
- memory-efficient training due to invertibility
- closed-form expression for $p_\theta(x)$

$$p_\theta(x) = p_z(G_\theta(x)) \det \nabla_x G_\theta(x) \approx p_X(x)$$

Training NFs

\[
\text{arg min}_{\theta} \mathbb{D}_{\text{KL}} (p_X \mid \mid p_\theta) = \text{arg min}_{\theta} \mathbb{E}_{x \sim p_X(x)} \left[ -\log p_\theta(x) + \log p_X(x) \right]
\]

\[
\approx \text{arg min}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{2} \left\| G_\theta(x_i) \right\|_2^2 - \log \left| \det \nabla_x G_\theta(x_i) \right| \right]
\]

- Kullback-Leibner (KL) divergence \( \mathbb{D}_{\text{KL}} \)
- training samples \( \{x^{(i)}\}_{i=1}^{n} \sim p_X(x) \)
- \( \det \nabla_x G_\theta(x) \) comes for free with affine coupling layers

---


Inference with NFs

Sampling from $p(x)$,

$$G_{\theta^*}^{-1}(z) \sim p(x), \quad z \sim p_z(z),$$

and fast density estimation,

$$p(x) = p_\theta(G_\theta(x)) \left| \det \nabla_x G_\theta(x) \right| \approx p_z(G_\theta(x)) \left| \det \nabla_x G_\theta(x) \right|$$
NFs—pros vs cons

- fast joint and conditional sampling
- exact and tractable density (likelihood) estimation
- allows for memory efficient gradient computation
- often have less representation power
- unable to alter dimensionality


Affine coupling layer

a basic "layer" with an analytic inverse
an invertible layer with input $u$

- arbitrary neural networks $s_\ell$, $t_\ell$

- output of $\ell$th layer $T_\ell(u)$

- orthogonal Householder reflections $Q_\ell$

---


Affine coupling layer—forward

\[
T_\ell(u) = \begin{bmatrix}
T^1_\ell(\tilde{u}_1) \\
T^2_\ell(\tilde{u}_1, \tilde{u}_2)
\end{bmatrix} = \begin{bmatrix}
\tilde{u}_1 \\
\tilde{u}_2 \circ \sigma(s_\ell(\tilde{u}_1)) + t_\ell(\tilde{u}_1)
\end{bmatrix},
\]

with \( \tilde{u} = \begin{bmatrix}
\tilde{u}_1 \\
\tilde{u}_2
\end{bmatrix} = Q_\ell u \)

sigmoid function \( \sigma(\cdot) \)

elementwise multiplication \( \circ \)

Affine coupling layer—inverse

\[
\tilde{u} = \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \end{bmatrix} = \begin{bmatrix} T_1^1(\tilde{u}_1) \\ \left( T_2^2(\tilde{u}_1, \tilde{u}_2) - t( T_1^1(\tilde{u}_1)) \right) \otimes \sigma \left( s(T_1^1(\tilde{u}_1)) \right) \end{bmatrix},
\]

\[
u = Q_\ell^{-1} \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \end{bmatrix}
\]

orthogonal learnable matrix \( Q_\ell^{-1} = Q_\ell^\top \)

elementwise division \( \otimes \)

Determinant of the Jacobian

\[
\nabla_u \mathbf{T}_\ell(u) = \nabla_u \begin{bmatrix} T^1_\ell(\tilde{u}_1) \\ T^2_\ell(\tilde{u}_1, \tilde{u}_2) \end{bmatrix} = \begin{bmatrix} I \\ \nabla_u T^2_\ell(\tilde{u}_1, \tilde{u}_2) \end{bmatrix} \begin{bmatrix} 0 \\ \text{diag}(\sigma(s_\ell(\tilde{u}_1))) \end{bmatrix}
\]

\[
\Rightarrow \log |\det \nabla_u \mathbf{T}_\ell(u)| = \text{sum}(\log \sigma(s_\ell(\tilde{u}_1)))
\]

sparsity pattern of Jacobian

Challenges of affine coupling layers

lower representation power due

- sparse triangular Jacobian

requires many layer compositions and orthogonal transforms

- to capture all pixel-wise dependencies

may become numerically non-invertible

- e.g. due to very small Jacobian singular values

---


Hierarchical coupling layers

dense triangular Jacobian

hierarchical architecture

▶ recursive affine coupling layers
▶ until input can not be split any further

Posterior inference with NFs

Train a NF to directly characterize the posterior density

\[ p_\theta(x) \approx p_{\text{post}}(x | y) \propto p_{\text{like}}(y | x) \, p_{\text{prior}}(x), \]

i.e.,

being able to sample from the posterior

estimate the posterior density
Training a NF for posterior inference

\[
\arg \min_{\theta} \mathbb{D}_{KL} \left( p_\theta \mid\mid p_{\text{post}}(\cdot \mid y) \right) \\
= \arg \min_{\theta} \mathbb{E}_{x \sim p_\theta(x \mid y)} \left[ -\log p_{\text{post}}(x \mid y) + \log p_\theta(x) \right].
\]

KL divergence, $\mathbb{D}_{KL}$, a distance between two distributions

distribution encoded by the NF, $p_\theta(x) = p_z(z) \mid \det \nabla_x T_\theta^{-1}(x)$, $z = T_\theta^{-1}(x)$
Posterior inference with NFs

purely physics-based approach

▶ tied to the physics
▶ no training data needed
▶ requires handcrafted priors

purely data-driven approach

▶ needs supervised pairs of model and data
▶ heavily relies on the training data to generalize
▶ ignores the physics
Physics-based variational inference

\[
\arg\min_{\theta} \mathbb{D}_{\text{KL}} (p_{\theta} \mid \mid p_{\text{post}}(\cdot \mid y)) \\
= \arg\min_{\theta} \mathbb{E}_{x \sim p_{\theta}(x)} \left[ -\log p_{\text{post}}(x \mid y) + \log p_{\theta}(x) \right] \\
= \arg\min_{\theta} \mathbb{E}_{z \sim p_{z}(z)} \left[ -\log p_{\text{post}}(T_{\theta}(z) \mid y) + \log p_{\theta}(T_{\theta}(z)) \right] \\
= \arg\min_{\theta} \mathbb{E}_{z \sim p_{z}(z)} \left[ -\log p_{\text{post}}(T_{\theta}(z) \mid y) + \log p_{z}(z) - \log \left| \det \nabla_{z} T_{\theta}(z) \right| \right] \\
= \arg\min_{\theta} \mathbb{E}_{z \sim p_{z}(z)} \left[ -\log p_{\text{post}}(T_{\theta}(z) \mid y) - \log \left| \det \nabla_{z} T_{\theta}(z) \right| \right].
\]

Physics-based variational inference

$$\min_{\theta} \mathbb{E}_{z \sim p_z(z)} \left[ \frac{1}{2\sigma^2} \left\| F(T_{\theta}(z)) - y \right\|^2 - \log p_{\text{prior}}(T_{\theta}(z)) - \log \left| \det \nabla_z T_{\theta}(z) \right| \right]$$

no training data required

cheap posterior sampling: $T_{\theta}(z) \sim p_{\theta}(x \mid y) \approx p_{\text{post}}(x \mid y)$

requires a prior density, $p_{\text{prior}}(x)$

repeated evaluations of $F$ and $\nabla F^\top$

specific to one observation, $y$
Data-driven variational inference

achieved via a block-triangular transport map, $G_\theta(y,x)$

training over model and data pairs, $(y,x) \sim \hat{p}_{y,x}(y,x)$

capable of sampling the posterior for all $y \sim p_y(y)$

provides an estimate to the posterior density
Data-driven variational inference

\[
\arg \max_{\phi} \mathbb{E}_{y,x \sim \hat{p}_{y,x}} [\log p_{\theta}(y, x)] = \arg \min_{\phi} \mathbb{E}_{y,x \sim \hat{p}_{y,x}} \left[ \frac{1}{2} \| G_{\theta}(y, x) \|^2 - \log \left| \det \nabla_{y,x} G_{\theta}(y, x) \right| \right] \\
\approx \arg \min_{\phi} \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{2} \| G_{\theta}(y_i, x_i) \|^2 - \log \left| \det \nabla_{y,x} G_{\theta}(y_i, x_i) \right| \right]
\]

NF, \( G_{\theta} : \mathcal{Y} \times \mathcal{X} \to \mathcal{Z}_y \times \mathcal{Z}_x \)

maximum likelihood estimate for \( \theta \) given data pairs, \( \{y_i, x_i\}_{i=1}^{n} \sim \hat{p}_{y,x}(y, x) \)
Block-triangular map

\[ G_\theta(y, x) = \begin{bmatrix} G_{\theta y}(y) \\ G_{\theta x}(y, x) \end{bmatrix}, \ \phi = \begin{bmatrix} \phi_y \\ \phi_x \end{bmatrix} \]

conditional sampling: \( G_{\theta x}^{-1}(G_{\theta y}(y), z) \sim p_{\text{post}}(x \mid y) \) for \( z \sim N(0, I) \)

posterior density estimation: \( p_G(x \mid y) = p_z(G_{\theta x}(y, x)) \left| \det \nabla_x G_{\theta x}(y, x) \right| \).

---


Posterior inference with conditional NFs

\[ z_y = G_{\theta_y}(y) \]

Posterior inference with conditional NFs

\[ G_{\theta_x}^{-1}(G_{\theta_y}(y), z_x) \sim p_{\text{post}}(x \mid y), \quad z_x \sim N(0, I) \]
Data-driven variational inference

$$\min_{\phi} \mathbb{E}_{y,x \sim p_{y,x}(y,x)} \left[ \frac{1}{2} \| G_{\theta}(y, x) \|^2 - \log |\det \nabla_{y,x} G_{\theta}(y, x)| \right]$$

cheap posterior sampling for all $y$: $G_{\theta,x}^{-1}(G_{\theta,y}(y), z) \sim p_{\text{post}}(x \mid y)$ for $z \sim N(0, I)$

does not involve the expensive forward operator, $F$

does not require a prior density, $p_{\text{prior}}(x)$

heavily relies on access to training pairs, $\hat{p}_{y,x} \neq p_{y,x}$

---

Multi-fidelity preconditioned scheme

initialize $T_{\phi_x}(z) := G_{\theta_x}^{-1}(G_{\theta_y}(y), z)$, followed by transfer learning

use $p_G(x) =$ density encoded by $G_{\theta_x}^{-1}$—as a (conditional) prior
Multifidelity preconditioned formulations

- insert the conditional NF in physics-based formulations and use transfer learning
  \[
  T_{\theta_x}(z) := G_{\theta_x}^{-1}(G_{\theta_y}(y), z)
  \]

- preconditioned maximum a posteriori (MAP) estimation
  \[
  \min_z \frac{1}{2\sigma^2} \| F(T_{\theta_x}(z)) - y \|^2_2 + \frac{1}{2} \| z \|^2_2
  \]

- preconditioned physics-based posterior learning
  \[
  \min_{\theta_x} \mathbb{E}_{z \sim N(0, I)} \left[ \frac{1}{2\sigma^2} \| F(T_{\theta_x}(z)) - y \|^2_2 - \log p_{\text{prior}}(T_{\theta_x}(z)) - \log \left| \det \nabla_z T_{\theta_x}(z) \right| \right]
  \]

---

2D Rosenbrock toy example
(left) prior, (right) low- and high-fidelity data

(left) without vs with preconditioning vs MCMC, (right) loss with and without preconditioning

Seismic compressed sensing example
(left) true model, (right) “observed” data
conditional mean (left) and pointwise standard deviation (right)
(left) true model, (right) “observed” data
conditional mean (left) and pointwise standard deviation (right)
(left) true model, (right) observed data
conditional mean before (left) and after (right) transfer learning
Seismic imaging examples
Seismic imaging forward model

\[ \delta d = J[m_0, q] \delta m + \eta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \]

linearized observed data \( \delta d \)
source signature \( q \)
smooth background model \( m_0 \)
unknown perturbation model \( \delta m \)
linearization error and noise \( \eta, \epsilon \)
Linearized Born modeling operator

\[ J[m_0, q] = \nabla_m PA^{-1}[m]q \bigg|_{m=m_0} \]

\[ = -PA^{-1}[m_0] \text{diag} \left( \nabla A[(m_0)A^{-1}[m_0]q] \right) \]

- discretized wave equation \( A[m] \)
- restriction operator \( P \)
Training dataset for the data-driven approach

Creating perturbation models $\delta m$

3075 m × 5120 m sections from the Parihaka dataset

selected from shallow sections

mainly consists of strong horizontal reflectors

artificial 125 m water layer to limit the near source imaging artifacts

Training dataset for the data-driven approach

For each reflectivity model $\delta m$ define

$$\mathbf{x} := \delta m$$

and

$$\mathbf{y} := \mathbf{J}^\top (\mathbf{J}\delta m + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$
Training dataset for the data-driven approach

Training pairs:

\[ \{y_i, x_i\}_{i=1}^n \sim p(y, x) \]

where

\[ (y_i, x_i) = (J^\top (J\delta m_i + \epsilon_i), \delta m_i) \]
Some data pairs from the training set
Low-fidelity reverse-time migrated image, $y$
Low-fidelity reverse-time migrated image, $y$
Low-fidelity reverse-time migrated image, $y$
Results