Denoising Score-Matching for Uncertainty Quantification in Inverse Problems

Anonymous Author(s)

Affiliation

Address

email

Abstract

Deep neural networks have proven extremely efficient at solving a wide range of inverse problems, but most often the uncertainty on the solution they provide is hard to quantify. In this work, we propose a generic Bayesian framework for solving inverse problems, in which we limit the use of deep neural networks to learning a prior distribution on the signals to recover. We adopt recent denoising score matching techniques to learn this prior from data, and subsequently use it as part of an annealed Hamiltonian Monte-Carlo scheme to sample the full posterior of image inverse problems. We apply this framework to Magnetic Resonance Image (MRI) reconstruction and illustrate how this approach not only yields high quality reconstructions but can also be used to assess the uncertainty on particular features of a reconstructed image.

1 Introduction

Generative modeling has enjoyed remarkable success in recent years with models such as Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] reaching extremely high quality results on complex high resolution images [Karras et al., 2020]. Yet, GANs are still prone to issues including unstable training and mode collapse, i.e. a lack of diversity in generated images. In addition, GANs only provide a convenient way to sample from the learned distribution, they do not give access to the density function itself. Another class of commonly used models, Variational Auto-Encoders (VAEs) [Kingma and Welling, 2014] have also been able to reach high quality sample resolutions [Vahdat and Kautz, 2020], but can only provide a bound on the density function of the model.

In some practical applications of these generative models however, the most desirable aspect is not necessarily to be able to sample from the model, or directly have access to its density functions, but have access to gradients of the log density function, i.e. its *score*. In particular, in this work we focus on the generic problem of performing Bayesian inference to solve an inverse problem (i.e. deconvolution, denoising, inpainting, etc), using a learned generative model as a Bayesian prior. Such problems can typically be solved by gradient-based inference methods such as Variational Inference (VI), and Langevin or Hamiltonian Monte-Carlo (HMC), but all these methods have in common that they *only rely on the score of the generative model*.

Recent work [Lim et al., 2020, Song and Ermon, 2019] has investigated various approaches to directly train a deep neural network (coined, a *score network*) to estimate the score of a density function instead of the density itself. One of the most promising and scalable approach relies on denoising score-matching, where the score network is trained using a denoising loss on an augmented dataset where various amounts of Gaussian noise is added to the data. Not only has this approach already

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

been demonstrated to reach convincing results in score estimation, but Song and Ermon [2020] for instance demonstrated that these score networks can be used to sample high resolution natural images through Langevin dynamics.

In this work, we develop the first application of deep denoising score matching to Magnetic Resonance Image (MRI) reconstruction and its use for Uncertainty Quantification (UQ) in the context of imaging inverse problems.

2 Related Works

Score Matching Score Matching was introduced by Hyvärinen [2005]. It was later revisited in Vincent [2011] and Alain and Bengio [2013] in the context of Deep Learning and Denoising Auto-Encoders (DAE) in particular. Recently, Song and Ermon [2019] applied these findings to images showing the potential of this approach combined with sampling techniques. They refined their work in Song and Ermon [2020] to take it to high dimension and provide some techniques for hyper-parameter setting.

Plug-and-Play Priors The idea of using a denoising model as a prior for inverse problems solving was introduced by Venkatakrishnan et al. [2013] who used non-learned denoisers to replace the proximity operator in the ADMM algorithm [Boyd, 2011]. More recently, Meinhardt et al. [2017] have made use of learned denoiser networks in Plug-and-Play inverse problems solving. Annealed HMC sampling can be seen as a fuzzy-version of these approaches with theoretical grounds.

Deep Uncertainty Quantification in Inverse Problems The idea of using generative models for inverse problems such as data imputation and denoising was for instance suggested in an early VAE work [Rezende et al., 2014]. In this approach, a generative network is first trained on high quality data, and inference is then performed in the latent space of the model typically using VI. The main advantage comes from the reduced dimensionality of the parameter space. Recent examples of this approach include [Wu et al., 2018, Böhm et al., 2019], and an application to MRI can for instance be found in [Edupuganti et al., 2019]. An important aspect of this approach is that it relies on the known likelihood of observations.

A different approach was proposed in Adler and Öktem [2018] based on a conditional Wasserstein Generative Adversarial Network (cWGAN). In this formulation, a generative model is trained to sample high quality images conditioned on degraded observations. At test time, independent samples from the posterior distribution induced by the GAN are obtaind by simply sampling different latent space variables. Similar frameworks based on other conditional generative models have alse been proposed, such as [Denker et al., 2020] using conditional Normalizing Flows, or [Tonolini et al., 2019] using conditional VAEs. While this conditional generative model approach allows for fast inference, thanks to ancestral sampling, it is worth noting that none of these models include an explicit *data consistency* step, nor do they make use of a test-time likelihood. As a direct consequence, for instance, these models cannot be used on data with different noise levels as seen during training.

3 Deep Denoising Score Matching for Posterior Inference

Our main objective in this work is to perform probabilistic inference over a Bayesian model typically represented as

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{y}$ are some measurements and the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ is the distribution of possible solutions $\boldsymbol{x}$ compatible with observations and a our a prior knowledge. The problem-specific likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ encodes the forward process of the model and accounts for observational noise, while the prior $p(\boldsymbol{x})$ encapsulates any a priori information we have on the solution of the problem. Our goal for UQ is to sample solutions $\boldsymbol{x}$ belonging to that posterior. Multiple inference techniques can be leveraged for sampling from this posterior $p(\boldsymbol{x}|\boldsymbol{y})$, but for high-dimensional problems modern techniques rely on gradient-based methods, including Variational Inference [Hoffman et al., 2013] and Langevin Diffusion, or HMC [Neal, 2011]. All of these techniques have in common that they only require having access to the *score* $\nabla_x \log p(x)$ of the target distribution. We illustrate in particular in Appendix B how Metropolis Hastings-calibrated HMC can be conducted with only knowledge of the

score.
Two terms will enter in the score of the posterior distribution in Equation 1, the score of the likelihood, and the score of the prior. In many problems, such as in the MRI problem presented later in this work, the likelihood score can be derived analytically, only the score of the prior remains unknown but can be learned from data by score matching.

Deep Denoising Score Matching As originally identified by Vincent [2011] and Alain and Bengio [2013], the score of a given target distribution P can be modeled using a DAE, i.e. by introducing an auto-encoding function $\boldsymbol{r} : \mathbb{R}^n \mapsto \mathbb{R}^n$ trained to reconstruct under an ℓ_2 loss a true $\boldsymbol{x} \sim P$ given a noisy version $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{n}$ with $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. An optimal denoiser $\boldsymbol{r}^\star$ would then be achieved for:

$$\boldsymbol{r}^\star(\boldsymbol{x}') = \boldsymbol{x}' + \sigma^2 \nabla_{\boldsymbol{x}} \log p_{\sigma^2}(\boldsymbol{x}') \tag{2}$$

where $p_{\sigma^2} = p * \mathcal{N}(0, \sigma^2)$. In other words, the optimal denoiser is closely related to the score we wish to learn and when the noise variance σ^2 tends to zero, should exactly match the score of the target density. In practice to learn this score efficiently, we adopt the residual noise-conditional denoising score matching technique proposed by Lim et al. [2020], and train a model to minimize the following Denoising Score Matching loss:

$$\mathcal{L}_{DSM} = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim P} \mathop{\mathbb{E}}_{\substack{\boldsymbol{u} \sim \mathcal{N}(0,I) \\ \sigma_s \sim \mathcal{N}(0, s^2)}} \parallel \boldsymbol{u} + \sigma_s \boldsymbol{r}_\theta(\boldsymbol{x} + \sigma_s \boldsymbol{u}, \sigma_s) \parallel_2^2 \tag{3}$$

In this formulation, the network $\boldsymbol{r}_\theta(\boldsymbol{x}, \sigma)$ is now directly modeling the score $\nabla_{\boldsymbol{x}} \log p_{\sigma^2}(\boldsymbol{x})$ of the Gaussian-convolved target distribution.

Annealed Hamiltonian-Monte Carlo Sampling Given the noise-conditional neural scores learned with the procedure described above, it is now possible to use a variety of inference methods to access the Bayesian posterior. In this work, we adopt an annealed HMC procedure which provides an efficient way to obtain parallel batches of independent samples from the target posterior despite the high dimensionality of the problem. This is a sampling procedure closely related to the Annealed Langevin Diffusion proposed in [Song and Ermon, 2019, 2020], but benefits from the faster Hamiltonian dynamics and Metropolis-Hastings calibration (see Appendix B).

To build our procedure, we consider a Gaussian-convolved version of our target density:

$$\log p_{\sigma^2}(\boldsymbol{x}|\boldsymbol{y}) = \log p_{\sigma^2}(\boldsymbol{y}|\boldsymbol{x}) + \log p_{\sigma^2}(\boldsymbol{x}) + cst \tag{4}$$

where σ^2 plays the role of the inverse temperature found in classical annealing. The likelihood can be obtained analytically, and in the case of a Gaussian likelihood takes the following form: $\log p_{\sigma^2}(\boldsymbol{y}|\boldsymbol{x}) = -\frac{\|\boldsymbol{x} - f(\boldsymbol{x})\|_2^2}{2(\sigma_n^2 + \sigma^2)} + cst$, where σ_n^2 is the noise variance in the measurements. As for the prior term, the noise-conditional score network introduced above already models the score of $\log p_{\sigma^2}$. This distribution is gradually annealed to low temperatures and the chain progressively moves towards a point in the target distribution.

4 Application to Bayesian Inverse Problems

The MRI Problem Magnetic Resonance Imaging (MRI) is a non-invasive modality used to probe soft tissues. Compressed sensing, introduced for MRI by Lustig et al. [2007] is used to reduce its significant acquisition time, and recently, deep learning approaches(see for example [Schlemper et al., 2018, Hammernik et al., 2018, Pezzotti et al., 2020]) have been shown to perform extremely well on the reconstruction problem. The idealized reconstruction problem is usually formalized in the following way, for the single coil setting, with uniform acquisition:

$$\boldsymbol{y} = M_\Omega \boldsymbol{F} \boldsymbol{x} + \boldsymbol{n} \tag{5}$$

where $\boldsymbol{y}$ is the acquired Fourier coefficients, also called the k-space data, M_Ω is a mask, $\boldsymbol{F}$ is the classical 2D Fourier transform, $\boldsymbol{x}$ is the real anatomical image, and $\boldsymbol{n} \sim \mathcal{N}(0, \sigma_n^2)$ is measurement noise (we set $\sigma_n = 0.1$ in our experiments). The data we consider is the single coil data with the Proton-Density (PD) contrast from the fastMRI dataset introduced in [Zbontar et al., 2018]. For comparison with deep learning approaches, we use the state-of-the-art Primal-Dual net enhanced with a U-net for image correction (UPDNet) from [Ramzi et al., 2020]. The under-sampling was done retrospectively using an acceleration factor of 4 and a random mask as described for knee images in [Zbontar et al., 2018].

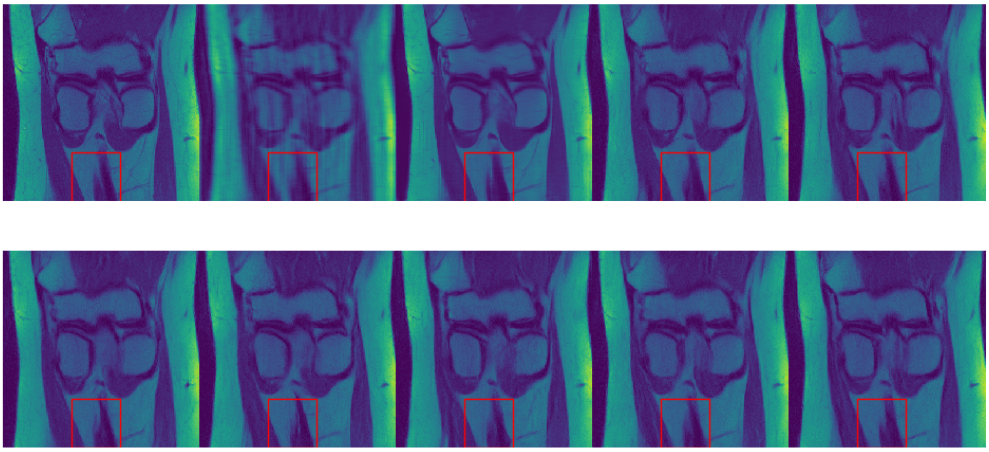

Figure 1: Bayesian posterior sampling for MRI reconstruction. The top leftmost image is the ground truth image. The top second to the left image is the zero-filled retrospectively under-sampled image $\boldsymbol{F}^T\boldsymbol{y}$. The top third to the left image is the reconstruction of the under-sampled image by the UPDNet. All the other images are denoised samples from the estimated posterior distribution obtained by a tempered HMC. The zero-filling achieves a PSNR of 25.55 dB, each sample 27.63 dB on average, the mean of the samples 30.04 dB and the neural network 32.15 dB. A zoom of the region in the red square is provided in appendix A.4. An animation of posterior samples is available online ▶.

Results We use a simple residual U-net inspired by Ronneberger et al. [2015] with ResNet building blocks from [He et al., 2016] as our score network. In order to promote the regularity of the learned scores, we regularize the spectral norm of each convolutional layer, and we find that setting the spectral norm to $\simeq 2$ yields the best results. More details about the network architecture and training can be found in appendix A. We then sample from the posterior following our annealed HMC procedure down to relatively low temperatures, and apply one final denoising step on the last sample from the chain, following the Expected Denoised Sample (EDS) scheme detailed in [Jolicoeur-Martineau et al., 2020]. Figure 1 compares samples from the MRI posterior to the ground truth (top left), zero-filled image (top second left), and UPDNet reconstruction (top center). We see that although individual samples carry slightly less details than the neural network reconstruction, confidence in any particular part of the image can be gauged by looking for variability or stability across multiple independent posterior samples. We highlight in particular the red region, where posterior samples show significant variability, indicating that this part of the image is poorly constrained by data. In contrast, a direct neural network reconstruction (top center) does not match the ground truth in that region, and does not provide an estimate of uncertainty which may lead a physician to misinterpret the image.

5 Conclusions and Discussions

We have presented in this work the first (to the best of our knowledge) instance of a framework for Bayesian inverse problems based on Deep Denoising Score Matching and applied to MRI reconstruction and Uncertainty Quantification. We illustrated the merits of this approach on an MRI example where the ability to sample from the full posterior highlights what features present in a reconstruction are not actually constrained by data.

We also showed that the Deep Denoising Score Matching approach can effortlessly be used for dimensions as high as $320 \times 320 \times 2$. It is therefore much more scalable than GAN- or VAE-based models. We note however that finding an optimal HMC annealing schedule and temperature-adaptive step-size proved difficult. In general we did not manage to lower the temperature below a certain level, prompting us to resort to a denoising step on the last samples from our HMC chains. This is a direction for further research.

References

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Mirza, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *Advances in neural information processing systems*, 2014. URL https://arxiv.org/pdf/1406.2661.pdf.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. pages 8107–8116, 2020. doi: 10.1109/cvpr42600.2020.00813.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, (Ml):1–14, 2014.

Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. pages 1–20, 7 2020. URL http://arxiv.org/abs/2007.03898.

Jae Hyun Lim, Aaron Courville, Christopher Pal, and Chin-Wei Huang. AR-DAE: Towards Unbiased Neural Entropy Gradient Estimation. 2020. URL http://arxiv.org/abs/2006.05164.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. (NeurIPS), 2019. URL http://arxiv.org/abs/1907.05600.

Yang Song and Stefano Ermon. Improved Techniques for Training Score-Based Generative Models. 2020. URL http://arxiv.org/abs/2006.09011.

Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6:695–708, 2005. URL http://jmlr.org/papers/volume6/hyvarinen05a/old.pdf.

Pascal Vincent. A connection between scorematching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. ISSN 08997667. doi: 10.1162/NECO{_}a{_}00142.

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data generating distribution. *1st International Conference on Learning Representations, ICLR 2013 - Conference Track Proceedings*, 15:3743–3773, 2013.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-Play priors for model based reconstruction. *2013 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings*, pages 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

Stephen Boyd. Alternating Direction Method of Multipliers. *IEEE Conference on Decision and Control*, 3:1–44, 2011. ISSN 1935-8237. doi: 10.1145/2783258.2783400.

Tim Meinhardt, Michael Moeller, and Daniel Cremers. Learning Proximal Operators : Using Denoising Networks for Regularizing Inverse Imaging Problems. (12):1781–1790, 2017.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *31st International Conference on Machine Learning, ICML 2014*, 4:3057–3070, 1 2014. URL http://arxiv.org/abs/1401.4082.

Ga Wu, Justin Domke, and Scott Sanner. Conditional Inference in Pre-trained Variational Autoencoders via Cross-coding. 5 2018. URL http://arxiv.org/abs/1805.07785.

Vanessa Böhm, François Lanusse, and Uroš Seljak. Uncertainty Quantification with Generative Models. (NeurIPS):1–9, 10 2019. URL http://arxiv.org/abs/1910.10046.

Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty Quantification in Deep MRI Reconstruction. pages 1–10, 2019. URL http://arxiv.org/abs/1901.11228.

Jonas Adler and Ozan Öktem. Deep Bayesian Inversion. 11 2018. URL http://arxiv.org/abs/1811.05910.

Alexander Denker, Maximilian Schmidt, Johannes Leuschner, Peter Maass, and Jens Behrmann. Conditional Normalizing Flows for Low-Dose Computed Tomography Image Reconstruction. 6 2020. URL http://arxiv.org/abs/2006.06270.

Francesco Tonolini, Jack Radford, Alex Turpin, Daniele Faccio, and Roderick Murray-Smith. Variational Inference for Computational Imaging Inverse Problems. 21:1–46, 4 2019. URL http://arxiv.org/abs/1904.06264.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013. ISSN 15324435.

Radford M. Neal. MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2011. doi: 10.1201/b10905-6.

Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging Michael. *Magnetic Resonance in Medicine*, 2007. ISSN 10014861. doi: 10.1002/mrm.21391.

Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony Price, and Daniel Rueckert. A Deep Cascade of Convolutional Neural Networks for MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018. doi: 10.1109/TMI.2017.2760978.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a Variational Network for Reconstruction of Accelerated MRI Data. *Magnetic Resonance in Medicine*, 3071(September 2017):3055–3071, 2018. doi: 10.1002/mrm.26977.

Nicola Pezzotti, Sahar Yousefi, Mohamed S Elmahdy, Jeroen Van Gemert, Christophe Schülke, Mariya Doneva, Tim Nielsen, Sergey Kastryulin, Boudewijn P F Lelieveldt, Matthias J P Van Osch, Elwin De Weerdt, and Marius Staring. An Adaptive Intelligence Algorithm for Undersampled Knee MRI Reconstruction: Application to the 2019 fastMRI Challenge. Technical report, 2020.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C Lawrence Zitnick, Michael P Recht, Daniel K Sodickson, and Yvonne W Lui. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. Technical report, 2018. URL https://arxiv.org/pdf/1811.08839.pdf.

Zaccharie Ramzi, Philippe Ciuciu, and Jean Luc Starck. Benchmarking MRI reconstruction neural networks on large public datasets. *Applied Sciences (Switzerland)*, 10(5), 2020. ISSN 20763417. doi: 10.3390/app10051816.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Technical report, 2015. URL http://lmb.informatik.uni-freiburg.de/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.90.

Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. pages 1–29, 2020. URL http://arxiv.org/abs/2009.05475.

Yuichi Yoshida and Takeru Miyato. Spectral Norm Regularization for Improving the Generalizability of Deep Learning. 2017. URL http://arxiv.org/abs/1705.10941.

Ernest K. Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:9764–9775, 2019.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. ISSN 00219606. doi: 10.1063/1.1699114.

W. K. Hastings. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444. doi: 10.1093/biomet/57.1.97.

A MRI experiments

A.1 Network architecture

The network is a 3-scale U-net with residual blocks composed of 3 convolutions followed by a batch normalisation. Each batch normalisation is followed by a ReLU non-linearity except the last one. A projection is used for the input of residual blocks whose number of input channels is not the same as that of output channels. Each scale uses 2 residual blocks for the downsampling path and 2 residual blocks for the upsampling path. Downsampling is performed via average pooling and upsampling is performed via up-convolution (as designed in [Ronneberger et al., 2015]), in order to avoid checkerboard artefacts. We use the following sequence of number of channels: [32, 64, 128]. In order to deal with MRI images, that are complex in nature, we concatenate the real and imaginary part of the image at the input of the network, forming effectively a 2-channel 320×320 image. For input noise level conditioning, we concatenate a noise standard deviation map to the input and in the lower scale of the network. Following the recommendation of Song and Ermon [2020], we also divide the output of the network by the absolute value of the noise power. Finally, we use Spectral Norm regularisation from [Yoshida and Miyato, 2017] in order to make sure the score does not take inconsistent values in low-density regions. The Spectral Norm regularisation indeed forces the spectral norm (maximum eigenvalue) of each layer to a certain value which in turn lowers the Lipschitz constant of the network, preventing it to blow in unseen regions. The influence of the Spectral Norm has been studied in the case of Plug-and-Play approaches by Ryu et al. [2019].

A.2 Network training

We use the Adam optimizer for network training, with a learning rate of 10^{-4}. We add a white Gaussian noise of variance σ_s^2 to the images scaled by a factor of 10^6. σ_s is drawn on-the-fly from a Gaussian distribution of variance $s = 50$. This means that at training time, the standard deviation of the noise can be negative, following the recommendation of Lim et al. [2020] to go from extrapolation to interpolation.

A.3 HMC procedure

The HMC procedure starts from a zero-filled reconstruction of the image with added white Gaussian noise of variance $\sigma_{init} = 100$. For the reduction of noise standard deviation at reduction step, we use a factor $\gamma = 0.995$, following the recommendation of Song and Ermon [2020]. We take a step size α dependent of the sampling temperature at step i, $\alpha = \epsilon \left(\frac{\sigma_i}{\sigma_0}\right)^{1.5}$ with $\epsilon = 10$.

A.4 Zoom on Reconstruction

The zoom on this part of the reconstruction shown in Figure 2, illustrates a failure of the Neural network reconstruction. It indeed produces a sharp artifact that might hinder the examiner's conclusions. The bayesian sampling allows us to see that this region is not necessarily very sharp.

B Score-Based Hamiltonian Monte Carlo

We used HMC for sampling from the posterior distribution. The sampling procedure follows the leapfrog integrator presented bellow:

$$\begin{aligned} \boldsymbol{m}_{t+\frac{\alpha}{2}} &= \boldsymbol{m}_t + \frac{\alpha}{2}\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t|\boldsymbol{y}) \\ \boldsymbol{x}_{t+\alpha} &= \boldsymbol{x}_t + \alpha \mathbf{M}^{-1}\boldsymbol{m}_{t+\frac{\alpha}{2}} \\ \boldsymbol{m}_{t+\alpha} &= \boldsymbol{m}_{t+\frac{\alpha}{2}} + \frac{\alpha}{2}\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_{t+\alpha}|\boldsymbol{y}) \end{aligned} \tag{6}$$

where α is the step size, $\boldsymbol{m}$ the momentum and $\mathbf{M}$ is a preconditioning matrix taking into account the space metric, but in our case the identity matrix. The gradient of the log posterior is computed as $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t|\boldsymbol{y}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}_t) + \nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}_t)$, using the inverse Bayes theorem in equation 1.

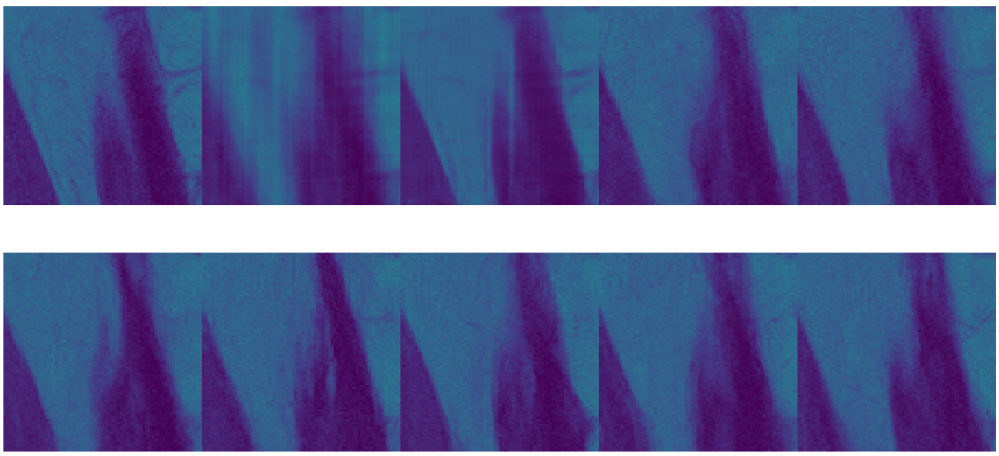

Figure 2: Zoom of Bayesian posterior sampling for MRI reconstruction. The ordering is the same as in Figure 1.

It is worth noting that the discretized HMC with final step size ϵ is not guaranteed to converge to the desired target distribution. This is the motivation for the Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970], which corrects for the discretization errors with a correction step. Concretely, a new step is proposed using Equation 6, but it is accepted with probability of the form $\min\{1, p(\boldsymbol{x}^*)q(\boldsymbol{x}_n|\boldsymbol{x}^*)/p(\boldsymbol{x}_n)q(\boldsymbol{x}^*|\boldsymbol{x}_n)\}$ where $\boldsymbol{x}_n$ is the last accepted point in the chain, $\boldsymbol{x}^*$ is the proposed update, p is the target density, and q is the proposal corresponding to taking a step of Equation 6. In the context of a score-only approach, we do not have access to p and therefore cannot compute directly this acceptance probability. We can nonetheless estimate from the scores themselves by using a path integral:

$$\log p(\boldsymbol{x}_*) - \log p(\boldsymbol{x}_n) = \int_0^1 \nabla_{\boldsymbol{x}} \log p(t * (\boldsymbol{x}_* - \boldsymbol{x}_n) + \boldsymbol{x}_n) \cdot (\boldsymbol{x}_* - \boldsymbol{x}_n) dt \tag{7}$$

This integral can be evaluated numerically to arbitrary precision (at the cost of additional score evaluations), but we find excellent agreement with the theoretical value on toy model densities like the two-moons distribution with a simple 4 points Simpson integration rule.

This illustrates that for gradient-based MCMC samplers, which also includes Metropolis Adjusted Langevin Algorithm, these algorithms can fully be implemented with only knowledge of the score.