Refining Amortized Posterior Approximations using Gradient-Based Summary Statistics

Rafael Orozco ROROZCO@GATECH.EDU
Georgia Institute of Technology

Ali Siahkoohi ALISK@RICE.EDU
Rice University

Mathias Louboutin MLOUBOUTIN3@GATECH.EDU
Georgia Institute of Technology

Felix J. Herrmann FELIX.HERRMANN@GATECH.EDU
Georgia Institute of Technology

Abstract

We present an iterative framework to improve the amortized approximations of posterior distributions in the context of Bayesian inverse problems, which is inspired by loop-unrolled gradient descent methods and is theoretically grounded in maximally informative summary statistics. Amortized variational inference is restricted by the expressive power of the chosen variational distribution and the availability of training data in the form of joint data and parameter samples, which often lead to approximation errors such as the *amortization gap*. To address this issue, we propose an iterative framework that refines the current amortized posterior approximation at each step. Our approach involves alternating between two steps: (1) constructing a training dataset consisting of pairs of summarized data residuals and parameters, where the summarized data residual is generated using a gradient-based summary statistic, and (2) training a conditional generative model—a normalizing flow in our examples—on this dataset to obtain a probabilistic update of the unknown parameter. This procedure leads to iterative refinement of the amortized posterior approximations without the need for extra training data. We validate our method in a controlled setting by applying it to a stylized problem, and observe improved posterior approximations with each iteration. Additionally, we showcase the capability of our method in tackling realistically sized problems by applying it to transcranial ultrasound, a high-dimensional, nonlinear inverse problem governed by wave physics, and observe enhanced posterior quality through better image reconstruction with the posterior mean.[1]

1. Introduction

We aim to solve Bayesian inverse problems where indirect observations $\mathbf{y}$ of parameters $\mathbf{x}$ are given by the forward operator $\mathcal{F}$ and noise $\boldsymbol{\varepsilon}$:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \varepsilon. \tag{1}$$

Our goal is to estimate the posterior distribution $p(\mathbf{x} \mid \mathbf{y})$ of parameters $\mathbf{x}$ given observation $\mathbf{y}$. Bayes' theorem relates the posterior to the likelihood function and prior distribution as: $p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ where the likelihood $p(\mathbf{y}|\mathbf{x})$ depends on the noise distribution $p(\boldsymbol{\varepsilon})$ and the forward operator $\mathcal{F}$. Challenges in this problem include: (1) high-dimensionality of unknown parameters, which necessitates the computation of numerous posterior samples to calculate statistics; (2) a non-linear forward operator $\mathcal{F}$, resulting in multi-modal posteriors; and (3) obtaining analytical expressions for non-Gaussian priors that capture our prior knowledge about the parameters. To

1. Code to reproduce experiments included as anonymized source code. Github repository to be released after the review process.

overcome these challenges, recent work has exploited deep neural networks by learning the high-dimensional and non-linear behaviour of these inverse problems and by approximating unknown prior distributions from examples. These methods typically fall under the umbrella of variational inference (VI) since the computational complexity of posterior sampling is exchanged for neural network optimization Jordan et al. (1999). There are two VI approaches: amortized VI and non-amortized VI Zhang et al. (2023).

Amortized VI can leverage deep generative models to approximate posterior distributions, offering wide applicability and online inference efficiency Radev et al. (2020). By online efficiency, we mean that amortized VI methods entail a computationally expensive pre-training phase (offline cost) but its cost at inference (online cost) is low because after training they can cheaply compute posterior inference results for many different observations $\mathbf{y}$ drawn from the same distribution as training data. In contrast, non-amortized VI does not have an offline cost thus all the computational cost is paid online at inference time by solving an optimization problem related to the potentially expensive operator $\mathcal{F}$ and its gradient. Non-amortized inference results can not be reused so the optimization will need to be repeated for every new observation. Since non-amortized VI is specialized to a single observation $\mathbf{y}$ its performance is typically better than its amortized counterpart since amortized VI averages performance over many observations; a phenomena denoted as the amortization gap Marino et al. (2018). Our work aims to bridge the gap with principled additional offline computation, preserving the fast online inference properties of amortized VI while achieving performance closer to non-amortized VI approaches.

2. Related work

Our method is related to the two-step process in Siahkoohi et al. (2021) starting with amortized VI then switching to non-amortized VI taking advantage of the implicitly learned prior. Siahkoohi et al. (2022) also performs an amortized then non-amortized step with the goal of correcting for distribution changes from the training data. Bayesian results in ultrasound imaging were previously shown Bates et al. (2022) but used a mean field approximation constraining it to point wise variances. The method we propose is not limited to Gaussian priors and recovers the full posterior covariance Figure 1(*b*). Deep-GEM Gao et al. (2021) is similar to our work since it iterates between normalizing flow training and gradient-based optimization. In contrast with Deep-GEM, our method trains an amortized normalizing flow, and at each iteration only one gradient step is needed; leading to an amortized procedure with efficient online inference.

3. Methods

3.1. Amortized variational inference

Reducing the costs of Bayesian inference in inverse problems with computationally costly forward operators can be achieved via amortized VI Radev et al. (2020). This is implemented by learning a parametric conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{y})$ (here we will use normalizing flows Dinh et al. (2016)) that approximates the posterior for a distribution of observations. The approximate posterior quality is measured by the Kullbeck-Leibler (KL) divergence calculated in expectation (amortized) over the distribution of observations $p(\mathbf{y})$. Amortized VI minimizes this expected KL divergence with respect to θ

$$\begin{aligned} \min_\theta \mathbb{E}_{p(\mathbf{y})}\Big[\mathbb{KL}\left(p(\mathbf{x} \mid \mathbf{y}) \mid\mid p_\theta(\mathbf{x} \mid \mathbf{y})\right)\Big] &= \min_\theta \mathbb{E}_{p(\mathbf{y})}\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\Big[-\log p_\theta(\mathbf{x} \mid \mathbf{y}) + p(\mathbf{x} \mid \mathbf{y})\Big] \\ = \min_\theta \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\Big[-\log p_\theta(\mathbf{x} \mid \mathbf{y}) + p(\mathbf{x} \mid \mathbf{y})\Big] \end{aligned} \tag{2}$$

where we have rewritten the terms to avoid an expectation over the posterior distribution in favor of an expectation over the joint distribution $p(\mathbf{x}, \mathbf{y})$. In practice, a dataset of N training examples

$\{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ from the joint distribution is used to approximate the expectation. For conditional normalizing flows (CNF), the above optimization problem simplifies to:

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \left(\|f_\theta(\mathbf{x}^{(n)}; \mathbf{y}^{(n)})\|_2^2 - \log |\det \mathbf{J}_{f_\theta}| \right), \tag{3}$$

where f_θ is a CNF implemented as in Ardizzone et al. (2019) and $\mathbf{J}_{f_\theta}$ is the Jacobian of the CNF. The success of the CNF in approximating the posterior depends on the amount of training data used to estimate the expectation in Equation (2). For high-dimensional inverse problems, limited training data leads to poor expectation estimates and therefore poor posterior approximations. To improve the accuracy of this posterior approximation, we will discuss the role of summary statistics in high-dimensional inference and how they ultimately lead to a iterative formulation for improving the posterior approximation without additional training data.

3.2. Summary statistics for high-dimensional inverse problems

Efficient inference procedures are difficult to implement when observations $\mathbf{y}$ are high-dimensional Cranmer et al. (2020). Thus efficient inference procedures typically use a summary statistic to summarize large observations. Alsing and Wandelt (2018) demonstrated that $\mathbf{y}$ can be summarized by the score of the log-likelihood at a fiducial $\mathbf{x}_0$. In this context, the fiducial $\mathbf{x}_0$ refers to a reliable estimate of the unknown $\mathbf{x}$. Furthermore, it has been proven that this score is maximally informative Deans (2002), as determined by the Fisher information it contains needed to infer $\mathbf{x}$. The score summary statistic is defined as

$$\bar{\mathbf{y}} := \nabla_{\mathbf{x}_0} \log p(\mathbf{y}|\mathbf{x}_0). \tag{4}$$

The score $\bar{\mathbf{y}}$ is maximally informative specifically around the fiducial $\mathbf{x}_0$ so approximate posterior learning with a score summary statistic has a modified implementation. First we change the target to be the update $\mathbf{\Delta x} = \mathbf{x} - \mathbf{x}_0$ and using Equation (3) we learn the conditional distribution $p(\Delta\mathbf{x}|\bar{\mathbf{y}})$, i.e.,

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \left(\|f_\theta(\mathbf{\Delta x}^{(n)}; \bar{\mathbf{y}}^{(n)})\|_2^2 - \log |\det \mathbf{J}_{f_\theta}| \right). \tag{5}$$

Then the approximate posterior is sampled by adding the learned conditional distribution to the fiducial, i.e., $p(\mathbf{x}|\mathbf{y}) \approx \mathbf{x}_0 + p_{\hat{\theta}}(\mathbf{\Delta x} \mid \bar{\mathbf{y}})$, where the conditional samples from $p_{\hat{\theta}}(\mathbf{\Delta x} \mid \bar{\mathbf{y}})$ are generated by the CNF with optimized weights $\hat{\theta}$. Using the score as a summary in the context of Bayesian inference has been implemented Adler and Öktem (2018) but because of finite training data and because of the amortization gap the amortized results can be lacking compared with non-amortized results Siahkoohi et al. (2022); Orozco et al. (2023).

3.3. Iterative amortized variational inference

To improve the amortized approximation, we take inspiration from loop-unrolled gradient descent and identify that the network trained in Equation (5) also takes the gradient as input (score is gradient of log-likelihood). Thus we suggest to interpret the fiducial $\mathbf{x}_0$ as the first in a sequence of L updating iterations $\mathbf{x}_0, \mathbf{x}_1 \ldots \mathbf{x}_L$ where the CNF has learned a probabilistic update to the fiducial $p_{\hat{\theta}_0}(\mathbf{\Delta x}_0 \mid \bar{\mathbf{y}}_0)$. We added subscript notation to emphasize that the weights $\hat{\theta}_0$, update $\mathbf{\Delta x}_0$ and score $\bar{\mathbf{y}}_0$ correspond to the 0th fiducial. To update the fiducial, we use a single point estimate from the learned distribution $p_{\hat{\theta}}(\mathbf{\Delta x}_0 \mid \bar{\mathbf{y}}_0)$. In this work, we concentrate on the approximate posterior mean $\mathbb{E}_{p_{\hat{\theta}_0}(\mathbf{\Delta x}_0|\bar{\mathbf{y}}_0)} [\mathbf{\Delta x}_0 \mid \bar{\mathbf{y}}_0]$:

$$\mathbf{x}_1 = \mathbf{x}_0 + \mathbb{E}_{p_{\hat{\theta}_0}(\mathbf{\Delta x}_0|\bar{\mathbf{y}}_0)} [\mathbf{\Delta x}_0 \mid \bar{\mathbf{y}}_0], \tag{6}$$

where the expectation is estimated as in Appendix Equation (7). We chose the posterior mean since it is the estimator with least variance Adler and Öktem (2018). The implications of using different point statistics is left for future work. According to Alsing and Wandelt (2018), the informativeness of the score depends on how close the fiducial $\mathbf{x}_0$ is to the ground truth $\mathbf{x}$. Therefore if the new fiducial $\mathbf{x}_1$ is closer to $\mathbf{x}$ than $\mathbf{x}_0$ then the score calculated at this new point $\bar{\mathbf{y}}_1$ will be a more informative summary statistic and lead to a better posterior approximation. We empirically verify this statement in Section 4.1.

By updating all fiducials $\mathbf{x}_0^{(n)}$ in the training dataset with Equation (6) and computing new scores $\bar{\mathbf{y}}_1^{(n)}$ we create a new dataset of pairs $\{(\boldsymbol{\Delta}\mathbf{x}_1^{(n)}, \bar{\mathbf{y}}_1^{(n)})\}_{n=1}^N$ that can be used to train a second CNF $f_{\hat{\theta}_1}$. In the algorithm we propose, this training and fiducial update procedure is continued alternated for L iterations as outlined in Appendix Algorithm 2.

At inference time, we take the trained networks $f_{\hat{\theta}_0}, f_{\hat{\theta}_1} \ldots, f_{\hat{\theta}_L}$ and the first fiducial $\mathbf{x}_0$ then iterate on calculating scores $\bar{\mathbf{y}}_i$ at the i-th fiducial and updating the fiducial with the posterior mean until iteration $L-1$. The final posterior approximation is the last fiducial added to samples from the last learned conditional distribution. The full inference algorithm is shown in Algorithm 1.

Algorithm 1: Inference phase

Input: Observation $\mathbf{y}$ and starting fiducial $\mathbf{x}_0$

for $i = 0$ **to** $L-1$ **do**

  $\bar{\mathbf{y}}_i = \nabla_{\mathbf{x}_i} \log p(\mathbf{y}|\mathbf{x}_i)$;

  $\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbb{E}_{p_{\hat{\theta}_i}(\boldsymbol{\Delta}\mathbf{x}_i|\bar{\mathbf{y}}_i)}\,[\boldsymbol{\Delta}\mathbf{x}_i \mid \bar{\mathbf{y}}_i]$;

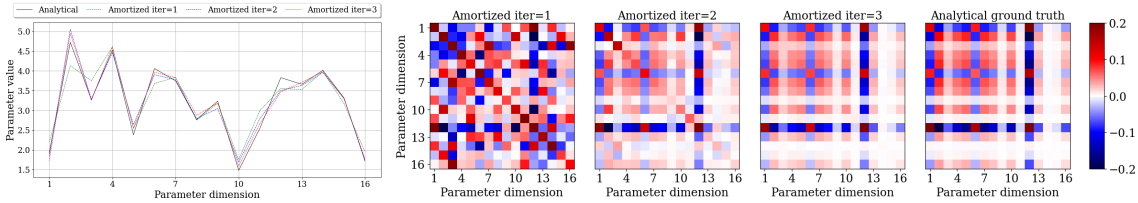
end

Output: Final approximate posterior: $p(\mathbf{x}|\mathbf{y}) \approx p_{\hat{\theta}_L}(\mathbf{x} \mid \mathbf{y}) = \mathbf{x}_L + p_{\hat{\theta}_L}(\boldsymbol{\Delta}\mathbf{x}_L \mid \bar{\mathbf{y}}_L)$

4. Results

4.1. Validation with Gaussian example

Our proposed method is grounded in theoretical concepts but its application to our ultrasound medical imaging inverse problem is difficult to validate due to a non-linear forward operator and an unknown prior. To build trust, we first validate our claims with an inverse problem that has an analytical posterior distribution to compare with. We solve an inverse problem with Gaussian prior on parameters and noise with a linear forward operator written as $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$ where $\mathbf{A} \in \mathbb{R}^{64\times 16}$. Given an observation $\mathbf{y}$ there is an analytical expression for the posterior covariance and the posterior mean $\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\,[\mathbf{x} \mid \mathbf{y}]$. See Hagemann et al. (2021) for derivation details.



(a) Posterior mean (b) Posterior covariance

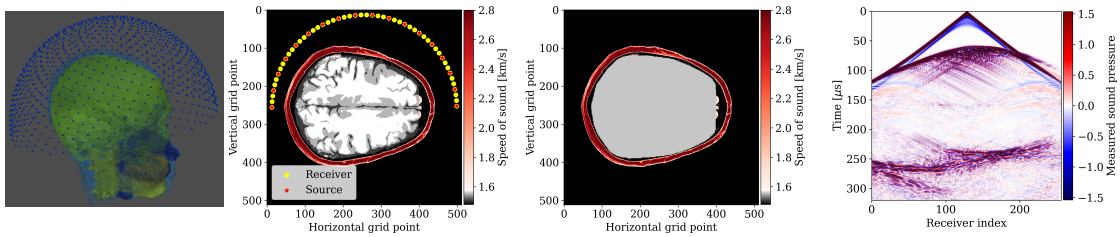
Figure 1: Our method iteratively improves its posterior approximation as compared to the analytical posterior.

We trained our method using $N = 1000$ training examples from the joint distribution and all starting fiducials $\mathbf{x}_0^{(n)}$ were set to zeros. In Figure 1, we observe that after each iteration, our posterior

mean and covariance is a better approximation to the ground truth. In Appendix, Section C, we show that this trend is consistent by showing the average approximation error for a test set.

4.2. High-dimensional ultrasound imaging

We evaluate our iterative method on transcranial ultrasound computed tomography (TUCT). TUCT involves estimating speed of sound of brain tissue from measurements of ultrasound waves impinging on internal brain tissue. For a schematic of the hardware involved in the real-life 3D imaging setup see Figure 2(a). In this work, we consider a 2D slice Figure 2(b) of the 3D setup. A Bayesian framework is desirable for TUCT because the observations are noisy and the forward operator has a null space due to limited-view receivers, in other words: waves can only be transmitted and received from the top of the simulation Figure 2(b). For details on TUCT, refer to Guasch et al. (2020); Marty et al. (2021).



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

Figure 2: Medical imaging inverse problem: (a) 3D imaging hardware; (b) 2D in-silico demo used in this work; (c) Example starting fiducial $\mathbf{x}_0$; (d) Simulated observations $\mathbf{y}$;

The TUCT inverse problem is high-dimensional (512×512 grid here) and the forward operator $\mathcal{F}$ is non-linear and is characterized by the wave equation, which is parameterized by varying speed of sound and density. We use finite-differences software Devito Louboutin et al. (2019) to solve the wave wave equation $\mathcal{F}$ and its Jacobian for scores/gradients. Our method requires samples from the prior $p(\mathbf{x})$ on brain speed of sound parameters that we generate by assigning acoustic values to the FASTMRI dataset Zbontar et al. (2018). We used MRI volumes from 250 subjects to make $N = 2475$ training 2D slices. Due to pathological cycle skipping in wave-based imaging Virieux and Operto (2009) the fiducial starts from the bone known (assumed to be calculated from X-ray as in Aubry et al. (2003)) and constant speed of sound for the internal tissue, refer to Figure 2(c) for an example starting fiducial.

After training our method for $L = 3$ iterations, we plot posterior statistics for an unseen observation $\mathbf{y}$ in Figure 3. Each iteration improves the amortized posterior approximation as quantified by the quality of the posterior mean. Due to the receivers being limited to the top half of the simulation area (see Figure 2(b)) the error and uncertainty concentrates at the bottom half of the parameters. Appendix Table 1 contains average quality metrics over 200 samples from a test set. We summarize by noting that at iteration three, the average PSNR was 43.27dB and on average the second iteration increased PSNR by 1.63dB $\pm$0.54dB and the third iteration increased PSNR by 2.82dB $\pm$0.91dB.

5. Conclusion

We showed the first instance of a gradient-based summary statistic paired with normalizing flows that enables iteratively refining amortized posterior distributions for large dimensional and non-linear Bayesian inverse problems. Our method exchanges additional offline computation to get better amortized approximations of Bayesian posteriors. Using a Gaussian inverse problem, we showed our

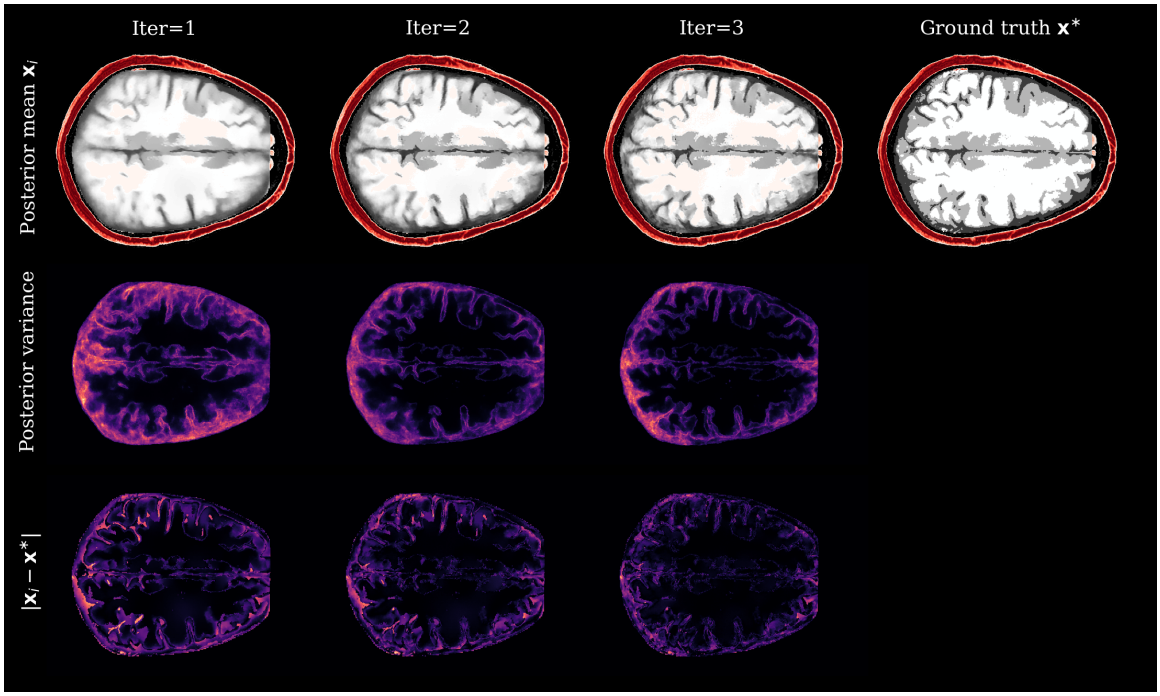


Figure 3: Our iterative method applied to a high-dimensional non-linear medical imaging problem.

method converges to the ground truth posterior. We also demonstrated our method scales to a high-dimensional medical imaging problem related to expensive non-linear PDE operators.

References

Jonas Adler and Ozan Öktem. Deep bayesian inversion. *arXiv preprint arXiv:1811.05910*, 2018.

Justin Alsing and Benjamin Wandelt. Generalized massive optimal data compression. *Monthly Notices of the Royal Astronomical Society: Letters*, 476(1):L60–L64, 2018.

Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Conditional invertible neural networks for guided image generation. 2019.

J-F Aubry, M Tanter, M Pernot, J-L Thomas, and M Fink. Experimental demonstration of noninvasive transskull adaptive focusing based on prior computed tomography scans. *The Journal of the Acoustical Society of America*, 113(1):84–93, 2003.

Oscar Bates, Lluis Guasch, George Strong, Thomas Caradoc Robins, Oscar Calderon-Agudo, Carlos Cueto, Javier Cudeiro, and Mengxing Tang. A probabilistic approach to tomography and adjoint state methods, with an application to full waveform inversion in medical ultrasound. *Inverse Problems*, 38(4):045008, 2022.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Matthew C Deans. Maximally informative statistics for localization and mapping. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, volume 2, pages 1824–1829. IEEE, 2002.

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

Angela Gao, Jorge Castellanos, Yisong Yue, Zachary Ross, and Katherine Bouman. Deepgem: Generalized expectation-maximization for blind inversion. *Advances in Neural Information Processing Systems*, 34:11592–11603, 2021.

Peimeng Guan, Jihui Jin, Justin Romberg, and Mark A Davenport. Loop unrolled shallow equilibrium regularizer (luser)–a memory-efficient inverse problem solver. *arXiv preprint arXiv:2210.04987*, 2022.

Lluís Guasch, Oscar Calderón Agudo, Meng-Xing Tang, Parashkev Nachev, and Michael Warner. Full-waveform inversion imaging of the human brain. *NPJ digital medicine*, 3(1):1–12, 2020.

Paul Hagemann, Johannes Hertrich, and Gabriele Steidl. Generalized normalizing flows via markov chains. *arXiv preprint arXiv:2111.12506*, 2021.

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.

Conlain Kelly and Surya R Kalidindi. Recurrent localization networks applied to the lippmann-schwinger equation. *Computational Materials Science*, 192:110356, 2021.

M. Louboutin, M. Lange, F. Luporini, N. Kukreja, P. A. Witte, F. J. Herrmann, P. Velesko, and G. J. Gorman. Devito (v3.1.0): an embedded domain-specific language for finite differences and geophysical exploration. *Geoscientific Model Development*, 12(3):1165–1187, 2019. doi: 10.5194/gmd-12-1165-2019. URL https://www.geosci-model-dev.net/12/1165/2019/.

Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pages 3403–3412. PMLR, 2018.

Patrick Marty, Christian Boehm, and Andreas Fichtner. Acoustoelastic full-waveform inversion for transcranial ultrasound computed tomography. In *Medical Imaging 2021: Ultrasonic Imaging and Tomography*, volume 11602, pages 210–229. SPIE, 2021.

Rafael Orozco, Mathias Louboutin, Ali Siahkoohi, Gabrio Rizzuti, Tristan van Leeuwen, and Felix Herrmann. Amortized normalizing flows for transcranial ultrasound with uncertainty quantification. *arXiv preprint arXiv:2303.03478*, 2023.

Patrick Putzky and Max Welling. Invert to learn to invert. *Advances in Neural Information Processing Systems*, 32, 2019.

Stefan T Radev, Ulf K Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp A Witte, and Felix J Herrmann. Preconditioned training of normalizing flows for variational inference in inverse problems. *arXiv preprint arXiv:2101.03709*, 2021.

Ali Siahkoohi, Gabrio Rizzuti, Rafael Orozco, and Felix J Herrmann. Reliable amortized variational inference with physics-based latent distribution correction. *arXiv preprint arXiv:2207.11640*, 2022.

Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Xin Zhang, Angus Lomas, Muhong Zhou, York Zheng, and Andrew Curtis. 3-d bayesian variational full waveform inversion. *Geophysical Journal International*, 234(1):546–561, 2023.

Appendix A. Training our method

In this work, we follow traditional loop-unrolling schemes (Putzky and Welling, 2019; Kelly and Kalidindi, 2021) and train separate conditional normalizing flows at each step but we could also reuse the same networks or define implicit layers Guan et al. (2022).

Algorithm 2: Training Phase

Input: Paired of observations $\{\mathbf{y}^{(n)}\}_{n=1}^{N_{train}}$ and corresponding parameters $\{\mathbf{x}^{(i)}\}_{n=1}^{N_{train}}$

for $n \leftarrow 1$ **to** N_{train} **do**
  Generate starting fiducial: $\mathbf{x}_0^{(n)}$;
  $\boldsymbol{\Delta}\mathbf{x}_0^{(n)} = \mathbf{x}^{(n)} - \mathbf{x}_0^{(n)}$;
  Summarize observation with gradient: $\bar{\mathbf{y}}_0^{(n)} = \nabla_{\mathbf{x}} \log p(\mathbf{y}^{(n)} \mid \mathbf{x}_0^{(n)})$;
  Add pairs to dataset: $\mathcal{D}_0^{(n)} = (\boldsymbol{\Delta}\mathbf{x}_0^{(n)}, \bar{\mathbf{y}}_0^{(n)})$;
end
for $j = 0$ **to** L **do**
  while *conditional normalizing flow* f_{θ_j} *is not converged* **do**
    Evaluate objective function on dataset $\mathcal{D}_j$ using Equation (5);
    Update θ_j with backpropagation;
  end
  for $n = 1$ **to** N_{train} **do**
    Update current guess: $\mathbf{x}_{j+1}^{(n)} = \mathbf{x}_j^{(n)} + \mathbb{E}\left[p_{\hat{\theta}_j}(\boldsymbol{\Delta}\mathbf{x} \mid \bar{\mathbf{y}}_j^{(n)})\right]$;
    $\boldsymbol{\Delta}\mathbf{x}_{j+1}^{(n)} = \mathbf{x}^{(n)} - \mathbf{x}_{j+1}^{(n)}$;
    Re-summarize observation with gradient: $\bar{\mathbf{y}}_{j+1}^{(n)} = \nabla_{\mathbf{x}} \log p(\mathbf{y}^{(n)} \mid \mathbf{x}_{j+1}^{(n)})$;
    Add pairs to dataset: $\mathcal{D}_{j+1}^{(n)} = (\boldsymbol{\Delta}\mathbf{x}_{j+1}^{(n)}, \bar{\mathbf{y}}_{j+1}^{(n)})$;
  end
end

Output: $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_L$ Posterior Networks

Appendix B. Sampling from conditional normalizing flow

To estimate the conditional expectation $\mathbb{E}_{p_{\hat{\theta}}(\boldsymbol{\Delta}\mathbf{x}|\bar{\mathbf{y}})}\left[\boldsymbol{\Delta}\mathbf{x} \mid \bar{\mathbf{y}}\right]$ we take the empirical mean over N_s samples. To calculate each conditional sample we pass newly sampled Gaussian noise through the inverse normalizing flow conditioned on $\bar{\mathbf{y}}$

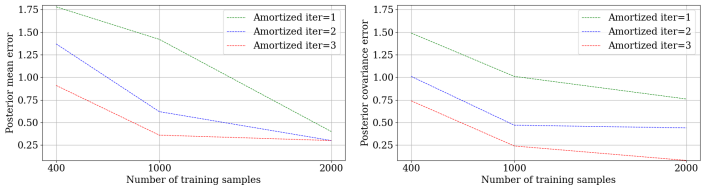
$$\mathbb{E}_{p_{\hat{\theta}}(\boldsymbol{\Delta}\mathbf{x}|\bar{\mathbf{y}})}\left[\boldsymbol{\Delta}\mathbf{x} \mid \bar{\mathbf{y}}\right] \approx \frac{1}{N_{\mathrm{s}}} \sum_{i=1}^{N_{\mathrm{s}}} \boldsymbol{\Delta}\mathbf{x}^{(n)} \text{ where } \boldsymbol{\Delta}\mathbf{x}^{(n)} = f_{\hat{\theta}}^{-1}(\mathbf{z}^{(n)}; \bar{\mathbf{y}}) \text{ and } \mathbf{z}^{(n)} \sim \mathcal{N}(0, I). \tag{7}$$

Posterior mean	PSNR ↑	SSIM ↑	RMSE ↓
Non-iterative ($\mathbf{x}_1$)	40.45	0.977	0.00964
Ours iter=2 ($\mathbf{x}_2$)	42.08	0.982	0.00806
Ours iter=3 ($\mathbf{x}_3$)	**43.27**	**0.985**	**0.00708**

Table 1: Image reconstruction quality metric comparison

Appendix C. Low training sample improvement

As discussed in Section 3.1, the main source of approximation error in amortized VI is from estimating the expectation over the joint distribution $p(\mathbf{x}, \mathbf{y})$ in Equation (2) with limited training samples. To evaluate whether our method can reduce estimation errors due to few training samples, we train the conditional normalizing flow with differing training dataset sizes: $N = 400, N = 1000, N = 2000$. As expected, increasing the amount of training data improves the approximation as shown by the black link in Figure 4, but we can also achieve similar improvement by using our iterative method (compare black line with red line). We emphasize that traditional amortized posterior inference would be limited to the black line in Figure 4 and more training data would be needed to improve the quality of the approximation whereas our method can provide better posterior approximations without additional training data.



(a) Posterior mean (b) Posterior covariance

Figure 4: Approximation errors due to small training datasets can be ameliorated with our iterative method.

Appendix D. Average performance for transcranial ultrasound

For 200 samples from an unseen test set, we simulate observations $\mathbf{y}$ and calculate the posterior at each iteration using our method. The average metrics for Peak Signal Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Root Mean Squared Error (RMSE) are shown in Table 1.