Efficient Bayesian Computational Imaging with a Surrogate Score-Based Prior

Berthy T. Feng **Katherine L. Bouman**
bfeng, klbouman @ caltech.edu
California Institute of Technology

Abstract

We propose a surrogate function for efficient use of score-based priors for Bayesian inverse imaging. Recent work turned score-based diffusion models into probabilistic priors for solving ill-posed imaging problems by appealing to an ODE-based log-probability function. However, evaluating this function is computationally inefficient and inhibits posterior estimation of high-dimensional images. Our proposed surrogate prior is based on the evidence lower-bound of a score-based diffusion model. We demonstrate the surrogate prior on variational inference for efficient approximate posterior sampling of large images. Compared to the exact prior in previous work, our surrogate prior accelerates optimization of the variational image distribution by at least two orders of magnitude. We also find that our principled approach achieves higher-fidelity images than non-Bayesian baselines that involve hyperparameter-tuning at inference. Our work establishes a practical path forward for using score-based diffusion models as general-purpose priors for imaging.

1 Introduction

Ill-posed image reconstruction requires a prior to enforce desired image statistics. From a Bayesian perspective, the prior influences the uncertainty and richness of the estimated image. Diffusion models represent rich image priors, but leveraging these priors for Bayesian image reconstruction remains a challenge. Recent work demonstrated how to turn score-based diffusion models into probabilistic priors (*score-based priors*) for Bayesian imaging [8]. However, it involves solving an ordinary differential equation (ODE) for every probability computation, requiring days to a week to reconstruct even a 32×32 image [8]. We present a method for Bayesian inference with a score-based prior that is both principled and computationally efficient.

We propose leveraging the evidence lower-bound of a score-based diffusion model [18; 10] as an efficient surrogate for the exact log-probability function. This function can be plugged into any inference algorithm that requires the value or gradient of the posterior log-density. When it is used in variational inference of an image posterior, we find at least two orders of magnitude in speedup of optimizing the variational distribution. Our time- and memory-efficiency improvements make it practical to perform inference with score-based priors.

In this paper, we describe our variational-inference approach to efficiently estimate a posterior with a surrogate score-based prior. We provide experimental results to validate the proposed surrogate prior, including high-dimensional posterior samples of sizes up to 256×256, a resolution infeasible with the exact prior. In the setting of accelerated MRI, we quantify time- and memory-efficiency improvements of the surrogate over the exact prior. We also demonstrate how our proposed approach achieves higher-quality image reconstructions than methods that deviate from true Bayesian inference.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

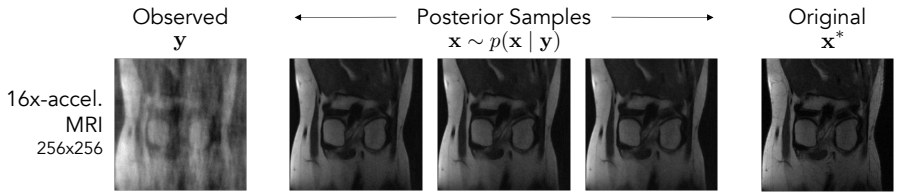


Figure 1: High-dimensional Bayesian inference with a surrogate score-based prior. We propose a surrogate prior for efficient use of score-based diffusion models as priors for Bayesian imaging. Here we show posterior samples for $16\times$-accelerated MRI of 256×256 knee images, approximated via variational inference with a surrogate score-based prior. Bayesian imaging at this image resolution is computationally infeasible with the previous ODE-based approach [8].

2 Background

2.1 Bayesian inverse imaging

Image reconstruction can be framed as an inverse problem in which a hidden image $\mathbf{x}^* \in \mathbb{R}^D$ must be recovered from measurements $\mathbf{y} \in \mathbb{R}^M$, where $\mathbf{y} = f(\mathbf{x}^*) + \epsilon$. For an ill-posed inverse problem, Bayesian imaging considers a posterior of possible images $\mathbf{x}$ given $\mathbf{y}$ whose log-density is

$$\log p(\mathbf{x} \mid \mathbf{y}) = \log p(\mathbf{y} \mid \mathbf{x}) + \log p(\mathbf{x}) + \text{const.} \tag{1}$$

Given a log-likelihood function $\log p(\mathbf{y} \mid \mathbf{x})$ and a prior log-probability function $\log p(\mathbf{x})$, we can use established techniques for sampling from the posterior, including variational inference (VI) [2].

2.2 Score-based diffusion models

Diffusion models for inverse imaging. Diffusion models [14; 10; 16; 17; 19] learn a rich image distribution that would be useful as a prior for inverse problems. Previous methods incorporate this prior in a non-Bayesian way, such as by projecting images onto a measurement subspace [20; 6; 4; 3; 5] or by following a gradient toward higher measurement likelihood [7; 11; 9; 12; 1; 15; 13]. Such methods require hyperparameter-tuning at inference and may not accurately sample the posterior.

Generative distribution. A score-based diffusion model transforms the simple distribution $\pi = \mathcal{N}(\mathbf{0}, \mathbf{I})$ into a complex image distribution through gradual denoising. This process, known as "reverse diffusion," is governed by a reverse-time stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \mathbf{s}_\theta(\mathbf{x}, t)\right] \mathrm{d}t + g(t) \mathrm{d}\bar{\mathbf{w}}, \quad t \in [0, T]. \tag{2}$$

Here $\bar{\mathbf{w}} \in \mathbb{R}^D$ denotes Brownian motion. $g(t) \in \mathbb{R}$ and $\mathbf{f}(\cdot, t) : \mathbb{R}^D \to \mathbb{R}^D$ are the diffusion and drift coefficients, respectively, and arise from a pre-defined forward-time diffusion process. $\mathbf{s}_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the *score model* learned by a neural network. Image generation is done by solving the reverse-time SDE starting with $\mathbf{x}(T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to get a clean image $\mathbf{x}(0) \sim p_\theta^{\text{SDE}}$. For any image $\mathbf{x}$, evaluating $p_\theta^{\text{SDE}}(\mathbf{x})$ is not tractable, and solving an ordinary differential equation (ODE) to evaluate $p_\theta^{\text{ODE}} \approx p_\theta^{\text{SDE}}$ [19] is computationally inefficient [8].

Evidence lower bound. Song et al. [18] derived an evidence lower-bound for p_θ^{SDE} such that $b_\theta^{\text{SDE}}(\mathbf{x}) \leq \log p_\theta^{\text{SDE}}(\mathbf{x})$ for any proposed image $\mathbf{x}$. Essentially, this lower-bound corresponds to how well the diffusion model is able to denoise a given image: an image with high probability under the diffusion model is easy to denoise, whereas a low-probability image is difficult. The lower-bound, or the negative "denoising score-matching loss" [18], is defined as

$$b_\theta^{\text{SDE}}(\mathbf{x}) := \mathbb{E}_{p_{0T}(\mathbf{x}' \mid \mathbf{x})}\left[\log \pi(\mathbf{x}')\right] - \frac{1}{2}\int_0^T g(t)^2 h(t) \mathrm{d}t, \quad \text{where} \tag{3}$$

$$h(t) := \mathbb{E}_{p_{0t}(\mathbf{x}' \mid \mathbf{x})}\left[\left\|\mathbf{s}_\theta(\mathbf{x}', t) - \nabla_{\mathbf{x}'} \log p_{0t}(\mathbf{x}' \mid \mathbf{x})\right\|_2^2 - \left\|\nabla_{\mathbf{x}'} \log p_{0t}(\mathbf{x}' \mid \mathbf{x})\right\|_2^2 - \frac{2}{g(t)^2} \nabla_{\mathbf{x}'} \cdot \mathbf{f}(\mathbf{x}', t)\right].$$

$p_{0t}(\mathbf{x}' \mid \mathbf{x})$ denotes the transition distribution from $\mathbf{x}(0) = \mathbf{x}$ to $\mathbf{x}(t) = \mathbf{x}'$. For a drift coefficient that is linear in $\mathbf{x}$, this transition distribution is Gaussian: $p_{0t}(\mathbf{x}' \mid \mathbf{x}) = \mathcal{N}(\mathbf{x}'; \alpha(t)\mathbf{x}, \beta(t)^2 \mathbf{I})$. This means that the gradient $\nabla_{\mathbf{x}'} \log p_{0t}(\mathbf{x}' \mid \mathbf{x})$ is directly proportional to the Gaussian noise that is subtracted from $\mathbf{x}'$ to get $\mathbf{x}$. In fact, Eq. 3 is closely related to the denoising score-matching objective used to efficiently train diffusion models [19].

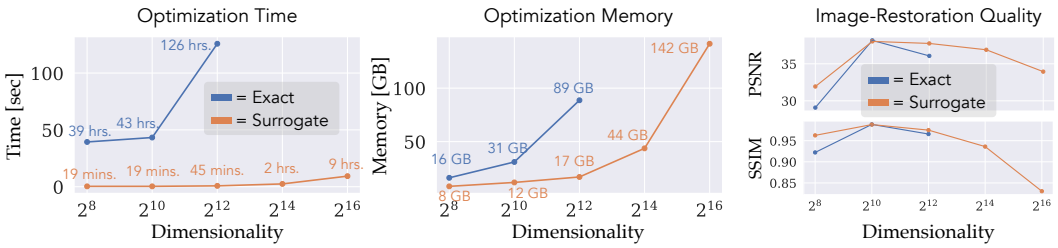


Figure 2: Efficiency of proposed surrogate vs. exact prior. For each image size, we estimated a posterior of images for $4\times$-accelerated MRI of a knee image, using a Gaussian variational distribution with diagonal covariance. For image sizes supported by the exact prior, the surrogate improved total optimization time by over $120\times$ while using less memory and scaling better with image size. "Image-Restoration Quality" verifies that faster optimization did not hurt the quality of samples.

3 Method

Given measurements $\mathbf{y} \in \mathbb{R}^M$ (with a known log-likelihood function) and a score-based diffusion model with parameters θ as the prior, our goal is to sample from the image posterior $p_\theta(\mathbf{x} \mid \mathbf{y})$. Following VI, we optimize the parameters of a variational distribution to approximate $p_\theta(\mathbf{x} \mid \mathbf{y})$.

Let q_ϕ denote the variational distribution with parameters ϕ, and we assume q_ϕ to have tractable log-probabilities. We wish to minimize the KL divergence from q_ϕ to the target posterior:

$$\phi^* = \arg\min_\phi D_{\mathrm{KL}}(q_\phi \| p_\theta(\cdot \mid \mathbf{y})) = \arg\min_\phi \mathbb{E}_{\mathbf{x}\sim q_\phi}\left[-\log p(\mathbf{y} \mid \mathbf{x}) - \log p_\theta^{\mathrm{SDE}}(\mathbf{x}) + \log q_\phi(\mathbf{x})\right]. \quad (4)$$

q_ϕ can be various types of distributions. It could be a Gaussian distribution with a diagonal covariance so that $\phi := [\mu^\top, \sigma^\top]^\top$, where $\mu \in \mathbb{R}^D$ and $\sigma \in \mathbb{R}^D$ ($\sigma > \mathbf{0}$) are the mean and pixel-wise standard deviation. As DPI showed [21], q_ϕ could also be a RealNVP normalizing flow with parameters ϕ.

To circumvent the intractability of the prior term $\log p_\theta^{\mathrm{SDE}}(\mathbf{x})$, we replace it with the surrogate $b_\theta^{\mathrm{SDE}}(\mathbf{x})$. This results in the following objective:

$$\phi^* = \arg\min_\phi \mathbb{E}_{\mathbf{x}\sim q_\phi}\left[-\log p(\mathbf{y} \mid \mathbf{x}) - b_\theta^{\mathrm{SDE}}(\mathbf{x}) + \log q_\phi(\mathbf{x})\right]. \quad (5)$$

We can also think of b_θ^{SDE} as replacing the intractable $\log p_\theta^{\mathrm{SDE}}$ in Eq. 4. Since $-\log p_\theta^{\mathrm{SDE}} \leq -b_\theta^{\mathrm{SDE}}$, our surrogate objective minimizes the upper-bound of a valid KL divergence involving p_θ^{SDE}.

Implementation details. The $b_\theta^{\mathrm{SDE}}(\mathbf{x})$ formula (Eq. 3) contains a time integral and expectation that can be estimated with numerical methods. Following Song et al. [18], we use importance sampling with N_t time samples $t \sim p(t)$ for the time integral and Monte-Carlo approximation with $N_{\mathbf{z}}$ noisy images $\mathbf{x}' \sim \mathcal{N}(\alpha(t)\mathbf{x}, \beta(t)^2\mathbf{I})$ for the expectation. In our experiments, we set $N_t = N_{\mathbf{z}} = 1$.

4 Experiments

In this section, we validate the efficiency improvements over [8], and we compare to diffusion-based inference methods. Please refer to the Appendices A and B for details about the experiment setups.

4.1 Efficiency improvements

In Fig. 2, we quantify the efficiency improvements of the surrogate prior for an accelerated MRI task at different image resolutions. We drew a test image from the fastMRI knee dataset [23] and resized it to 16×16, 32×32, 64×64, 128×128, and 256×256. For each image size, we trained a score model on training images of the corresponding size from the fastMRI dataset of single-coil knee scans. We then optimized a Gaussian distribution with diagonal covariance to approximate the posterior. We find at least two orders of magnitude in time improvement with the surrogate prior.

4.2 Bayesian approach vs. diffusion-based approaches

Being grounded in Bayesian inference helps us obtain a more accurate posterior *and* images that more accurately reflect the ground-truth image than the diffusion-based approaches mentioned in

Sec. 2.2. We compare to three diffusion-based baselines: **SDE+Proj** [20], **Score-ALD** [11], and Diffusion Posterior Sampling (**DPS**) [7]. All baselines involve measurement-weight hyperparameters. Our approach is variational inference with the surrogate prior and a RealNVP variational distribution.

4.2.1 Accuracy of posterior

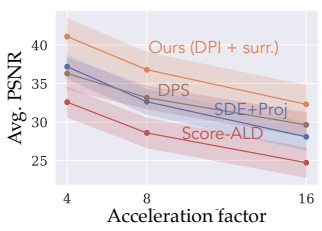
We tested how well each method could recover a simple bimodal 2D posterior. The prior is a bimodal mixture-of-Gaussians and the forward model a linear projection with Gaussian noise, making the posterior a known bimodal mixture-of-Gaussians. Each method was given the true score function of the prior. We considered a reasonable search space of hyperparameters for the diffusion-based baselines, but none correctly recovered the bimodal posterior. As shown in Tab. 1, even the best KL divergence obtained by the diffusion-based baselines does not rival that of VI. Hyperparameter values giving the "best" KL divergence for baselines can only be found with knowledge of the ground-truth, whereas our method automatically finds a better KL divergence by following the Bayesian posterior formula.

	KL (↓)	time/step (↓)
DPI + exact	**0.030**	130 ms
Ours: DPI + surr.	0.037	**22 ms**
DPS (oracle)	0.064	
Score-ALD (oracle)	0.10	
SDE+Proj (oracle)	0.12	

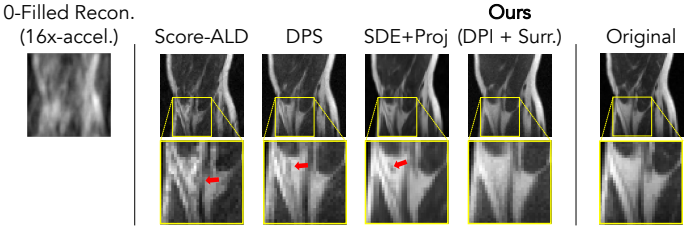
Table 1: Quantitative comparison of estimated posteriors. A two-component Gaussian mixture model was fit to estimated samples. "Ours" achieves much lower KL div. (i.e., forward KL from estimated to true bimodal posterior) than diffusion-based baselines at their best. Our surrogate is more efficient than the exact score-based prior without sacrificing much accuracy.

4.2.2 Image-reconstruction quality

We find that our approach achieves higher-fidelity reconstructions in addition to more-reliable uncertainty. We performed multiple MRI tasks at three acceleration rates and compared our approach to the diffusion-based baselines. The score model $\mathbf{s}_\theta$ was trained on 64×64 fastMRI knee images and stayed fixed across all methods. Our method achieves a marked improvement in PSNR and SSIM over the three baselines (Fig. 3), improving PSNR by 2.7 to 8.5 dB.



(a) Image-restoration metrics.



(b) Example image reconstructions for $16\times$ acceleration.

Figure 3: Accelerated MRI of knee images. **(a)** For each accel. factor ($4\times$, $8\times$, $16\times$), we estimated posteriors for ten images. For each method, we computed the average PSNR and SSIM of 128 estimated posterior samples (line plot shows average result across the ten tasks; shaded region shows one std. dev. above and below average). **(b)** An example of $16\times$-accel. MRI. The cropped region exemplifies how diffusion-based baselines hallucinate more features than necessary.

5 Conclusion

We have presented a surrogate function that provides efficient access to score-based priors for Bayesian inference. Specifically, the evidence lower-bound $b_\theta^{\text{SDE}}(\mathbf{x}) \leq \log p_\theta^{\text{SDE}}(\mathbf{x})$ serves as a proxy for the log-prior of an image in the Bayesian log-posterior. Our experiments with variational inference show at least two orders of magnitude in runtime improvement and significant memory improvement over the ODE-based prior. This enables inference of images previously too large for a strictly Bayesian approach, such as 256×256 pixels. We also establish that a principled approach like ours outperforms baselines on posterior approximation and image restoration, evidence that following a Bayesian approach results in more-reliable image reconstructions.

References

[1] Alexandre Adam, Adam Coogan, Nikolay Malkin, Ronan Legin, Laurence Perreault-Levasseur, Yashar Hezaveh, and Yoshua Bengio. Posterior samples of source galaxies in strong gravitational lenses with score-based priors. *arXiv preprint arXiv:2211.03812*, 2022. 2

[2] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. 2

[3] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *ICCV*. IEEE, 2021. 2

[4] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical Image Analysis*, 80:102479, 2022. 2

[5] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022. 2

[6] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022. 2

[7] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k. 2, 4

[8] Berthy T Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L Bouman, and William T Freeman. Score-based diffusion models as principled priors for inverse imaging. In *International Conference on Computer Vision (ICCV)*. IEEE, 2023. 1, 2, 3

[9] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. URL https://arxiv.org/pdf/2206.09012.pdf. 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[11] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jonathan I Tamir. Robust compressed sensing mri with deep generative priors. *NeurIPS*, 2021. 2, 4

[12] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. 2

[13] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023. 2

[14] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[15] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ. 2

[16] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, pages 11895–11907, 2019. 2

[17] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI*, page 204, 2019. URL http://auai.org/uai2019/proceedings/papers/204.pdf. 2

[18] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2, 3

[19] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS. 2

[20] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *ICLR*, 2022. URL https://openreview.net/forum?id=vaRCHVj0uGI. 2, 4, 8

[21] He Sun and Katherine L Bouman. Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging. In *AAAI*, pages 2628–2637, 2021. 3, 7

[22] Muhammad Usman and Philipp G Batchelor. Optimized sampling patterns for practical compressed mri. In *SAMPTA'09*, pages Poster–session, 2009. 7

[23] 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. 3, 8

A Accelerated MRI

In this section, we describe the forward model of accelerated MRI that was used in our experiments. Accelerated MRI collects sparse spatial-frequency measurements in κ-space of an underlying anatomical image. As the acceleration rate increases, the number of measurements decreases. In accelerated MRI, the forward model can be written as

$$\mathbf{y} = \mathbf{M} \odot \mathcal{F}(\mathbf{x}^*) + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}^2 \mathbf{I}), \tag{6}$$

where $\mathbf{x} \in \mathbb{C}^D$ and $\mathbf{y} \in \mathbb{C}^M$. $\mathcal{F}$ denotes the 2D Fourier transform, and $\mathbf{M} \in \{0,1\}^D$ is a binary sampling mask that reduces the number of non-zero measurements to $M << D$. Often $\sigma_{\mathbf{y}}$ is assumed to be small (e.g., corresponding to an SNR of at least 30 dB). We use Poisson-disc sampling [22] to obtain a sampling mask. $16\times$-acceleration, for example, corresponds to a sampling mask with only $1/16$ nonzero elements.

Experimental setup. In our experiments, we assumed that $|\sigma_{\mathbf{y}}|$ is 0.05% of the DC (zero-frequency) amplitude. This corresponds to a maximum SNR of 40 dB. The only exception is for comparison to baselines (Fig. 3), since baseline methods do not account for measurement noise. In this case, we let $|\sigma_{\mathbf{y}}| = 0.1\%$ of the DC amplitude along the horizontal direction of the true image, which amounts to a very low level of noise.

B Experiment details

For the sake of reproducibility, we detail the experimental setup behind each figure. Some common implementation details are that the exact prior ($\log p_\theta^{\text{ODE}}$) was always estimated with 16 trace estimators. The RealNVP variational distribution had 32 affine-coupling layers unless stated otherwise.

B.1 Variational distributions

RealNVP. The architecture is determined by the number of affine-coupling layers and the width of each layer. For images up to 64×64, we use 32 affine-coupling layers and set the number of hidden neurons in the first layer to $1/8$ of the image dimensionality (e.g., $32 \cdot 32 \cdot 3/8$ for 32×32 RGB images). We use batch normalization in the network. Our implementation is a JAX-based adaptation of the original DPI [21] PyTorch implementation.[1]

Gaussian. Other experiments use a multivariate Gaussian distribution with a diagonal covariance matrix as the variational distribution. The parameters are the mean image and the pixel-wise standard deviation. We initialize the mean at 0.5 and the standard deviation at 0.1 for all pixels. To sample, we take the absolute value of the standard deviation and construct the diagonal covariance matrix.

B.2 MRI efficiency experiment (Fig. 2)

Score model. For each image size, the score model was an NCSN++ architecture with 64 filters in the first layer and trained with the VP SDE with $\beta_{\text{min}} = 0.1$ and $\beta_{\text{max}} = 10$.

Variational optimization. For each task (i.e., each image size and prior), the variational distribution was a multivariate Gaussian with diagonal covariance. The batch size was 64, learning rate 0.0002, and gradient clip 1. A convergence criterion based on the loss value is difficult to define due to high variance of the loss (we used 1 time sample to estimate $b_\theta(\mathbf{x})$). We defined a convergence criterion based on the change in the mean of the variational distribution. Specifically, every 10000 steps, we evaluated a snapshot of the variational Gaussian and computed $\delta = \|\mu_{\text{curr}} - \mu_{\text{prev}}\| / \|\mu_{\text{prev}}\|$, where μ_{curr} and μ_{prev} are the current and previous snapshot means, respectively. If $\delta < \varepsilon$ for some threshold ε two snapshots in a row, then the optimization was considered converged. Since convergence rate depends on the image size and the prior used, we set a different ε for each task:

- 16×16 (surrogate): $\varepsilon = 0.002$
- 32×32 (surrogate): $\varepsilon = 0.003$
- 64×64 (surrogate): $\varepsilon = 0.005$

[1] https://github.com/HeSunPU/DPI

- 128×128 (surrogate): $\varepsilon = 0.007$
- 256×256 (surrogate): $\varepsilon = 0.009$
- 16×16 (exact): $\varepsilon = 0.0025$
- 32×32 (exact): $\varepsilon = 0.0027$
- 64×64 (exact): $\varepsilon = 0.005$

We were conservative in defining the convergence and checked that optimization under the surrogate actually achieved better sample quality than optimization under the exact prior (see Fig. 2).

Data. The test image is from the fastMRI [23] single-coil knee test dataset and was resized to 64×64 with antialiasing.

B.3 256x256 MRI example (Fig. 1)

The $4\times$-acceleration result is from the efficiency experiment (Fig. 2) on the 256×256 test image. The $16\times$-acceleration result came from a similar setup, where the variational distribution was Gaussian with diagonal covariance. Optimization was done with a batch size of 64, learning rate of 0.00001, and gradient clip of 0.0002. We ran optimization for 270K steps (optimization for $4\times$-acceleration was done in 100K steps with the convergence criterion).

B.4 Accuracy of posterior (Tab. 1)

Variational optimization. For both the exact score-based prior and the surrogate score-based prior, the variational distribution was a RealNVP with 16 affine-coupling layers, and it was optimized for 12000 iterations with a batch size of 2560 and learning rate of 10^{-5}. For the surrogate score-based prior, the lower-bound was approximated with $N_t = N_z = 1$ (i.e., 1 time sample and 1 noise sample).

Baselines. For this simple 2D experiment, we implemented the diffusion-based baselines exactly according to their proposed algorithms. For SDE+Proj, we tested the following values for the measurement weight λ: `linspace(0.001, 0.5, num=100)`. For Score-ALD, we distilled all hyperparameters into one global hyperparameter $1/\gamma_T$ and tested the following values for γ_T: `linspace(100, 0.8, num=100)`. For DPS, we tested the following values for the scale parameter ζ: `exp(linspace(log(0.001), log(0.15), num=100))`.

Evaluation. Since the diffusion-based approaches only provide samples (not probability densities), we approximated the probability density function (PDF) from the estimated posterior samples. For each method, we fit a two-component Gaussian mixture model (GMM) to 10000 samples. The forward KL divergence was approximated with the log-density function of the fitted GMM and the log-density function of the true posterior, evaluated on these 10000 samples.

B.5 Image-restoration metrics (Fig. 3)

Score model. The score model is the same as the one used for the 64×64 image in the MRI efficiency experiment (Fig. 2).

Variational optimization. The variational distribution was a RealNVP. Optimization was done with a learning rate of 0.00001 and gradient clip of 0.0002. We used the same convergence criterion as the one used in the MRI efficiency experiment with $\varepsilon = 0.005$.

Baseline hyperparameters. For SDE+Proj, we used the `projection` CS solver provided by Song et al. [20] with the hyperparameters `snr=0.517, coeff=1`. For Score-ALD, we used the `langevin` CS solver with the hyperparameters `n_steps_each=3, snr=0.212, projection_sigma_rate=0.713`. For DPS, we used `scale=0.5`. This was the best scale out of $[10, 1, 0.9, 0.5, 0.3, 0.1, 0.001]$ for a test image in terms of PSNR with respect to the true image.