Bayesian Inference in Physics-Driven Problems with Adversarial Priors

Abstract

Generative adversarial networks (GANs) have found multiple applications in the solution of inverse problems in science and engineering. These applications are driven by the ability of these networks to learn complex distributions and to map the original feature space to a low-dimensional latent space. In this manuscript we consider the use of GANs as priors in physics-driven Bayesian inference problems. Within this approach the posterior distribution is learnt by mapping the problem to the latent space of the GAN and then using an HMC sampler for efficient sampling. We apply this approach to solving linear and nonlinear inverse problems, including an example with experimental data acquired from an application in biophysical imaging. Furthermore, we analyze the weak convergence of the approximate prior to the true prior and elucidate its dependence on the capacity of the network and the number of training samples.

1 Introduction

Bayesian inference is often used to quantify the uncertainty in inferring parameters and fields for problems that are constrained by physical principles [1, 2, 3]. Within this approach, Bayes rule is used to assert that the conditional distribution of the inferred parameters is proportional to the product of the likelihood of the measurement and the prior distribution of the parameters. Thereafter, any point estimate for the posterior-distribution can be evaluated by sampling from this distribution. There are two challenges in applying Bayesian inference to practical problems. The first is the so-called curse of dimensionality which restricts its application to problems with small number of independent parameters (in the tens). The second is the difficulty associated with constructing complex prior distributions that need to be learned from data.

Generative adversarial networks (GANs) [4] are algorithms based on deep neural networks that can address the challenges described above. Through adversarial learning, the generative component of a GAN learns complex distributions from i.i.d. samples and can therefore be used to represent the prior distribution of parameters. Further, a GAN derives its efficiency by mapping the feature space to a lower-dimensional latent space. In doing so it also provides a lower-dimensional representation for the parameters to be inferred, thus making the inference problem tractable.

Significant recent work has been done on using GANs for solving inverse problems. This includes using GANs for solving problems in compressed sensing [5], to learn regularizers [6], to learn the posterior distribution in a linear inverse problem [7], and as priors in simple, linear inverse problems [8]. In this manuscript we describe two new developments along these lines. We focus on the case where a Wasserstein GAN (WGAN) [9, 10] is used to represent the prior density in Bayesian inference. Our contributions are: (1) a proof of convergence of the prior density learned by the WGAN to the underlying true distribution with increasing model capacity and training data. This result provides the foundation for the use of WGAN in Bayesian inference to estimate the posterior density. (2) The application of this approach to non-linear physics-informed inverse problems in heat conduction and elasticity. For the inverse elasticity problem we present results on experimental data.

Second Workshop on Solving Inverse Problems with Deep Networks (NeurIPS 2020), Online.

2 Problem formulation

We consider the problem of inferring a field $\boldsymbol{x} \in \Omega_x \subset \mathbb{R}^{N_x}$ from the measurement $\boldsymbol{y} \in \Omega_y \subset \mathbb{R}^{N_y}$. Let $\boldsymbol{f} : \Omega_x \mapsto \Omega_y$ be the forward map i.e., $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, which is assumed to be known either exactly or approximately. In practice, the measurements may be corrupted by noise, $\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\eta}$, where $\hat{\boldsymbol{y}}$ is the noisy measurement and $\boldsymbol{\eta}$ is the noise. The well-posedness of $\boldsymbol{f}$ does not guarantee the well-posedness of its inverse, making the inference of $\boldsymbol{x}$ from $\hat{\boldsymbol{y}}$ challenging. Bayesian inference provides a way to tackle the lack of well-posedness.

We assume a prior distribution p_X on $\boldsymbol{x}$. Let $p^l(\boldsymbol{y}|\boldsymbol{x})$ be the likelihood of $\boldsymbol{y}$ given an instance of $\boldsymbol{x}$. Then, we get the posterior distribution of $\boldsymbol{x}$ given $\hat{\boldsymbol{y}}$ by Bayes' rule

$$p_X^{\text{post}}(\boldsymbol{x}|\hat{\boldsymbol{y}}) = \frac{1}{Q} p^l(\hat{\boldsymbol{y}}|\boldsymbol{x}) p_X(\boldsymbol{x}) = \frac{1}{Q} p_\eta(\hat{\boldsymbol{y}} - \boldsymbol{f}(\boldsymbol{x})) p_X(\boldsymbol{x}), \tag{1}$$

where p_η is the model distribution for noise, and Q is the prior-predictive distribution of $\boldsymbol{x}$. The posterior encapsulates the uncertainty in $\boldsymbol{x}$ given the measured data. However, the dimension N_x of $\boldsymbol{x}$ is typically very large. This makes the estimation of statistical quantities computationally challenging.

This problem can be resolved by mapping the $\boldsymbol{x}$ to a latent space variable $\boldsymbol{z} \in \mathbb{R}^{N_z}$ sampled from a distribution p_Z, with $N_z \ll N_x$. We achieve this by training a WGAN with a generator $\boldsymbol{g}_{N_\theta}(\boldsymbol{z}; \boldsymbol{\theta})$ and discriminator $d_{N_\phi}(\boldsymbol{x}; \boldsymbol{\phi})$, where $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$, $\boldsymbol{\phi} \in \mathbb{R}^{N_\phi}$ are the trainable network parameters. The training makes use of a dataset $\mathcal{S} = \{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}\}$ of N i.i.d. samples from the true distribution p_X to learn the true distribution. Specifically, by assuming that a) the discriminator is smooth, and b) the set of derivatives of the discriminator with respect to its parameters form a dense subset of the space of bounded continuous functions in the limit of infinite capacity, we prove the following statistical error estimate for any $\mathcal{F} \in C_b(\Omega_x)$ in Appendix A,

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^N \sim p_Z^N} \left[\left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_X} [\mathcal{F}(\boldsymbol{x})] - \frac{1}{N} \sum_{i=1}^{N} \mathcal{F}(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}^{(i)}; \boldsymbol{\theta}_N^{*,N_\theta})) \right)^2 \right] < \frac{3}{N} \mathop{\text{Var}}_{\boldsymbol{x} \sim p_X} [\mathcal{F}(\boldsymbol{x})] + \frac{C}{(N_\phi)^{2\alpha}} \tag{2}$$

where p_Z^N is the distribution of $\boldsymbol{z}^N := (\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(N)})$ while C and α are positive constants depending on $\mathcal{F}$. The result above demonstrates how the point estimate $\mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_X} [\mathcal{F}(\boldsymbol{x})]$ can be approximated by a finite sum of samples drawn from the latent space and pushed through the generator.

Under the additional assumption that $\boldsymbol{f}$ and p_η are continuous, we can show that in the limit of infinite capacity ($N_\phi \to \infty$) and with infinite sampling ($N \to \infty$), we have (see Appendix B)

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_X^{\text{post}}} [\mathcal{F}(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_Z^{\text{post}}} [\mathcal{F}(g_\infty(\boldsymbol{z}))], \tag{3}$$

for every $\mathcal{F} \in C_b(\Omega_X)$, where g_∞ is the generator obtained in the limit and $\boldsymbol{z}$ is drawn from the posterior distribution of the latent space

$$p_Z^{\text{post}}(\boldsymbol{z}|\hat{\boldsymbol{y}}) = \frac{1}{Q} p_\eta(\hat{\boldsymbol{y}} - \boldsymbol{f}(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}, \boldsymbol{\theta}))) p_Z(\boldsymbol{z}). \tag{4}$$

Thus our approach to estimating any point estimate for the posterior involves: (a) training the generator of the GAN with i.i.d. samples drawn from p_X, (b) training an MCMC algorithm using (4) to generate samples from p_Z^{post}, and (c) using these samples in the RHS of (3) to approximate the desired point estimate. We note that all these calculations are restricted to the latent space and the significantly smaller dimension of $\boldsymbol{z}$ makes the computation of various statistical quantities tractable.

3 Results

Inverse heat conduction - inferring thermal conductivity We consider a non-linear coefficient inversion problem arising in thermal imaging. The forward model is given by the steady-state heat conduction equation (see Appendix C.1). The goal of coefficient inversion problem is to recover the thermal conductivity $\boldsymbol{x}(\boldsymbol{s})$ at each location given the noisy (and potentially partial) measurement of temperature $\boldsymbol{y}(\boldsymbol{s})$.

We consider two distinct instances of the dataset $\mathcal{S}$: one is obtained by sampling from an underlying parametric description, and the other is not. The parametric dataset allows the computation of the

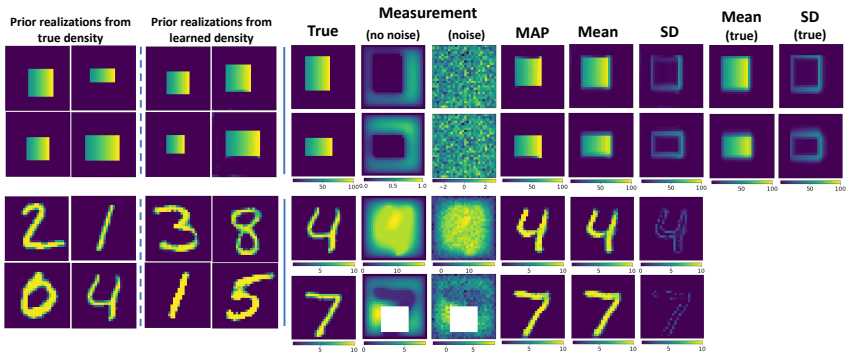

Figure 1: Prior realizations and inference of thermal conductivity. Right panel : (row 1-2) recovery from noisy parametric measurements, (row 3) recovery from noisy non-parametric measurement, (row 4) recovery from noisy non-parametric partial measurement.

true posterior statistics, thereby enabling the verification of estimated statistics. For the parametric dataset, we consider conductivity with a background of 1, and a rectangular sub-domain where it varies linearly from 50 units on the left edge to 100 units on the right edge. The coordinates of the the lower left and upper right corners of the rectangle parameterize this distribution. The dataset $\mathcal{S}$ contains 10,000 images generated by sampling each parameter from a uniform distribution. For the non-parametric case, we consider MNIST dataset. Realizations from the true prior density as well as the density learned by the WGAN are shown in the first two panels of Figure 1. In the right panel of Figure 1 we have shown the true thermal conductivity, the measured temperature (with and without noise), the fields inferred using our method (MAP, mean and standard deviation) and the noisy temperature. For the parametric case, the true mean and SD are determined by Monte Carlo sampling in the parametric space. We observe that (a) the MAP is close to the true distribution in every case, even in the presence of significant noise; (b) the estimated mean and SD are close to the true values for the case where the latter can be determined; (c) the SD is large in regions where the conductivity has sharp gradients; (d) even when a large amount of measured signal is occluded (last row), the strong prior for the problem is able to accurately recover the true conductivity.

Inverse Radon transform We demonstrate the efficacy of the proposed Bayesian inference strategy in the context of inverse Radon transforms. Given an input image $\boldsymbol{x} \in \mathbb{R}^{N\times N}$, the forward maps leads to an output $\boldsymbol{y} \in \mathbb{R}^{N\times N}$ comprising N^2 one-dimensional Radon transforms, which are essentially integrals of the image density along straight lines passing though the image (see Appendix C.2). For the current problem, the inferred field is generated using the modified Shepp-Logan head phantom [11], which is composed of a union of ten ellipses with distinct constant relative densities. The total density at any point in the phantom ranges from 0 (air cavity) to 1 (bone).

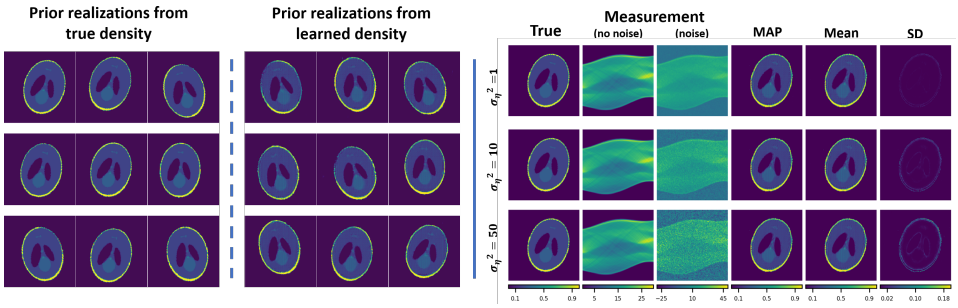

Figure 2: Prior realizations and inference of the Shepp-Logan phantom from noisy Radon transforms.

We generated a data set $\mathcal{S}$ of 10,000 phantom images of dimension 128×128 by randomly perturbing the axes lengths, axis inclination and densities of each ellipse, and further rotating and linearly shifting the whole phantom. Care was taken to ensure that the perturbed images did not violate the underlying spatial topology of the phantom, and that the density at each point stayed within the valid range. A few samples from this data set are shown in the the left panel of Figure 2. The output

of the forward map, also called a sinogram, is depicted in the second column of the right panel in Figure 2 for the phantom in column one of the same panel. A WGAN is trained using the data set $\mathcal{S}$ to approximate the prior of the inferred field. The realizations from the learned prior shown in the second panel of Figure 2, once again demonstrate the strength of GANs in capturing the target distribution. We choose a sample from the true prior (not in the training set) and add Gaussian noise to the measurement. We consider three different noise levels, as shown in each row of the right panel in Figure 2. For each case, we recover the distribution of the likely phantom image using the MCMC approach, with the MAP estimate, mean and pixel-wise standard deviations shown in the last three columns of the left panel in Figure 2. Even for the highest noise level in the measurement, we get an excellent recovery of the phantom, with a relatively low variance concentrated along the boundaries of the ellipses.

Elasticity imaging We provide results for an imaging problem involving experimental measurements. Elasticity imaging is a technique of inferring mechanical properties of tissue from displacement data collected via different medical imaging modalities [12]. The forward problem is given by elasticity problem which solves for the displacement field of an incompressible linear elastic solid (see Appendix C.3. We are interested in recovering the shear modulus field $\boldsymbol{x}$, given a noisy observation of displacement field $\boldsymbol{y}$. For this study, we used experimental data obtained from a physical phantom experiment [13]. The phantom was manufactured from a mixture of gelatin, agar, and oil and contained a spherical inclusion with an elevated shear modulus compared to the background. The phantom was subjected to uniaxial loading and the interior deformation was measured using ultrasound.

The sample set $\mathcal{S}$ contained 3,000 images of elliptical inclusions centered around different locations inside discretized into 56^2 grid points. The ratio of the shear modulus of inclusion to that of the background was varied between 1:1 and 8:1 to account for a wide range of possibilities. A WGAN was trained using this sample set and the learned GAN prior was used in Bayesian inference in conjunction with the experimentally measured displacement field (shown in the leftmost panel of Figure 3). The reconstruction results are shown in the middle panel of Figure 3 and reveal the circular inclusion. These results compare very well with the true physical quantities for the inclusion (see Appendix D). Since the measurement data for this study was obtained experimentally, we do not know the true measurement noise. Given this, we tested the robustness of our algorithm with different assumed values of measurement noise. As shown in the Figure 3, the algorithm produces consistent reconstruction results across different noise values, with elevated pixel-wise standard deviation for the case where higher value of noise is assumed.

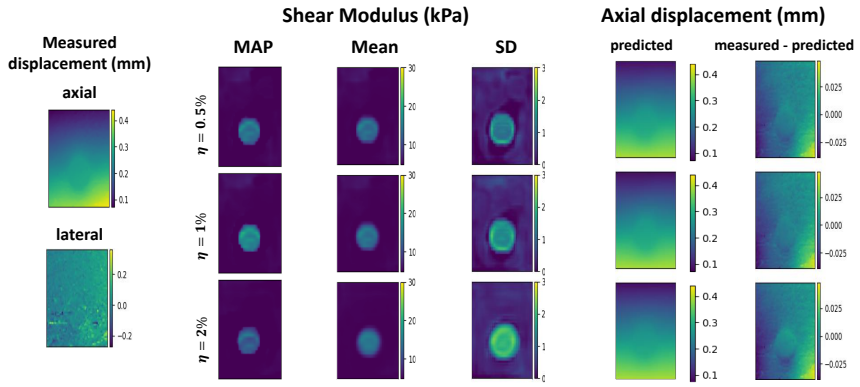

Figure 3: Recovery of shear modulus field from noisy measurements of the displacement field.

4 Conclusions

We have demonstrated how WGANs can be used to learn the prior in a range of Bayesian inference problems. Since GANs inherently map the inferred field to a latent space with a significantly lower dimension, we are able to alleviate the curse of dimensionality, while being able to capture complex distributions. Coupled with an HMC algorithm to determine the posterior, we have presented three physics-driven problems to show the robustness of our method to quantify the uncertainty in determining the underlying inferred field, particularly when the measurement with relatively high noise, which might even be partially occluded.

References

[1] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[2] Masoumeh Dashti and Andrew M Stuart. The bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, pages 1–118, 2016.

[3] Adriano Polpo, Julio Stern, Francisco Louzada, Rafael Izbicki, and Hellinton Takada, editors. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 239 of *Springer Proceedings in Mathematics & Statistics*. Springer International Publishing, Cham, 2018.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[6] Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, pages 8507–8516, 2018.

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

[8] Dhruv Patel and Assad A Oberai. Bayesian inference with generative adversarial network priors. *arXiv preprint arXiv:1907.09987*, 2019.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[11] Peter Aundal Toft. *The Radon Transform - Theory and Implementation*. PhD thesis, Technical University of Denmark, 1996.

[12] Paul E Barbone and Assad A Oberai. A review of the mathematical and computational foundations of biomechanical imaging. In *Computational Modeling in Biomechanics*, pages 375–408. Springer, 2010.

[13] Theo Z Pavan, Ernest L Madsen, Gary R Frank, Antonio Adilton O Carneiro, and Timothy J Hall. Nonlinear elastic behavior of phantom materials for elastography. *Physics in Medicine & Biology*, 55(9):2679, 2010.

A Error estimate for WGAN predictions

Let the generator of the WGAN be $\boldsymbol{g}_{N_\theta}(\boldsymbol{z};\boldsymbol{\theta})$, where

$$\boldsymbol{g}_{N_\theta} : \mathbb{R}^{N_z} \times \mathbb{R}^{N_\theta} \mapsto \Omega_X \subset \mathbb{R}^{N_x}. \tag{5}$$

Similarly, the discriminator $d_{N_\phi}(\boldsymbol{x};\boldsymbol{\phi})$ is the function

$$d_{N_\phi} : \Omega_X \times \mathbb{R}^{N_\phi} \mapsto \mathbb{R}. \tag{6}$$

Consider the loss function being optimized during the WGAN training

$$\mathcal{L}_N(\boldsymbol{\theta},\boldsymbol{\phi}) = \frac{1}{N}\sum_{i=1}^{N} d_{N_\phi}(\boldsymbol{x}^{(i)};\boldsymbol{\phi}) - \frac{1}{N}\sum_{i=1}^{N} d_{N_\phi}\left(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}^{(i)};\boldsymbol{\theta});\boldsymbol{\phi}\right), \tag{7}$$

where $\{\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},...,\boldsymbol{x}^{(N)}\}$ and $\{\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},...,\boldsymbol{z}^{(N)}\}$ are a particular set of N i.i.d. realizations of $\boldsymbol{x} \sim p_X$ and $\boldsymbol{z} \sim p_Z$, respectively. Denote the joint probability distribution of $\boldsymbol{x}^N := (\boldsymbol{x}^{(1)},...,\boldsymbol{x}^{(N)})$, $\boldsymbol{z}^N := (\boldsymbol{z}^{(1)},...,\boldsymbol{z}^{(N)})$ and $(\boldsymbol{x}^N,\boldsymbol{z}^N)$ as

$$p_X^N = \prod_{i=1}^{N} p_X, \qquad p_Z^N = \prod_{i=1}^{N} p_Z, \qquad p_{X,Z}^N = \prod_{i=1}^{N} p_{X,Z},$$

respectively, where $p_{X,Z}$ is the joint distribution of $(\boldsymbol{x},\boldsymbol{z})$.

Define the optimal parameter values

$$\boldsymbol{\theta}_N^{*,N_\theta}, \boldsymbol{\phi}_N^{*,N_\phi} = \underset{\boldsymbol{\phi}\in\mathbb{R}^{N_\phi}}{\arg\max}\left(\underset{\boldsymbol{\theta}\in\mathbb{R}^{N_\theta}}{\arg\min}(\mathcal{L}_N(\boldsymbol{\theta},\boldsymbol{\phi}))\right), \tag{8}$$

for a fixed N and a fixed architecture of the WGAN. Then we have the following result for convergence of the WGAN.

Theorem A.1. *Consider the loss function* (7) *being optimized to train the WGAN, with* (8) *denoting the optimal parameters for a particular number of samples N and a particular finite network architecture. Let us assume the following conditions are true:*

A1 The domain Ω_X is closed and bounded.

A2 The samples for $\boldsymbol{x}$ and $\boldsymbol{z}$ are chosen independently, with $p_{X,Z} = p_X p_Z$.

A3 The discriminator is a C^1 function of it inputs and parameters $\boldsymbol{\phi}$.

A4 Let the derivates of the discriminator with respect to its parameters, denoted by

$$w_k^{N_\phi} := \frac{\partial d_{N_\phi}}{\partial \phi_k} \qquad 1 \le k \le N_\phi. \tag{9}$$

We assume that given $\mathcal{F} \in C_b(\Omega_X)$, there exists $\alpha > 0$ and integers $\widetilde{N}, \widetilde{N}_\phi > 0$ such that for all $N \ge \widetilde{N}$ and $N_\phi \ge \widetilde{N}_\phi$,

$$\|\mathcal{F} - w_{k'}^{N_\phi}(.;\boldsymbol{\phi}_N^{*,N_\phi})\|_\infty < \frac{C_{\mathcal{F}}}{(N_\phi)^\alpha} \quad \underline{\textit{for some}} \quad 1 \le k' \le N_\phi, \tag{10}$$

where $\widetilde{N}, \widetilde{N}_\phi$ and the constant $C_{\mathcal{F}}$ depend only on $\mathcal{F}$ and the activation function used in the networks.

Then, given an $\mathcal{F} \in C_b(\Omega_X)$ and the corresponding $\widetilde{N}, \widetilde{N}_\phi, \alpha$ and $C_{\mathcal{F}}$ from assumption A4, the following error estimate holds true

$$\underset{\boldsymbol{z}^N\sim p_Z^N}{\mathbb{E}}\left[\left(\underset{\boldsymbol{x}\sim p_X}{\mathbb{E}}[\mathcal{F}(\boldsymbol{x})] - \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}^{(i)};\boldsymbol{\theta}_N^{*,N_\theta}))\right)^2\right] < \frac{3}{N}\underset{\boldsymbol{x}\sim p_X}{Var}[\mathcal{F}(\boldsymbol{x})] + \frac{6C_{\mathcal{F}}^2}{(N_\phi)^{2\alpha}} \tag{11}$$

for $N \ge \widetilde{N}$, $N_\phi \ge \widetilde{N}_\phi$.

Proof. Since $\boldsymbol{\theta}_N^{*,N_\theta}, \boldsymbol{\phi}_N^{*,N_\phi}$ satisfy (8), the necessary optimality conditions are

$$\frac{\partial \mathcal{L}_N}{\partial \theta_k}(\boldsymbol{\theta}_N^{*,N_\theta}, \boldsymbol{\phi}_N^{*,N_\phi}) = 0 \qquad \forall\, 1 \le k \le N_\theta, \tag{12}$$

$$\frac{\partial \mathcal{L}_N}{\partial \phi_k}(\boldsymbol{\theta}_N^{*,N_\theta}, \boldsymbol{\phi}_N^{*,N_\phi}) = 0 \qquad \forall\, 1 \le k \le N_\phi. \tag{13}$$

Using (7) in (13) and the notation introduced in (9), we get

$$\frac{1}{N}\sum_{i=1}^{N} w_k^{N_\phi}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_N^{*,N_\phi}) = \frac{1}{N}\sum_{i=1}^{N} w_k^{N_\phi}\left(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}^{(i)}; \boldsymbol{\theta}_N^{*,N_\theta}); \boldsymbol{\phi}_N^{*,N_\phi}\right) \qquad \forall\, 1 \le k \le N_\phi. \tag{14}$$

Given $\mathcal{F} \in C_b(\Omega_X)$, we use (14) and Assumption A4 to get

$$\begin{aligned}
\mathcal{E}^2_{N,N_\theta,N_\phi}(\mathcal{F}) :=& \Big(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_X}[\mathcal{F}(\boldsymbol{x})] - \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}^{(i)}; \boldsymbol{\theta}_N^{*,N_\theta}))\Big)^2 \\
=& \Big(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_X}[\mathcal{F}(\boldsymbol{x})] - \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}(\boldsymbol{x}^{(i)}) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}(\boldsymbol{x}^{(i)}) - \frac{1}{N}\sum_{i=1}^{N} w_{k'}^{N_\phi}(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_n^{N_\phi}) \\
&+ \frac{1}{N}\sum_{i=1}^{N} w_{k'}^{N_\phi}(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}^{(i)}; \boldsymbol{\theta}_N^{*,N_\theta}); \boldsymbol{\phi}_n^{N_\phi}) - \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}^{(i)}; \boldsymbol{\theta}_N^{*,N_\theta}))\Big)^2 \\
<& 3\left(\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_X}[\mathcal{F}(\boldsymbol{x})] - \frac{1}{N}\sum_{i=1}^{N}\mathcal{F}(\boldsymbol{x}^{(i)})\right)^2 + \frac{3C_{\mathcal{F}}^2}{(N_\phi)^{2\alpha}} + \frac{3C_{\mathcal{F}}^2}{(N_\phi)^{2\alpha}},
\end{aligned} \tag{15}$$

for $N \ge \widetilde{N}, N_\phi \ge \widetilde{N}_\phi$, where $\widetilde{N}, \widetilde{N}_\phi, \alpha$ and k' are as defined in Assumption A4. Using assumption A2, we can split the joint probability distribution of $(\boldsymbol{x}^N, \boldsymbol{z}^N)$ as

$$p_{X,Z}^N = \prod_{i=1}^{N} p_{X,Z} = \prod_{i=1}^{N}(p_X p_Z).$$

Since $\mathcal{F}$ is bounded, we have

$$\mathop{\mathrm{Var}}_{\boldsymbol{x}\sim p_X}[\mathcal{F}(\boldsymbol{x})] < \infty, \quad \mathop{\mathrm{Var}}_{\boldsymbol{z}\sim p_Z}\left[\mathcal{F}(\boldsymbol{g}_{N_\theta}(\boldsymbol{z}; \boldsymbol{\theta}_N^{*,N_\theta}))\right] < \infty.$$

Taking expectation on both sides of (15) with respect to $p_{X,Z}^N$ and using the linearity of expectation and the standard Monte Carlo error estimate, we have for $N \ge \widetilde{N}, N_\phi \ge \widetilde{N}_\phi$

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^N\sim p_Z^N}\left[\mathcal{E}^2_{N,N_\theta,N_\phi}(\mathcal{F})\right] < \frac{3}{N}\mathop{\mathrm{Var}}_{\boldsymbol{x}\sim p_X}[\mathcal{F}(\boldsymbol{x})] + \frac{6C_{\mathcal{F}}^2}{(N_\phi)^{2\alpha}}. \tag{16}$$

□

Remark A.1. *We make a few comments about the assumptions of the above theorem.*

- *Assumption A1 is meaningful since the training data is typically scaled to a suitable hypercube before being used with the network.*
- *Assumptions A1 and A3 ensure that $w_k^{N_\phi} \in C_b(\Omega_X)$. Note that A3 can be guaranteed if we take a C^1 activation function in the discriminator.*
- *The space $C_b(\Omega_X)$ is separable, i.e., it is guaranteed to contain a countable dense subset. Assumption A4 implies that in the limit of infinite capacity ($N_\phi \to \infty$) and with sufficient number of samples ($N \to \infty$), the set $\{w_k^{N_\phi}\}$ converges to such a countable dense subset of $C_b(\Omega_X)$. Note that this is indeed a subset due to Assumption A3.*
- *Another implicit implication of Assumption A4 is that the trained WGAN steers away from modal collapse.*

B Convergence to the Bayesian posterior

Let us assume that in the limit of infinite capacity ($N_\phi \to \infty$) and infinite sampling ($N \to \infty$), the generator $\boldsymbol{g}_{N_\theta}(\boldsymbol{z}; \boldsymbol{\theta}_N^{*,N_\theta})$ converges to $\boldsymbol{g}_\infty(\boldsymbol{z})$. Then in this limit, the result of Theorem A.1 implies that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_X}[\mathcal{F}(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim p_Z}[\mathcal{F}(\boldsymbol{g}_\infty(\boldsymbol{z}))] \tag{17}$$

for any $\mathcal{F} \in C_b(\Omega_X)$. Thus, the approximate prior weakly converges to the true prior.

Furthermore, let us assume that the forward map $\boldsymbol{f}$ noise density p_η are both continuous functions. Consider the function

$$\mathcal{G}(\boldsymbol{x}) = \frac{\mathcal{F}(\boldsymbol{x}) p_\eta\big(\hat{\boldsymbol{y}} - \boldsymbol{f}(\boldsymbol{x})\big)}{Q}. \tag{18}$$

Then clearly $\mathcal{G} \in C_b(\Omega_X)$. Substituting (18) in (17), and using (1) gives us

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_X^{\text{post}}}[\mathcal{F}(\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim p_Z^{\text{post}}}[\mathcal{F}(\boldsymbol{g}_\infty(\boldsymbol{z}))]\,, \tag{19}$$

where $\boldsymbol{z}$ is sampled from the posterior distribution of the latent space (4). In other words, we obtain an expression for evaluating the statistics with respect to the posterior distribution. Since (19) holds for any $\mathcal{F} \in C_b(\Omega_X)$, we get the weak convergence of the posterior distribution.

C Forward models

The various forward models considered in this work are described below.

C.1 Heat conduction

Consider the following steady-state heat conduction equation

$$-\nabla \cdot (\boldsymbol{x}(\boldsymbol{s})\nabla \boldsymbol{y}(\boldsymbol{s})) = \boldsymbol{b}(\boldsymbol{s}), \qquad \boldsymbol{s} \in (0,1)^2 \tag{20}$$

$$\boldsymbol{y}(\boldsymbol{s}) = 0, \qquad \boldsymbol{s} \in \partial(0,1)^2 \tag{21}$$

where $\boldsymbol{x}(\boldsymbol{s})$, $\boldsymbol{y}(\boldsymbol{s})$, and $\boldsymbol{b}(\boldsymbol{s})$ denote thermal conductivity, temperature, and the heat source respectively. The goal of coefficient inversion problem is to recover the thermal conductivity at each location given the noisy (and potentially partial) measurement of temperature. For our simulations, we take the source term to be constant, with $\boldsymbol{b}(\boldsymbol{s}) = 100$.

C.2 Radon transform

The forward map for this problem is given by one-dimensional Radon transforms, which are integrals of the phantom density along straight lines passing through the phantom image. Given an input $\boldsymbol{x}$ image of size $N \times N$, the forward map evaluates the line integral of N equally spaced lines though the image inclined at an angle ψ for N equally spaced angles in $(0°, 180°)$,

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y} \in \mathbb{R}^{N\times N}, \quad \boldsymbol{y}_{i,j} = \int \ell_{t_i,\psi_j} \boldsymbol{x} d\ell, \quad \forall\, 1 \le i,j \le N \tag{22}$$

where ℓ_{t_i,ψ_j} is the line through the image inclined at an angle ψ_j and at a signed-distance of t_i from the center of the image.

C.3 Elasticity imaging

Consider the following partial differential equation,

$$\nabla \cdot \boldsymbol{\sigma} = \boldsymbol{0} \qquad \text{in } \Omega \tag{23}$$

$$\boldsymbol{y} = \boldsymbol{y}_D \qquad \text{on } \Gamma_D \tag{24}$$

$$\boldsymbol{\sigma} \cdot \boldsymbol{n} = \boldsymbol{\tau} \qquad \text{on } \Gamma_N \tag{25}$$

where, $\boldsymbol{\sigma} = 2\boldsymbol{x}(\nabla^s \boldsymbol{y} + (\nabla \cdot \boldsymbol{y})\mathbb{1})$ for plane-stress incompressible linear elastic solid, $\boldsymbol{x} \in \mathbb{R}^{N\times N}$ is the discretized shear modulus and $\boldsymbol{y} \in \mathbb{R}^{N\times N\times 2}$ is the discretized displacement field. We are interested in recovering the shear modulus field $\boldsymbol{x}$, given a noisy observation of displacement field $\boldsymbol{y}$.

D Errors in the Elasticity Imaging Reconstructions

The important physical parameters of the inclusion and their reconstructed values are:

1. Horizontal dimension: true = 10 mm, MAP-based reconstruction = 10.22$\pm$0.49 mm, mean-based reconstruction = 10.06$\pm$0.27 mm.
2. vertical dimension: true = 10 mm, MAP-based reconstruction=10.30$\pm$0.35 mm, mean-based reconstruction=10.30$\pm$0.35 mm.
3. Mean true shear modulus: true = 10.7 kPa, MAP-based reconstruction = 11.05$\pm$1.32 kPa, mean-based reconstruction=10.94$\pm$0.62 kPa