Probabilistic seismic interpolation with the implicit Prior of a Deep Denoiser
Matteo Ravasi, KAUST

SUMMARY

Geophysicists have long recognized the importance of quantifying the uncertainty associated with geophysical inverse problems, whether they are used for processing, imaging, or parameter estimation purposes. The inability to create representative prior and proposal distributions has, however, hindered the widespread adoption of acceptance-rejection algorithms, such as those from the family of Monte-Carlo Markov Chain methods. We present a flexible approach to probabilistic sampling that leverages the ability of denoising neural networks to provide direct access to the gradient of the log-probability of interest. The proposed algorithm can produce high-quality, diverse samples from both unconditional and conditional probability distributions, the latter being of particular interest when solving probabilistic inverse problems. A successful application is presented in the context of seismic interpolation on both synthetic and field data.

INTRODUCTION

It is widely recognized that deep neural networks can be trained to act as powerful denoisers (Zhang et al., 2017, 2018). Recently, such networks have been embedded into various proximal algorithms to solve scientific inverse problems - leading to new inversion frameworks commonly referred to as Plug-and-Play priors (Venkatakrishnan et al., 2013; Meinhardt et al., 2017) and Regularization by Denoising (RED – Cohen et al., 2021). Early successful applications of such concepts in geophysics can be found in seismic interpolation (Zhang et al., 2020; Xu et al., 2022), deblending (Luiken et al., 2022), and inversion (Romero et al., 2022; Izzatullah et al., 2022) problems. Nevertheless, despite the fact that the process of adding random noise to a signal is stochastic in nature, such solvers are tailored to provide deterministic, maximum-a-posterior solutions without any information about the uncertainty associated with the problem to be solved.

In this work, we leverage the capabilities of properly trained denoising neural networks to develop a coarse-to-fine gradient ascent algorithm, which can produce high-probability samples of a distribution of interest. Following Kadkhodaie and Simoncelli (2021), the proposed algorithm capitalizes on the property of a poorly known statistical theorem stating that ‘a minimum mean squared error denoiser acting on signals corrupted by additive Gaussian noise may be interpreted as computing the gradient of the log of the density of noisy signal’. As such, one can start from a random noise realization and slowly walk into the manifold of plausible solutions learned by a denoiser, by repeatedly applying such denoiser alongside introducing a new small corruption to avoid getting stuck in local maxima of the distribution. By choosing different starting noise realizations, different samples of the distribution can be built. A simple modification of the algorithm allows for sampling of conditional distributions, which in our case represent the posterior distribution of the inverse problem of interest. The overall sampling procedure is displayed in Figure 1 and showcased on a seismic interpolation problem with both synthetic and field data.

THEORY

We are interested to learn the prior distribution $p(x)$ of multi-dimensional signals (i.e., seismic wavefields) from a set of training samples, $X = \{x_0, x_1, \ldots, x_n\}$, as well as to use it to solve linear inverse problems of the form $d = Gx + \varepsilon$ in a stochastic manner (i.e., sampling realizations of the posterior distribution).

Provided a reasonable degree of similarity between the available samples, we can expect them to lie in a low-dimensional manifold (and to have constant or slowly-varying probability), whilst distorted version of such samples (e.g., perturbed by additive random noise) lie off the manifold and therefore have low or zero probability. Assuming now that we can access noisy observations of the form $y = x + n$ where $n \sim \mathcal{N}(0, \sigma^2 I)$, the probability density function of such observations is directly linked to that of the original samples via:

$$
p_{\sigma}(y) = \int p(y|x)p(x)dx = \int g_{\sigma}(y-x)p(x)dx.
$$

As the last expression of equation 1 is in the form of a convolution, $p_{\sigma}(y)$ can be interpreted as a Gaussian-blurred version of $p(x)$, with the level of blurring $g_{\sigma}(x)$ being directly proportional to the standard deviation of the additive noise.

Given such noisy observations, the minimum mean squared error estimate of the clean signal can be equivalently represented by the conditional mean of the posterior density:

$$
\hat{x}(y) = \int x p(x|y)dx
$$

independently on the choice of the actual denoiser. By exploiting a result from the literature on Empirical Bayesian estimation (Miyasawa, 1961), Kadkhodaie and Simoncelli (2021) showed that in the case of Gaussian additive noise one can rewrite equation 2 (with the help of equation 1) into a more direct form:

$$\
\hat{x}(y) = y + \sigma^2 \nabla_x \log p_{\sigma}(y)
$$

This result provides a direct link between the gradient of the noisy probability distribution, the noisy realization, and its denoised counterpart. More precisely, the gradient of $p_{\sigma}(y)$ is equivalent to the residual of the denoiser $f_{\sigma}(y) = \hat{x}(y) - y$ (i.e., the difference between the denoised and noisy samples), where the subscript is used to emphasize the fact that the denoiser is provided with a noisy sample with standard deviation $\sigma$. In
Seismic interpolation with the Prior of a Deep Denoiser

\[ y_0 \sim \mathcal{N}(G^Hd, \sigma_0^2I) \]

\[ y_{t-1} \]

\[ \text{noise estimate:} \quad f_\sigma(y_{t-1}) = \sigma^2 \nabla \log p(y_{t-1}) \]

\[ y_t = y_{t-1} + h_t f(y_{t-1}) + \gamma z \]

\[ y_t \]

\[ y_{N_{1k}} \]

Figure 1: Schematic illustrations of the coarse-to-fine gradient ascent method to sample the posterior distribution of the seismic interpolation problem.

practice, the denoiser can be learned in a supervised manner provided access to pairs of clean-noisy training data, and later used in equation 3. A detailed description of the actual denoiser used in the work is postponed to the Numerical Examples section.

We can now devise a simple stochastic gradient ascent algorithm to sample from the unknown prior distribution \( p(x) \). Starting from a random realization \( y_0 \sim \mathcal{N}(0, \sigma_0^2I) \), where \( \sigma_0 \) is a user-defined initial standard deviation, this algorithm simply takes advantage of the gradient information to move the initial guess towards regions of high-probability of \( p(x) \). Note that, as the process evolves, the noise reduces and ultimately converges to the unknown \( p(x) \) (i.e., the algorithm produces samples belonging to the manifold of the actual prior of interest). Because of equation 1, we can also interpret this process as first optimizing a blurred version of the original probability density function (and recovering large-scale features) and subsequently introducing details as the underlying probability function sharpens. The algorithm can be compactly written as:

\[ y_t = y_{t-1} + h_t f(y_{t-1}) + \gamma z \]

(4)

where \( h_t \) is an iteration-dependent step-size, which controls the fraction of noise correction taken into account at iteration \( t \). Note that an additional amount of noise \( z \sim \mathcal{N}(0, I) \) is added at every iterations, whose extent is controlled by the parameter \( \gamma \). This additional noise allows the method to avoid getting stuck in local maxima; moreover, it ensures stochastic exploration of the manifold, yielding a more diverse (higher entropy) family of solutions.

The choice of \( h_t \) and \( \gamma \) is crucial to the success of the algorithm: they must in fact ensure that the effective noise level decreases throughout iterations despite the injection of additional noise. By doing so, the observable dimensionality of the sought after manifold increases, enabling the synthesis of more and more details in the generated samples. To be able to consistently choose such parameters during iterations, one must have access to the variance of the noise remaining in the sample. As suggested in Mohan and Simoncelli (2020), the magnitude of the denoising residual can be directly used as it provides a good estimate of such a quantity. More precisely, \( h_t \) is chosen to slowly increase through iterations \( h_t = (h_{ft})/(1 + h_t(t - 1)) \), whilst \( \gamma \) is derived directly from the condition on the effective noise variance reduction \( \gamma = ((1 - \beta h_t)^2 - (1 - h_t)^2) \sigma_0^2 \) where \( \beta \in [0, 1] \) is a user-defined parameter controlling the level of injected noise (\( \beta = 1 \), no noise). In general, the smaller \( \beta \) the slower the sampling process; however, a small \( \beta \) usually produces higher-quality samples.

Finally, the algorithm in equation 4 can be extended to the case of conditional sampling, producing samples from the conditional (or posterior distribution), \( p(x|d) \). Geometrically, this can be seen as a constrained sampling problem, corresponding to drawing samples which sit at the intersection of the prior manifold and a constraining hyperplane. Mathematically, this entails using the gradient of the conditional distribution:

\[ \sigma^2 \nabla \log p(x|d) = (I - G^H G) f(y) + G^H (d - G y) \]

(5)

We refer the reader to Kadkhodaie and Simoncelli (2021) for a detailed derivation. Similarly, the starting guess must be modified to include our knowledge of the observed data as follows. \( y_0 \sim \mathcal{N}(G^H d, \sigma_0^2 I) \).

NUMERICAL EXAMPLES

The proposed methodology is applied to the problem of seismic interpolation. Although this is an inverse problem with a simple modelling operator (i.e., \( G \) is a restriction operator), its posterior distribution can be very easily assessed: as the observed data does not provide any information of what the solution should look like in the missing traces, the posterior in such traces is fully driven by the prior. As such, one would expect more variability (i.e., higher uncertainty) in the solution for those gaps that have been filled during the stochastic sampling process.

To begin with, we consider a synthetic dataset based on a subsurface velocity model that mimics the V olve field (see Ravasi et al. (2022) for more details on the data creation process). Second, we consider the V olve field dataset. In both cases, we assume densely sampled sources (as usually the case in ocean-bottom acquisition systems) and coarsely sampled receivers. Whilst we show here a scenario with irregular sampling (including a rather large gap in the near offset of the shot gather used in the interpolation process), similar results have been also achieved in the case of regularly subsampled data.

First, a bias-free network is trained on patches extracted from the available common-receiver gathers and amplitude normalized between 0 and 1. At each step of the training process, a batch of patches is corrupted with random noise with variable standard deviation (ranging from 0 to 0.4) and the network is
Seismic interpolation with the Prior of a Deep Denoiser

After 200 epochs, the denoising capabilities of the network are excellent for both the synthetic and field data. The choice of bias-free networks is motivated by their ability to adapt to a variety of noise levels (even slightly outside of those seen during training) as explained in Mohan and Simoncelli (2020), which is of great benefit in the context of our stochastic sampling algorithm. For the network architecture, we have adopted a standard UNet-like architecture with residual blocks; the benefit of using a fully convolution network is that the size of the input can change from training to inference, which we exploit here by training on rather small patches whilst solving the interpolation problem on an entire shot gather.

After training, the denoising network can be embedded in both the unconditional and conditional sampling processes described above. Starting from the unconditional process, a noise realization of the same size of the training patches is fed into the network trained on the synthetic data, and after 51 iterations a plausible wavefield is produced with features that resemble both linear and curved arrivals (Figure 2). Finally, 50 random noise realizations of the size of an entire shot gather are created for both the synthetic and field data and used as starting guess for the conditional version of the proposed algorithm. In both cases, the available data is sampled every 50m (whilst the training data is sampled every 25m); we also introduce an additional irregular subsampling where 60% of the available traces are discarded. Figures 3 and 4 show the shot gathers with all available traces to be used as comparison, their subsampled counterpart, the reconstruction mean over all realizations, and its corresponding mean-normalized standard deviation. The reconstruction mean suggests that the trained denoiser acts as a strong implicit prior for the interpolation problem, making it possible to successfully reconstruct missing gaps in the data from nearby available information, ultimately producing shot gathers with the same spatial resolution of the receiver gathers used in training. The mean-normalized standard deviation highlights how different noise realizations have converged to slightly different solutions, which mainly differ from each other in areas with larger gaps. Moreover, by looking at three different realizations near the largest gap (bottom row of Figures 3 and 4), we can conclude that the reconstructed wavefields are plausible and ‘seismic-like’ in all cases. This result opens doors to the creation of truly multi-realization seismic processing flows, where uncertainties are propagated between steps.

Figure 2: Generation process for a single patch obtained by evaluating equation 4 for \( T = 51 \) iterations using the pre-trained denoising network for the synthetic data and the following parameters, \( \sigma_0 = 1.0, \sigma_T = 0.005, h_0 = 0.01, \) and \( \beta = 0.5 \).

DISCUSSION AND CONCLUSIONS

Whilst we have shown that the current algorithm can produce high-quality samples from the posterior distribution of a seismic interpolation problem, it currently requires a large number of iterations to converge. We envision that the following two modifications may lead to a more efficient version of the algorithm. First, whilst the adopted network is a blind denoiser (i.e., unaware of the noise level of the input), one could alternatively use non-blind denoisers (e.g., Zhang et al., 2021) to control the amount of noise to be removed at each iteration. In such a case, a different strategy may however be required to estimate the variance of noise at each iteration. For example, one could pre-train an auxiliary neural network to estimate the variance of noise from the input sample of the denoiser and feed this directly into the non-blind denoiser. Second, the current sampling process is run completely independently for each initial noise realization; one could instead consider a tree-based sampling strategy, where one (or few) initial samples are run through a number of iterations, and later split into multiple samples (by adding different realizations of noise \( z \) at a given iteration). A comparison between the current approach and one including the proposed modifications is currently under investigation.

In this work, we have presented a new method that, by leveraging Miyasawa’s empirical Bayes estimator, can extend the usage of denoising neural networks to the solution of probabilistic inverse problems. We showcase this method on the problem of seismic interpolation; after training a denoising UNet network on the available seismic data in the well-sampled domain, a carefully designed gradient ascent process is used to produce many plausible realizations of interpolated seismic data whilst successfully filling the missing gaps in the input data.

ACKNOWLEDGMENTS

The author thanks KAUST and the DeepWave Consortium sponsors for supporting this research, as well as Equinor and the Volve license partners for releasing the field dataset.
Figure 3: Seismic interpolation results for a shot gather of the synthetic data. Top (from left to right): true shot gather, subsampled shot gather, reconstructed shot gather (mean over 50 realizations), and corresponding mean-normalized standard deviation (white: small, red: large). Bottom: close-up of true data, and three realizations and traces corresponding to the black and blue vertical lines in areas of high and low uncertainty.

Figure 4: Seismic interpolation results for a shot gather of the field data. Key as in Figure 3.