Solving Inverse Problems with Score-Based Generative Priors learned from Noisy Data

Asad Aali[1], Marius Arvinte[1], Sidharth Kumar[1], Jonathan I. Tamir[1]

[1]*Chandra Family Department of Electrical and Computer Engineering, The University of Texas at Austin*, Austin, TX, USA
Email: {asad.aali, arvinte, sidharth.kumar, jtamir}@utexas.edu

Abstract—We present SURE-Score: an approach for learning score-based generative models using training samples corrupted by additive Gaussian noise. When a large training set of clean samples is available, solving inverse problems via score-based (diffusion) generative models trained on the underlying fully-sampled data distribution has recently been shown to outperform end-to-end supervised deep learning. In practice, such a large collection of training data may be prohibitively expensive to acquire in the first place. In this work, we present an approach for approximately learning a score-based generative model of the clean distribution, from noisy training data. We formulate and justify a novel loss function that leverages Stein's unbiased risk estimate to jointly denoise the data and learn the score function via denoising score matching, while using only the noisy samples. We demonstrate the generality of SURE-Score by learning priors and applying posterior sampling to ill-posed inverse problems in two practical applications from different domains: compressive wireless multiple-input multiple-output channel estimation and accelerated 2D multi-coil magnetic resonance imaging reconstruction, where we demonstrate competitive reconstruction performance when learning at signal-to-noise ratio values of 0 and 10 dB, respectively.

Index Terms—Score, Diffusion, Generative, MIMO, MRI, SURE, Inverse Problems

I. INTRODUCTION

Recent advances in score-based (diffusion) generative modeling [1], [2] have helped substantially improve the capabilities of solving ill-posed imaging inverse problems using fewer measurements and with higher reconstruction fidelity in various domains such as medical imaging [3]–[6], digital communications [7], [8], image super-resolution [9], and more.

However, learning high-quality score-based generative models for distributions over real-world signals currently assumes a large database of fully-sampled and noise-free training samples are available [10]. In many application domains, acquiring such a training set is impossible in practice, because noise is inherently present in the sensors used to acquire measurements. For example, noise is always present in single or multi-coil magnetic resonance imaging (MRI) due to thermal noise in the radio-frequency (RF) hardware as well as losses due to the body. Similarly, noise is present in communication transceivers, especially in wireless scenarios where mobile users operate devices under cost and energy constraints. Learning score-based generative models using only noisy samples is

thus an important research problem. In this work, we propose a solution to this problem by combining denoising using Stein's unbiased risk estimate (SURE) [11] with the denoising score matching (DSM) objective used in diffusion model training [12], to achieve joint denoising and score learning with a single deep learning model.

Our framework connects the two already-similar objectives by leveraging Tweedie's single-step denoising formula [13] at a properly chosen noise level. We show that this objective is naturally compatible with the DSM formulation and allows for a simple modification which introduces two additional terms and a scalar weight in the training function, in the case of training data corrupted by independent and identically distributed (i.i.d.) Gaussian noise. We evaluate the effectiveness of our proposed approach by using the learned models as priors in two different applications: compressive multiple-input multiple-output (MIMO) channel estimation and accelerated MRI reconstruction. Our results show that the priors learned using the proposed formulation can be reliably used for solving linear inverse problems, even when the training dataset is captured in a poor signal-to-noise ratio (SNR) of 0 dB.

A. Related Work

There is an extensive body of prior work related to denoising of structured signals; see [14] for a recent review paper. Most relevant to our work is the use of the self-supervised SURE objective [11], which introduces an unbiased estimate of the supervised denoising loss purely using noisy data. Using the SURE loss as a training objective for deep neural network denoisers has been previously investigated in a standalone denoising setting [15]–[17], and the connection to multi-level denoising score matching has only been recently explored in [18], where learning the score function on a restricted subset of noise levels is addressed.

Several other recent methods have used self-supervised learning for blind denoising without ground truth training data. One example is Noise2Noise [19], where multiple noisy measurements of the same sample are used to train a deep neural denoiser. The work in [20] further introduces the Noise2Self framework, which requires only a single noisy measurement of each sample in the training set, similar to our proposed method. The recent work in [17] introduces the Noise2Score framework for self-supervised learning of the score function in moderate and high SNR settings, which leaves the problem of learning score functions in low SNR

settings open. More recently, the work in [21] has investigated learning a score-based generative model from noisy diffusion MRI samples, still requiring access to multiple noisy scans of the same subject.

Pre-trained generative priors have also been used in solving linear inverse problems, surpassing classical compressed sensing approaches [22], [23]. Recently, a series of works have shown that score-based generative models produce competitive reconstruction performance when used to solve accelerated MRI [3]–[6] in a supervised learning setting, and that the score can be learned implicitly from MMSE denoisers [24]. The work in [7] applies the same ideas to learning priors for MIMO channel estimation from a limited number of pilot measurements. The recent work of [25] explores self-supervised learning of score models for MRI reconstruction by splitting the measurement data into distinct subsets, similar in approach to Noise2Self.

B. Contributions

Our contributions in this work are the following:

- We formulate SURE-Score: a framework for jointly learning a denoiser and score-based generative model using the same deep neural network. The loss connects SURE, DSM, and Tweedie's rule in a single objective.
- We evaluate SURE-Score by learning score-based generative models from noisy data for MIMO channels and MRI. Even though the distributional structure in the two domains is different, our results show that SURE-Score can learn accurate priors even at an SNR of 0 dB.
- Our simulation results show that MIMO channels acquired at 0 dB can be used for compressed channel estimation with a performance drop of at most 8 dB in normalized mean square error (NMSE) compared to fully supervised. We also show that self-supervised denoising and reconstruction at $5\times$ accelerated MRI is possible with about 15% error increase when learning a prior from noisy samples acquired at 10 dB.

We use lowercase and uppercase letters to denote vectors (including scalars) and matrices, respectively. We use p_X to denote the probability distribution function (p.d.f.) of the random variable X. We use $\underline{x}$ to denote the flattened version of X, and X^H to denote the Hermitian of X. Unless explicitly stated, all signals are assumed to be complex-valued.

II. SYSTEM MODEL AND BACKGROUND

A. Linear Inverse Problems

We model signals corrupted by additive white Gaussian noise as follows:

$$\tilde{x} = x + w, \tag{1}$$

where $x \in \mathbb{C}^N$ is sampled from the distribution p_X. The noise $w \sim \mathcal{CN}(0, \sigma_w^2 I)$ is i.i.d. complex-valued Gaussian distribution, I is the identity matrix, and σ_w^2 is the noise power. In the general case of linear inverse problems, the measurements are given by:

$$y = Ax + n, \tag{2}$$

where $x \sim p_X$ is the ground-truth signal, $y \in \mathbb{C}^M$ are the observed measurements, $n \sim \mathcal{CN}(0, \sigma_n^2 I)$ is the measurement noise, and $A \in \mathbb{C}^{M \times N}$ is the forward operator. The goal of solving linear inverse problems is to estimate the specific signal x, given the measurements y, the forward operator A, and statistical models of the distributions of x and n.

B. Score-Based Generative Modeling

The score function ψ_X of a probability distribution p_X is defined as the gradient of its logarithm with respect to the sample [26]:

$$\psi_X(x) = \nabla_x \log p_X(x). \tag{3}$$

Learning the score function of the probability distribution avoids the challenges of normalization and is useful to downstream tasks such as sampling from the learned distribution or solving inverse problems [3]. The goal of score-based generative modeling is to learn the score function ψ_X, given a training dataset sampled from p_X.

Denoising score matching [12] learns the score of a family of perturbed signal distributions through a noise-conditional score network $s_\theta(x; \sigma)$. When conditioned on a noise level σ, the network is trained to predict the score of the distribution $p_{\tilde{X}}$ of perturbed real-valued signals $\tilde{x} = x + z$, where $z \sim \mathcal{N}(0, \sigma^2 I)$. The DSM loss is given by:

$$L_{\text{DSM},\theta}(x; \sigma) = \left\| s_\theta(\tilde{x}; \sigma) - \nabla_{\tilde{x}} \log p_{\tilde{X}|X}(\tilde{x}|x; \sigma) \right\|_2^2. \tag{4}$$

When noiseless data are available for training and choosing the added noise z as Gaussian i.i.d., the gradient of the conditional distribution can be expressed in closed form as $-z/\sigma^2$. Score-based generative models are trained using denoising score matching at multiple noise levels simultaneously [1]. Following [27], we weigh the loss at each level by σ^2 to normalize the magnitude of the DSM loss across all noise levels. This yields the expression for the multilevel DSM training loss:

$$L_{\text{DSM},\theta} = \mathbb{E}_{\substack{x \sim p_X \\ \sigma \sim p_\Sigma \\ z \sim \mathcal{N}(0, \sigma^2 I)}} \sigma^2 \left\| s_\theta(x + z; \sigma) + \frac{z}{\sigma^2} \right\|_2^2. \tag{5}$$

The distribution of training noise levels p_Σ is a uniform distribution across geometrically distributed noise levels between $\sigma_{\min}$ and $\sigma_{\max}$. We use the guidelines in [27] and implement the noise-conditional score network as $s_\theta(x; \sigma) = s_\theta(x)/\sigma$, where $s_\theta(x)$ is a deep neural network with learnable weights θ.

C. Solving Inverse Problems using Annealed Langevin Dynamics

The score-based generative model s_θ learned during training does not use any knowledge of the linear operator A or received measurements y in the setting of (2), but is still a powerful tool for solving linear inverse problems. The algorithm we use to generate a point estimate x_{est} is based on posterior sampling, proven in [28] to be sample optimal for linear inverse problems. This is a stochastic algorithm that draws a single sample from the posterior distribution:

$$x_{\text{est}} \sim p_{X|Y}(x|y). \tag{6}$$

838

We use a modified version of annealed Langevin dynamics [1] to sample from the posterior distribution. This is an iterative algorithm that starts from a random $x_0 \sim N(0, I)$ and requires access to the noise-conditional score function $\psi_{X_t|Y}(x_t|y; \sigma)$ at each step. For brevity, we omit σ_t from the notation in the following. The update rule is given by:

$$x_{t+1} = x_t + \alpha_t \cdot \psi_{X_t|Y}(x_t|y) + \sqrt{2\beta\alpha_t}\eta_t, \qquad (7)$$

where α_t is a time-varying step size, $\eta_t \sim \mathcal{N}(0, I)$ is annealing noise sampled and added at each update step, and β is a temperature hyper-parameter controlling the level of annealing noise. When $\beta = 1$, the algorithm in (7) becomes annealed Langevin dynamics. Following [27], we exponentially decay the step size as $\alpha_t = \alpha_0 \cdot (\sigma_t/\sigma_{\min})^2$, defining the schedule via two hyper-parameters. The term σ_t comes from the geometric distribution of training noise levels beginning from $\sigma_{\max}$ and ending at $\sigma_{\min}$.

The conditional score function $\psi_{X_t|Y}$ can be further expanded via Bayes' rule as:

$$\begin{aligned}
\psi_{X_t|Y}(x_t|y) &= \nabla_{x_t} \log p_{X_t|Y}(x_t|y) \\
&= \nabla_{x_t} \log \frac{p_{Y|X_t}(y|x_t)p_{X_t}(x_t)}{p_Y(y)} \qquad (8) \\
&= \psi_{Y|X_t}(y|x_t) + \psi_{X_t}(x_t),
\end{aligned}$$

where the denominator term vanishes due to not depending on x. The first term is related to the linear formulation of the inverse problem in (2). Because the noise is Gaussian, as $\sigma_t \to 0$, this term becomes:

$$\psi_{Y|X_t}(y|x_t) = \frac{A^{\mathrm{H}}(y - Ax_t)}{\sigma_n^2}, \qquad (9)$$

where A^{H} is the Hermitian transpose (adjoint) of A. We approximate the dependency on the iteration-dependent noise level σ_t in the first term of (8) by additionally annealing this term with a factor γ_t as in [3]:

$$\psi_{Y|X_t}(y|x_t) \approx \frac{A^{\mathrm{H}}(y - Ax_t)}{\sigma_n^2 + \gamma_t^2}, \qquad (10)$$

where we choose $\gamma_t^2 = \sigma_t^2$ in this work. The dependency of the second term in (8) on σ_t is included in a pre-trained noise-conditional score-based model $s_\theta(x; \sigma_t)$, which can readily approximate this term. The final update rule for sampling from the posterior distribution is given by:

$$x_{t+1} = x_t + \alpha_t \left(\frac{A^{\mathrm{H}}(y - Ax_t)}{\sigma_n^2 + \gamma_t^2} + s_\theta(x_t; \sigma_t) \right) + \sqrt{2\beta\alpha_t}\eta_t. \qquad (11)$$

We emphasize that our algorithm only approximates posterior sampling.

D. Stein's Unbiased Risk Estimate

It is clear from Section II-B that learning a prior via denoising score matching requires a clean dataset of training set $x \sim p_X$. However, in our setting we only have access to a training set of noisy samples $\tilde{x}$ sampled according to (1). Ideally, we wish to learn a denoising model $g_\phi(x)$ with parameters ϕ, such that it maps perturbed samples $\tilde{x}$ to clean samples x. One approach is to use Stein's unbiased risk estimate [29] to obtain an unbiased estimate of the ground-truth supervised denoising loss $\|g_\phi(\tilde{x}) - x\|_2^2$ for each sample x. When the sample is corrupted by i.i.d. Gaussian noise at noise level σ_w, the SURE expression is given by:

$$L_{\mathrm{SURE}, \phi}(\tilde{x}; \sigma_w) = \|\tilde{x} - g_\phi(\tilde{x})\|_2^2 + 2\sigma_w^2 \cdot \mathrm{div}_{\tilde{x}}(g_\phi(\tilde{x})), \quad (12)$$

where $\mathrm{div}_x(g_\phi(x))$ represents the divergence of the vector field (any function that maps inputs to outputs of the same dimension) g_ϕ evaluated at the point x, and is given by:

$$\mathrm{div}_x(g_\phi(x)) = \sum_i \frac{\partial[g_\phi(x)]_i}{\partial x_i}. \qquad (13)$$

Evaluating all partial derivatives in (13) requires a number of function evaluations proportional to the input dimension, which does not scale to real-world high-dimensional signals. Following [30], we approximate the divergence term using its Monte Carlo approximation with $b \sim \mathcal{N}(0, I)$ being i.i.d. Gaussian noise sampled for each sample in each batch, and the approximation given by:

$$\mathrm{div}_x(g_\phi(x)) \approx b^T \left(\frac{g_\phi(x + \epsilon b) - g_\phi(x)}{\epsilon} \right), \qquad (14)$$

for small ϵ. The SURE loss does not involve the ground-truth sample x and has the remarkable property that its expectation taken over $\tilde{x} \sim \mathcal{N}(x, \sigma_w^2 I)$ satisfies [11]:

$$\mathbb{E}_{\tilde{x}} L_{\mathrm{SURE}, \phi}(\tilde{x}; \sigma_w) = \mathbb{E}_x \|g_\phi(\tilde{x}) - x\|_2^2, \qquad (15)$$

thus it can be used as a training function for the model g_ϕ, when only a noisy set of training samples is available [30], and a mini-batch of noisy samples is used. Note that the SURE loss can be applied when a single noisy version of each sample is available in the dataset, which matches our setting.

III. PROPOSED METHOD: SURE-SCORE

Our proposed method learns the score of the perturbed distribution of minimum mean square error (MMSE) denoised samples. While this is distinct from learning the score of the true perturbed distribution of clean samples at arbitrary noise levels, our formulation couples SURE learning and denoising score matching via Tweedie's rule for Gaussian corruptions such that the two objectives are consistent with each other at a specific noise level. We call our method SURE-Score. A block diagram of the training flow is shown in Figure 1.

Naively, learning the score of perturbed MMSE denoised samples could be done with two different functions, each with its own set of learnable parameters, and applied sequentially:

1) A function g_ϕ that learns the MMSE denoiser. This function can be learned purely from noisy data using the SURE training objective described in Section II-D.
2) A score-based generative model s_θ trained on the outputs of g_ϕ. If sufficient training data are available, it could be split between the two stages, otherwise the same training set used to learn g_ϕ can be re-used in the second stage.

839

$$L_{\text{SURE-Score},\theta}(\tilde{x};\sigma,\sigma_w) = \sigma^2 \left\| s_\theta(g_\theta(\tilde{x};\sigma_w) + z;\sigma) + \frac{z}{\sigma^2} \right\|_2^2 + \lambda \|\tilde{x} - g_\theta(\tilde{x};\sigma_w)\|_2^2 + 2\lambda\sigma_w^2 \cdot \text{div}_{\tilde{x}}(g_\theta(\tilde{x};\sigma_w)) \tag{16}$$

$$= \sigma^2 \left\| s_\theta(\tilde{x} + \sigma_w^2 s_\theta(\tilde{x};\sigma_w) + z;\sigma) + \frac{z}{\sigma^2} \right\|_2^2 + \lambda \left\| \sigma_w^2 s_\theta(\tilde{x};\sigma_w) \right\|_2^2 + 2\lambda\sigma_w^2 \cdot \text{div}_{\tilde{x}}(\tilde{x} + \sigma_w^2 s_\theta(\tilde{x};\sigma_w)). \tag{17}$$
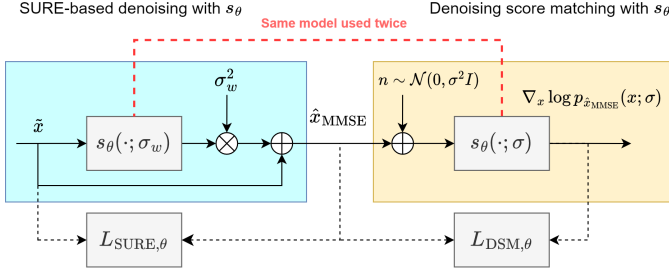

Fig. 1. Flow of SURE-Score during training. The same deep neural network s_θ is used first for denoising and subsequently for denoising score matching. The terms in (17) are obtained from the input, $\hat{x}_{\text{MMSE}}$, and the final output.

Inspired by the recent work in [17], we take advantage of the inherent connection between the two steps given by Tweedie's rule [13] which allows us to efficiently re-use the same network parameters for both purposes and reduce storage requirements. Assuming that a score-based generative model $s_\theta(\tilde{x};\sigma)$ is available and already learned, the MMSE denoiser in the case of AWGN is given by its *Tweedie re-parameterization* as [17]:

$$g_\theta(\tilde{x};\sigma) = \tilde{x} + \sigma^2 \cdot s_\theta(\tilde{x};\sigma). \tag{18}$$

Conversely, given access to an already trained MMSE denoiser $g_\theta(\tilde{x};\sigma)$ its score re-parameterization is given by [17]:

$$s_\theta(\tilde{x};\sigma) = \frac{g_\theta(\tilde{x};\sigma) - \tilde{x}}{\sigma^2}, \tag{19}$$

where the above is due to the existence and uniqueness of the MMSE estimator under finite means and variance of the perturbed distribution of $\tilde{x}$, hence Tweedie's rule can be directly inverted for $\sigma > 0$.

The previous identities suggest that a score-based generative model for a specific noise level and the MMSE denoiser for the same noise level are interchangeable. Thus, we propose to learn them *jointly* using the following loss function at a single noise level σ, with training data corrupted at noise level σ_w:

$$L_{\text{SURE-Score},\theta}(\tilde{x};\sigma,\sigma_w) = L_{\text{DSM},\theta}(g_\theta(\tilde{x};\sigma_w);\sigma) + \lambda \cdot L_{\text{SURE},\theta}(\tilde{x};\sigma_w), \tag{20}$$

where $\lambda > 0$ is a hyper-parameter controlling the relative weighting between the two losses. Note that in the above, the DSM term uses the output of the denoiser g_θ – this leads to a recurrent call of the same underlying network. Each of the two terms in the above corresponds to one of two objectives: the first term learns the score-based generative model for the

perturbed distribution of MMSE denoised samples, while the second term learns the MMSE denoiser itself.

The Tweedie re-parameterization introduces a consistency between the two objectives at noise level σ_w, that is when $\sigma = \sigma_w$ in (20): the more accurately the true score function is approximated at noise level σ_w, the more accurately denoising becomes, and vice versa. This can be viewed as a coarse version of the recently introduced consistency objective in [31], where we ensure that removing and adding back noise at noise level σ_w leads to consistently learned gradients of the perturbed score function. In practice, both terms at $\sigma = \sigma_w$ are bounded and non-zero due to the non-zero residuals of MMSE denoising. Additionally, the above consistency is only introduced for $\sigma = \sigma_w$, but in practice we find that this does not hinder performance at other noise levels, where extrapolation is sufficiently accurate.

Expanding the expected versions of the two losses in (20) yields the expression in (16). Replacing g_θ with its Tweedie re-parameterization yields our SURE-Score loss function in (17), which only depends on the score network s_θ and regularization λ. The training loss we use in practice is a mini-batch version of (17), with expectations equal to those in (5).

Note that while all the derivations so far have relied on the assumption that $\tilde{x}$ is corrupted by i.i.d. Gaussian noise, both the SURE loss and the Tweedie re-parameterization can be readily extended to other exponential family distributions [17], [32], making our proposed approach easily applicable to other common noise models as well.

Finally, our training and reconstruction algorithms require the selection of several hyper-parameters without access to ground-truth for tuning. In particular, we inherit all hyper-parameters related to both denoising score matching and to SURE; notable among them is the step size schedule in DSM given by p_Σ, which is known to depend on the dataset [27], and the Monte Carlo approximation ϵ, which is known to depend on the noise level σ_w [18], [30]. We also must choose λ to balance the two losses in (17). Note that setting $\lambda = 0$ in (17) is not equivalent to naively training the score model due to the nesting of the function s_θ.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We test our approach on two linear inverse problems: estimating the MIMO wireless channel matrix from compressed pilot measurements, and multi-coil MRI reconstruction from limited k-space measurements. We compare our approach to the following baselines: (i) supervised score-model training, (ii) naive DSM training on the noisy data, (iii) BM3D, (iv)

840

Noise2Self, and (v) Noise2Score. In (iii-iv), we apply DSM training on top of the denoised dataset.

We qualitatively evaluate score models learned from training data at different SNR levels (given by SNR^w). We also quantitatively evaluate posterior reconstruction performance. Following convention, we report normalized mean squared error (NMSE) in dB for channel estimation and normalized root mean squared error (NRMSE) for MRI, each at particular sub-sampling ratios and for different values of σ_w^2 and σ_n^2.

A. Model Architecture

We use a deep neural network based on the RefineNet model [33], following the NCSNv2 architecture in [27][1]. To handle complex-valued signals for both applications, we treat the real and imaginary parts as two real-valued channels of the same sample. As suggested by other works [16], we set $\epsilon = 10^{-3}$ in (14). We use the Adam optimizer with a fixed learning rate of 10^{-4} for all experiments. The parameter λ in (20) was set to $L_{\text{DSM},\theta}/L_{\text{SURE},\theta}$ using the first batch at the start of training. Though it may not be optimal, we found this heuristic worked well across our experiments. Minimizing the loss in (16) involves calculating the gradient through the nesting of the function s_θ; therefore we did not use Stopgrad in its calculation. We did not observe numerical instability in our particular experiments.

B. MIMO Channel Modeling and Estimation

We consider a point-to-point MIMO baseband communication scenario where the transmitter and receiver are equipped with N_t and N_r antennas, respectively. Channel estimation involves estimating the channel matrix $H \in \mathbb{C}^{N_r \times N_t}$, given measurements Y of the form:

$$Y = HP + N, \qquad (21)$$

where the matrix $P \in \mathbb{C}^{N_t \times N_p}$ is the measurement matrix known ahead of time (commonly called a pilot matrix), and N is additive white Gaussian noise with variance σ_n^2. Equation (21) can be vectorized as:

$$\underline{\mathbf{y}} = \left(P^H \otimes I_{N_r}\right)\underline{\mathbf{h}} + \underline{\mathbf{n}}, \qquad (22)$$

where I_{N_r} represents the square identity matrix of size N_r. Thus, this problem falls in the general class of linear, noisy inverse problems given by (2). We use randomly chosen quadrature phase shift keying (QPSK) signals for the entries of P. Channel estimation is an under-determined inverse problem when the pilot density ratio $\alpha = N_p/N_t < 1$. We normalize channels during training and inference by dividing by $\mathbb{E}_{H_{\text{train}}}[\|H\|_2^2]$. The SNR for a distribution of channels is defined as: $\mathbb{E}_H[\|H\|_2^2]/\sigma_w^2 = 1/\sigma_w^2$.

We train a score-based model s_θ with 32 hidden layers in the first block using the proposed SURE-Score framework on perturbed clustered delay line (CDL) channels $\tilde{H}$ at $\text{SNR}^w = 0$ dB, in particular from the CDL-C non-line-of-sight distribution. We simulate 10000 (possibly noisy) samples for training for each method where $N_r = 16$ and $N_t = 64$.
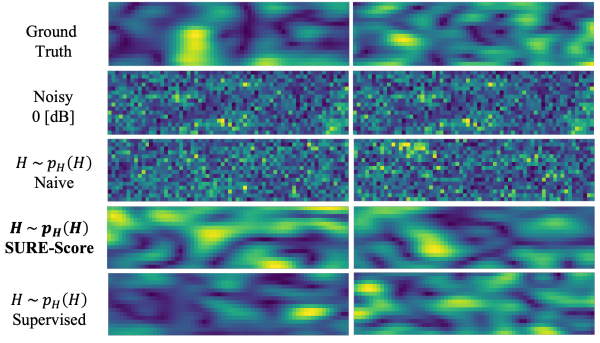
Fig. 2. Prior sampling performance for three methods: naive (ii), SURE-Score at $\text{SNR}^w = 0$ dB, and supervised (i). Each column shows a different realization of CDL-C channels.

1) Prior Sampling: To sample unconditionally from the learned prior $p_{\hat{x}_{\text{MMSE}}}$, we use the update in (11) with $A = 0$, eliminating the measurement-dependent term. Figure 2 shows CDL-C samples from the training set, noisy samples at SNR^w, and samples using three learned priors: (i) naive DSM, (ii) SURE-Score, and (iii) DSM with clean data. Qualitatively, we can see that prior samples generated from a score model naively trained on noisy channels are corrupted. However, both supervised and SURE-Score samples qualitatively match the signal structure in the reference samples. This is further validated in our posterior sampling experiments.
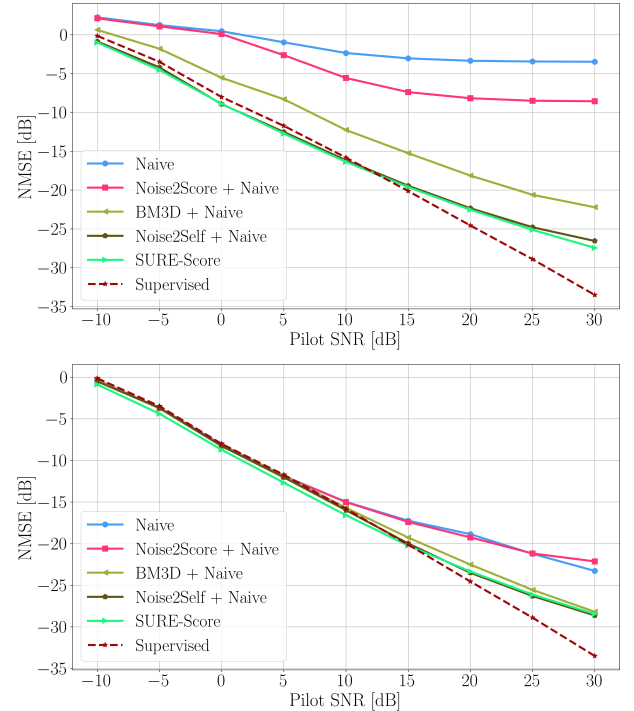

Fig. 3. Channel estimation performance at $\alpha = 0.6$ (38 pilots) using score models trained on CDL-C channels at SNR^w: 0 dB (top) and 10 dB (bottom).

2) Posterior Reconstruction: We use the annealed Langevin dynamics update in (11) with the data consistency term based on (22). We simulate a pilot density of $\alpha = 0.6$, i.e., 38 pilots.
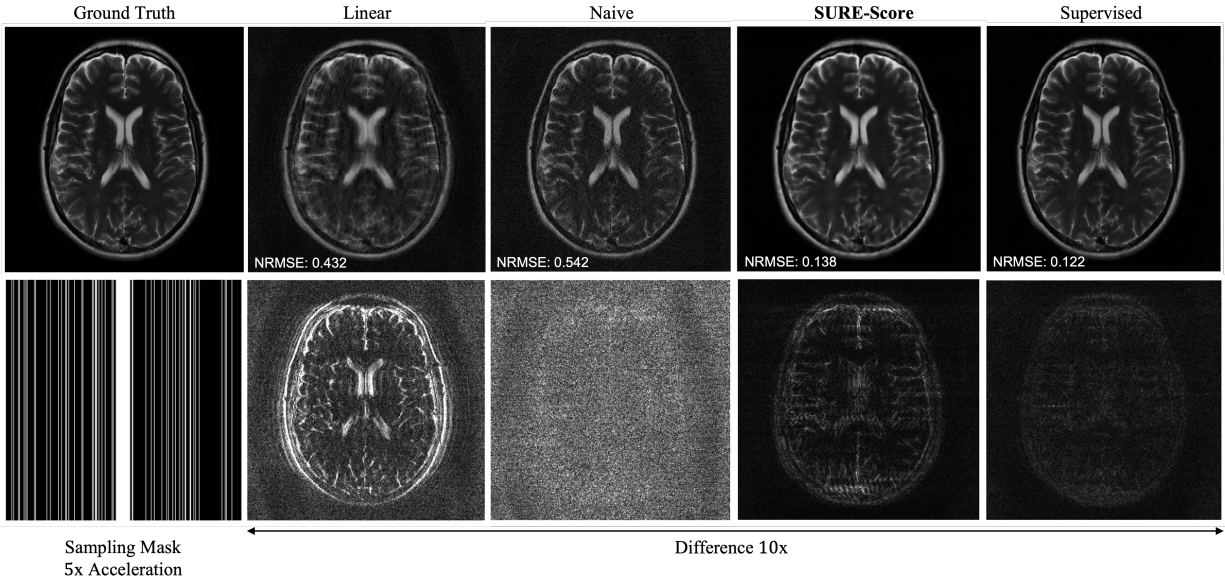
Fig. 4. Multi-coil MRI reconstruction at acceleration factor of $5\times$. From left to right: fully sampled ground truth, linear reconstruction, posterior sampling after naively training on noisy data at SNR^w, posterior sampling after training with SURE-Score, and posterior sampling after training with noise-free data. The bottom row shows the sampling pattern and difference images for each method, respectively.

Figure 3 summarizes results from posterior sampling using different training methods and at different pilot SNR levels, where we assume training samples $\tilde{H}$ contain noise with SNR of 0 dB and 10 dB in the two plots. We observe that naive training plateaus in estimation performance because of over-fitting to the noisy training set. Noise2Score and BM3D suffer at lower SNR^w and improve at higher SNR, consistent with results reported in the literature [17]. SURE-Score performs close to optimal with respect to supervised DSM except at higher pilot SNR. Noise2Self as a denoising pre-processing step also performed well. We observe a performance gap at high pilot SNR, likely due to the performance limits of the learned MMSE denoiser for our particular architecture and the finite amount of training data.

C. Multi-Coil MRI Reconstruction

Multi-coil MRI data are acquired in the frequency domain by placing multiple RF coils around the imaging anatomy. We assume a vectorized image $x \in \mathbb{C}^N$, with N_c RF receive sensitivity profiles each represented by a diagonal matrix $S_i \in \mathbb{C}^{N \times N}$, and Fourier sampling $F_\alpha \in \mathbb{C}^{\alpha N \times N}$. The acquired k-space data from the ith coil is given by:

$$y_i = F_\alpha S_i x + n_i, \tag{23}$$

where n_i is additive white Gaussian noise with variance σ_w^2. Following convention, we define the acceleration factor $1/\alpha$ relative to a single fully sampled coil, such that the total number of acquired measurements is $\alpha \cdot N_c \cdot N$. We can represent the MRI multi-coil forward model according to (2).

We train score-based models with 128 hidden layers in the first block using 10000 multi-coil T2-weighted brain MRI scans from the FastMRI dataset [34]. The SURE-Score model

was trained on a subset of the data (2000 samples) due to training time constraints. We first normalize all slices in the training set by performing a root sum-of-squares (RSS) reconstruction using k-space data from a central window of size 24×24. We divide the k-space of each slice by the 95th magnitude percentile of this low-resolution reconstruction. We define the SNR as $1/\sigma_w^2$.

1) Denoising Performance: We evaluate denoising performance using the Tweedie re-parameterization in (18) for score models trained (i) naively, (ii) with SURE-Score, and (iii) using noise-free data (supervised). Table I lists the NRMSE (mean and standard deviation) over 100 validation slices for SNR^w levels of 0 dB and 10 dB. The denoising performance of SURE-Score nearly matches supervised learning, indicating consistency at σ_w between the SURE objective and the score-matching objective. Thus, denoising score matching at multiple levels does not degrade denoising performance at σ_w.

Denoising Performance (NRMSE)		
SNR^w	0 dB	10 dB
Naive	2.48 ± 0.24	0.70 ± 0.07
Supervised	0.21 ± 0.01	0.14 ± 0.01
SURE-Score	0.23 ± 0.01	0.16 ± 0.01

TABLE I: MRI DENOISING PERFORMANCE OF SCORE MODELS TRAINED WITH NOISE-CORRUPTED DATA AND DIFFERENT OBJECTIVE FUNCTIONS.

2) Posterior Reconstruction: We use annealed Langevin dynamics as previously described in (7). We simulate an acceleration factor of five using a vertical sampling mask with fully sampled central k-space and uniform random sub-sampling elsewhere. We assume high SNR and set $\sigma_n = 0$ during inference.

Table II displays reconstruction NRMSE (mean and standard deviation) over a validation set of 50 slices. We compare

linear reconstruction, naive training, supervised training, and SURE-Score. We observe similar trends as in our MIMO experiments. Figure 4 compares example reconstructions after training with noisy data at SNR^w. Naive training leads to noise amplification, while SURE-Score qualitatively and quantitatively improves over this result.

Reconstruction Performance (NRMSE)		
SNR^w	0 dB	10 dB
Linear	0.53 ± 0.19	0.53 ± 0.19
Naive	1.42 ± 0.35	0.71 ± 0.15
Supervised	0.16 ± 0.05	0.16 ± 0.05
SURE-Score	0.22 ± 0.06	0.19 ± 0.05

TABLE II: COMPARISON OF RECONSTRUCTION NRMSE (MEAN AND STANDARD DEVIATION) OVER THE VALIDATION SET FOR FASTMRI.

V. CONCLUSION

Through our results, we can observe the impact of SURE-based denoising on prior sampling as well as for solving inverse problems in the wireless and MR domain while learning from datasets corrupted with a noise-to-signal ratio of up to 0 dB. We also observe that self-supervised techniques like SURE-Score can match supervised denoising performance. While our algorithm matches the convergence speed of supervised score models in terms of epochs, we find the runtime per iteration increases due to an additional pass through the network at each step. Another challenge, like in other self-supervised methods, is choosing hyper-parameters without access to ground truth data. Finally, our work currently assumes white Gaussian noise corruption but could be extended to arbitrary exponential families using Generalized SURE [32].

REFERENCES

[1] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[3] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir, "Robust compressed sensing mri with deep generative priors," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 938–14 954, 2021.

[4] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," *arXiv preprint arXiv:2111.08005*, 2021.

[5] H. Chung and J. C. Ye, "Score-based diffusion models for accelerated mri," *Medical Image Analysis*, vol. 80, p. 102479, 2022.

[6] G. Luo, M. Blumenthal, M. Heide, and M. Uecker, "Bayesian mri reconstruction with joint uncertainty estimation using diffusion models," *Magnetic Resonance in Medicine*, 2023.

[7] M. Arvinte and J. I. Tamir, "Mimo channel estimation using score-based generative models," *IEEE Transactions on Wireless Communications*, 2022.

[8] N. Zilberstein, C. Dick, R. Doost-Mohammady, A. Sabharwal, and S. Segarra, "Annealed langevin dynamics for massive mimo detection," *IEEE Transactions on Wireless Communications*, 2022.

[9] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *arXiv preprint arXiv:2201.11793*, 2022.

[10] F. Koehler, A. Heckett, and A. Risteski, "Statistical efficiency of score matching: The view from isoperimetry," *arXiv preprint arXiv:2210.00726*, 2022.

[11] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[12] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[13] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.

[14] M. Elad, B. Kawar, and G. Vaksman, "Image denoising: The deep learning revolution and beyond–a survey paper–," *arXiv preprint arXiv:2301.03362*, 2023.

[15] S. Soltanayev and S. Y. Chun, "Training deep learning based denoisers without ground truth data," *Advances in neural information processing systems*, vol. 31, 2018.

[16] C. A. Metzler, A. Mousavi, R. Heckel, and R. G. Baraniuk, "Unsupervised learning with stein's unbiased risk estimator," *arXiv preprint arXiv:1805.10531*, 2018.

[17] K. Kim and J. C. Ye, "Noise2score: tweedie's approach to self-supervised image denoising without clean images," *Advances in Neural Information Processing Systems*, vol. 34, pp. 864–874, 2021.

[18] M. Arvinte and J. Tamir, "Score-based generative models for wireless channel modeling and estimation," in *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[19] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2965–2974.

[20] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," in *International Conference on Machine Learning*. PMLR, 2019, pp. 524–533.

[21] T. Xiang, M. Yurt, A. B. Syed, K. Setsompop, and A. Chaudhari, "Ddm2: Self-supervised diffusion mri denoising with generative diffusion models," *arXiv preprint arXiv:2302.03018*, 2023.

[22] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing mri," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[23] A. Alkhateeb, G. Leus, and R. W. Heath, "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 2909–2913.

[24] Z. Kadkhodaie and E. Simoncelli, "Stochastic solutions for linear inverse problems using the prior implicit in a denoiser," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 13 242–13 254.

[25] Z.-X. Cui, C. Cao, S. Liu, Q. Zhu, J. Cheng, H. Wang, Y. Zhu, and D. Liang, "Self-score: Self-supervised learning on score-based models for mri reconstruction," *arXiv preprint arXiv:2209.00835*, 2022.

[26] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[27] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," *Advances in neural information processing systems*, vol. 33, pp. 12 438–12 448, 2020.

[28] A. Jalal, S. Karmalkar, A. Dimakis, and E. Price, "Instance-optimal compressed sensing via posterior sampling," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4709–4720.

[29] K.-C. Li, "From stein's unbiased risk estimates to the method of generalized cross validation," *The Annals of Statistics*, pp. 1352–1377, 1985.

[30] S. Ramani, T. Blu, and M. Unser, "Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms," *IEEE Transactions on image processing*, vol. 17, no. 9, pp. 1540–1554, 2008.

[31] G. Daras, Y. Dagan, A. G. Dimakis, and C. Daskalakis, "Consistent diffusion models: Mitigating sampling drift by learning to be consistent," *arXiv preprint arXiv:2302.09057*, 2023.

[32] Y. C. Eldar, "Generalized sure for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 471–481, 2008.

[33] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[34] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno *et al.*, "fastmri: An open dataset and benchmarks for accelerated mri," *arXiv preprint arXiv:1811.08839*, 2018.