



Subject Areas:

signal processing, electrical engineering, artificial intelligence

Keywords:

deep generative models, Bayesian inference, inverse modeling, active inference, medical imaging, ultrasound, MRI, radar, diffusion models

Author for correspondence:

Ruud J.G. van Sloun

e-mail: R.J.G.v.Sloun@tue.nl

Deep Generative Models for Bayesian Inference on High-Rate Sensor Data: Applications in Automotive Radar and Medical Imaging

Tristan S.W. Stevens¹, Jeroen Overdevest^{1,2},
Oisín Nolan¹, Wessel L. van Nierop¹,
Ruud J.G. van Sloun¹, Yonina C. Eldar³

¹Department of Electrical Engineering, University of Technology Eindhoven, the Netherlands

²Signal Processing Algorithms, NXP Semiconductors, the Netherlands

³Faculty of Mathematics and Computer Science, Weizmann institute of Science, Israel

Deep generative models have been studied and developed primarily in the context of natural images and computer vision. This has spurred the development of (Bayesian) methods that use these generative models for inverse problems in image restoration, such as denoising, inpainting, and super-resolution. In recent years, generative modeling for Bayesian inference on sensory data has also gained traction. Nevertheless, the direct application of generative modeling techniques initially designed for natural images on raw sensory data is not straightforward, requiring solutions that deal with high dynamic range signals acquired from multiple sensors or arrays of sensors that interfere with each other, and that typically acquire data at a very high rate. Moreover, the exact physical data-generating process is often complex or unknown. As a consequence, approximate models are used, resulting in discrepancies between model predictions and the observations that are non-Gaussian, in turn complicating the Bayesian inverse problem. Finally, sensor data is often used in real-time processing or decision-making systems, imposing stringent requirements on, e.g., latency and throughput. In this paper, we will discuss some of these challenges and offer approaches to address them, all in the context of high-rate real-time sensing applications in automotive radar and medical imaging.

© The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, provided the original author and source are credited.

1. Introduction

Active array sensing techniques are at the core of numerous advanced technologies, playing a critical role in fields such as automotive radar and medical imaging. These sensor arrays consist of multiple individual elements that actively emit signals and capture the reflected waves (sound or light) to obtain accurate representations of the scenery. The culmination of many sensory signals leads to massive data rates and corresponding challenges [1–3]. These systems are required to provide high-resolution and real-time imaging, all while dealing with interference, noise and a changing environment.

Many of the challenges faced in array sensing can be effectively reformulated as inverse problems, which seek to estimate unknown parameters from observations. In the context of array sensing, this typically means reconstructing images or (distance, velocity or angular) measures from the raw data captured by the sensors. However, due to the underdetermined nature of these inverse problems, strong priors and knowledge of the underlying processes at hand are key for reliable reconstructions. This necessitates advanced modeling techniques that extract and capture meaningful information from the raw sensory data.

Recent advances in **Deep Generative Models (DGMs)** have shown great potential in solving problems with high-dimensional data by learning and exploiting the data manifold in domains ranging from medical imaging [4], computer vision [5] and natural language processing [6]. Specifically, they seek to model the distribution of data and subsequently sample from it, which can serve as a signal prior and aid in the inverse problem solving. However, applying these techniques directly to raw sensory data presents additional challenges due to the **high dynamic range (HDR)** and rapid data acquisition rates, which impose stringent requirements on latency and throughput, further complicating the use of **DGMs**. Consequently, these challenges necessitate the development of tailored techniques involving generative models that effectively manage the unique characteristics of sensory data while adhering to the requirements of high-rate real-time sensing applications.

In this paper, we review ongoing work in application of deep generative models to sensory data. We start with some background on sensing applications, **DGMs** and finally Bayesian inference with **DGMs** in Section 2. Then, we discuss the two main challenges, namely model mismatches and real-time inference in Sections 3 and 4, respectively. In both of these sections, we discuss methods that mitigate these challenges, ranging from improved forward models with the use of **DGMs**, to accelerated inference techniques through approaches such as *deep unfolding*, *temporal inference* and *active compressed sensing*.

2. Background

(a) Sensing applications

Many sensor modalities, such as medical imaging and radar sensing, share a common objective, namely the measurement of an *unknown* channel impulse response. These channels are often deeply complex, with signals exhibiting significant dynamic range variations in combination with high sample rates. These characteristics of sensory data challenge accurate signal recovery. For instance, in both ultrasound imaging and radar, the generative model of the sensory data is not exactly known due to the highly variable reflectivity of objects within the sensor's line of sight and the presence of multipath propagation. Extracting a signal-of-interest from raw data is a challenging task for existing model-based techniques, e.g., due to the lack of modeling capacity, non-Gaussian (structured) noise or other undesirable sensing phenomena such as interference or aberration. Another challenge faced when dealing with sensory data are high data rates, for example in real-time imaging applications. **Compressed Sensing (CS)** [7–10] has emerged as a powerful technique for reducing data rates by compressing the size of measurement vector \mathbf{y} necessary to recover the signal-of-interest \mathbf{x} , through intelligent design of the sensing matrix

\mathbf{A} and exploitation of known signal statistics. A common application of CS is via subsampling, which aims to recover the full signal-of-interest from a subset of possible measurements, typically fewer than required by the Nyquist-Shannon theorem. This has been successfully applied to reducing data rates in many domains, such as ultrasound imaging [11–13], radar [14,15], and Magnetic Resonance Imaging (MRI) [16].

Throughout this work, we refer to the following forward model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} + \boldsymbol{\epsilon}, \quad \mathbf{x} \in \mathbb{R}^M, \{\mathbf{y}, \mathbf{n}, \boldsymbol{\epsilon}\} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{N \times M} \quad (2.1)$$

where the observations \mathbf{y} are contaminated with *structured noise* \mathbf{n} and thermal noise $\boldsymbol{\epsilon}$, respectively. The thermal noise is assumed to be AWGN, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$ where σ_ϵ^2 represents its variance. In contrast, structured noise encapsulates all model errors (mismatch), including possibly non-linear effects, distortions and multipath components in the observations \mathbf{y} . The nature of \mathbf{n} depends on the sensing scenario. In some applications, \mathbf{n} may be independent of \mathbf{x} , such as structured interference from external sources, commonly encountered in automotive applications of radar. However, in other scenarios, such as diffraction or multipath scattering in ultrasound imaging, \mathbf{n} is inherently a function of \mathbf{x} , e.g. $\mathbf{n}(\mathbf{x})$. Explicitly modeling (and performing inference on) the true forward physics model of \mathbf{n} is often challenging. Given the complexity of capturing potential dependencies between \mathbf{n} and \mathbf{x} , we assume independence, i.e., $p(\mathbf{n}|\mathbf{x}) \approx p(\mathbf{n})$ and instead learn the marginal distribution $p(\mathbf{n})$ in a fully data-driven fashion, as discussed in Section 3.

For our goal of inferring the underlying signal-of-interest \mathbf{x} from observations \mathbf{y} , we make use of the forward model in (2.1) combined with statistical priors. To establish these priors, we resort to Deep Learning (DL) which has been proven to be effective for tasks that require accurate statistical models learned from the data itself. While black box approaches often fail when the trained networks are subjected to out-of-distribution data [17], recently, Deep Generative Models (DGMs) have shown exceptional capabilities when used in conjunction with Bayesian theory. By conditioning the generative process on observations, DGMs provide a robust framework for solving inverse problems, such as signal recovery in the presence of structured noise.

In the following, we provide a brief introduction to DGMs and their role in posterior sampling. We then highlight their growing use in various sensing applications, demonstrating their potential to enhance the accuracy and robustness of signal recovery in these challenging environments.

(b) Deep generative models

Generative models try to understand and model the underlying distribution of data and have an effective way of sampling new data points from this distribution. DGMs are a class of generative models that have specifically gained traction due to their ability to model high-dimensional data. At the core of DGMs are general parameterized function approximators, usually neural networks. These networks are trained on many examples from a training dataset, representing the data distribution. DGMs are able to effectively capture the structure of the data manifold as they leverage the property that all data points lie on a lower-dimensional manifold embedded in the high-dimensional data space [18].

Nonetheless, modeling distributions with DGMs poses several challenges. Probability density functions are constrained to be non-negative and integrate to one, which limits the choice of neural architectures. Variational Autoencoders (VAEs) [19] circumvent the intractability of density estimation by approximating it with a variational lower bound. Generative Adversarial Networks (GANs) [20] are another class of DGMs that learn the data distribution implicitly through an adversarial objective. Normalizing Flows (NFs) [21] take a different approach altogether by transforming a simple base distribution into the target distribution through a series of invertible transformations.

Here, we focus on a more recent development in the field of DGMs, namely Diffusion Models (DMs). These models indirectly model the underlying distribution $p_{\text{data}}(\mathbf{x})$ through the score

function $\nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})$, which is the gradient of the probability density function with respect to the data itself. Unlike likelihood-based methods, this circumvents the need of directly modeling the probability density function. Furthermore, it leads to an interpretable denoising score-matching objective, which allows us to parameterize the score function with any neural network $s_{\theta}(\mathbf{x})$ and train it as follows:

$$\arg \min_{\theta} \mathbb{E}_{p_{\text{data}}(\mathbf{x})} [\|s_{\theta}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x})\|_2^2]. \quad (2.2)$$

This effectively results in a function that points back towards the data manifold and can be used to sample from the data distribution p_{data} . DMs model this sampling procedure through the reversal of a corruption process, also known as forward diffusion, which progressively adds increasing levels of **additive white Gaussian noise (AWGN)** until the sample is completely transformed from the original data distribution $\mathbf{x}_0 \equiv \mathbf{x} \sim p_{\text{data}}$ to a Gaussian noise sample $\mathbf{x}_{\mathcal{T}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, with diffusion time $\tau \in [0, \mathcal{T}]$. This continuous forward process $\mathbf{x}_0 \rightarrow \mathbf{x}_{\tau} \rightarrow \mathbf{x}_{\mathcal{T}}$ can be formalized using a **Stochastic Differential Equation (SDE)**:

$$d\mathbf{x} = f(\tau)\mathbf{x} + g(\tau)d\mathbf{w}, \quad (2.3)$$

where $\mathbf{w} \in \mathbb{R}^d$ is a standard Wiener process, $f(\tau) : [0, \mathcal{T}] \rightarrow \mathbb{R}$ and $g(\tau) : [0, \mathcal{T}] \rightarrow \mathbb{R}$ are the drift and diffusion coefficients, which contribute to the deterministic and stochastic aspects of the **SDE**, respectively. Naturally, we are interested in the reversal of this process, which leads to the reverse diffusion process which has shown to result in a reverse-time **SDE** as follows [22]:

$$d\mathbf{x} = \left[f(\tau)\mathbf{x} - g(\tau)^2 \underbrace{\nabla_{\mathbf{x}_{\tau}} \log p(\mathbf{x}_{\tau})}_{\text{score}} \right] d\tau + g(\tau)d\bar{\mathbf{w}}_{\tau}, \quad (2.4)$$

where $d\tau$ and $d\bar{\mathbf{w}}$ are now processes running backwards in diffusion time. Conveniently, the *score function* emerges from this reverse diffusion process and can accordingly be substituted with the learned score model from (2.2) to gradually remove noise and sample from p_{data} . Moreover, to facilitate the training process, the score model is conditioned on the diffusion time step τ , resulting in a **Noise Conditional Score Network (NCSN)** $s_{\theta}(\mathbf{x}_{\tau}, \tau)$ which is able to jointly evaluate the score of all perturbed data distributions $\forall \tau \in [0, \mathcal{T}]$ [23].

(c) Posterior sampling

To reconstruct corrupted or incomplete incoming sensory data, according to (2.1), using generative models, we resort to a probabilistic framework with deep generative models serving as foundation for inferring the underlying data. The act of posterior sampling centers around the idea of incorporating both prior information $p(\mathbf{x})$ with incoming observations \mathbf{y} according to Bayes' rule. Many posterior sampling algorithms have been proposed for various generative modeling architectures, for example the conditional Wasserstein **GAN** [24]. Here, however, we focus on posterior sampling with **DMs** in order to provide the necessary background for the methods to follow.

As **DMs** generate samples using gradients of probability density functions, we start by using Bayes' rule to formulate the posterior score function:

$$\underbrace{\nabla_{\mathbf{x}_{\tau}} \log p(\mathbf{x}_{\tau} | \mathbf{y})}_{\text{posterior}} = \underbrace{\nabla_{\mathbf{x}_{\tau}} \log p(\mathbf{x}_{\tau})}_{\text{prior}} + \underbrace{\nabla_{\mathbf{x}_{\tau}} \log p(\mathbf{y} | \mathbf{x}_{\tau})}_{\text{likelihood}}. \quad (2.5)$$

This expression factorizes the posterior distribution into a prior distribution which we model with **DGMs** and a likelihood term, which is a known distribution given our understanding of the physical acquisition process of the observed sensory data \mathbf{y} , capturing how the true signal \mathbf{x} is corrupted by factors such as sensor noise, distortions and resolution limits. In order to achieve posterior sampling with pre-trained **DMs**, one can substitute the score function in (2.4) with the factorization of (2.5) leading to a conditional reverse-time diffusion process. The posterior score is then approximated as $\nabla_{\mathbf{x}_{\tau}} \log p(\mathbf{x}_{\tau} | \mathbf{y}) \approx s_{\theta}(\mathbf{x}_{\tau}, \tau) + \nabla_{\mathbf{x}_{\tau}} \log p(\mathbf{y} | \mathbf{x}_{\tau})$.

Unfortunately, the structured noise-perturbed likelihood $p(\mathbf{y}|\mathbf{x}_\tau)$ is intractable, in contrast to the noiseless case $p(\mathbf{y}|\mathbf{x}_0)$. Various posterior sampling methods for **DMs** have been proposed to estimate this quantity [25–27]. A widely used approach is **Diffusion Posterior Sampling (DPS)** [28], which leverages the posterior mean that is derived via first order Tweedie’s [29]:

$$p(\mathbf{x}_0|\mathbf{x}_\tau) \approx \mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau] = \frac{1}{\alpha_\tau} (\mathbf{x}_\tau + \sigma_\tau^2 \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)) \quad (2.6)$$

$$\approx \frac{1}{\alpha_\tau} (\mathbf{x}_\tau + \sigma_\tau^2 s_\theta(\mathbf{x}_\tau, \tau)) := \mathbf{x}_{0|\tau}, \quad (2.7)$$

where $\mathbf{x}_{0|\tau}$ represents the one-step denoising from diffusion step τ . The first approximation corresponds to the **Minimum Mean Squared Error (MMSE)** estimator for $p(\mathbf{x}_0|\mathbf{x}_\tau)$ [30], while the second substitutes the score function with the trained **NCSN**. Further, we reparameterize the **SDE** in (2.3) using signal and noise rates, α_τ and σ_τ , which can be derived from the noise scheduling $f(\tau)$, $g(\tau)$ [22], as $\mathbf{x}_\tau = \alpha_\tau \mathbf{x}_0 + \sigma_\tau \mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Finally, we approximate a tractable posterior score, by starting from (2.5) and substituting the approximate gradient of the log likelihood using (2.7) as follows:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y} | \mathbf{x}_\tau) \quad (2.8)$$

$$\approx s_\theta(\mathbf{x}_\tau, \tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y} | \mathbf{x}_{0|\tau}). \quad (2.9)$$

The exact implementation of this posterior sampling framework varies based on the specific application. To illustrate this, we will provide examples from ultrasound, radar, and **MRI** in the following sections.

3. Model mismatch

Model mismatch is a critical challenge in inverse problems involving real-world sensory data, where the assumed forward model deviates from the actual, often more complex, data acquisition process. While we often assume a known and accurate forward process $p(\mathbf{y} | \mathbf{x})$ as described in (2.1), this assumption rarely holds in practice.

DGMs have demonstrated strong performance in inverse problems when the forward model is fully known. However, in sensory data, the acquisition process often involves unknown propagation effects such as multi-path scattering, or sensor-specific distortions. Factors that are difficult to capture directly with a simple forward model or to learn directly from data using **DGMs**.

To close this gap, we discuss several key approaches. Firstly, in Section 3(a), we examine the concept of *structured noise*, where model errors are explicitly captured with a **DGM**, relaxing the reliance on a perfectly known forward model. Secondly, in Section 3(b), we address the challenge posed by the **high dynamic range (HDR)** of raw sensory data, which can complicate the training and application of generative models. Finally, in Section 3(c), we incorporate *model-based score functions* that leverage prior knowledge about the signal or sensing physics, enabling more robust guidance during inference. Throughout this section, the practical application of these concepts will be illustrated through two detailed examples in the domains of ultrasound imaging and automotive radar.

(a) Structured noise

One approach to dealing with model mismatch and other sensing and imaging artifacts is to model these error terms as structured noise. Logically, the structured noise cannot be captured using parametric probability distributions (such as Gaussian). Therefore, we resort to **DGMs** to learn its structure from data. This approach can effectively mitigate model mismatch in ultrasound and radar applications as we show in the following section.

For radar interference mitigation and multipath dehazing, similar source separation techniques have been applied through the use of joint posterior sampling using **DGMs** [31,32].

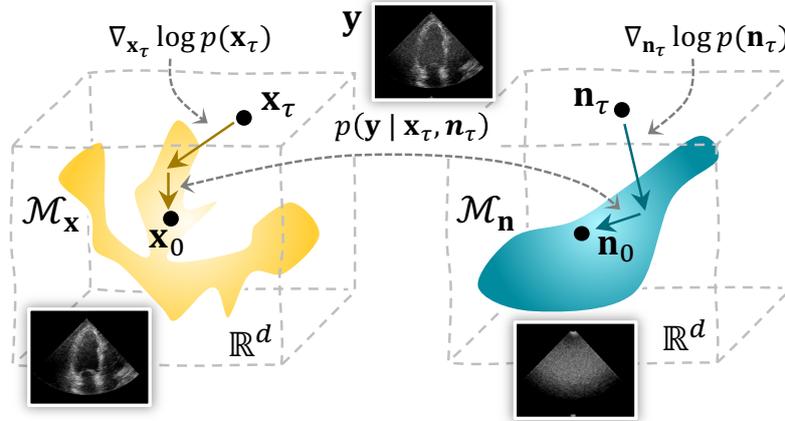


Figure 1: Overview of the proposed joint posterior sampling method for removing structured noise using DMs. During the sampling process, the solutions for both signal and structured noise move toward their respective data manifold \mathcal{M} through the score functions. At the same time, the data consistency term derived from the joint likelihood $p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{n}_\tau)$ ensures solutions that are in line with the (structured) noisy measurements. Figure adopted from [33].

The ill-posed problems are tackled by introducing two parallel generative processes that are conditioned on \mathbf{y} to create a joint posterior sampling process using DMs [33]. Using Bayes' rule, samples are drawn from the joint posterior distribution $p(\mathbf{x}, \mathbf{n}|\mathbf{y})$ to obtain estimates of both signal and structured noise components, \mathbf{x} and \mathbf{n} :

$$(\mathbf{x}, \mathbf{n}) \sim p(\mathbf{x}, \mathbf{n}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{n}) \cdot p(\mathbf{x}) \cdot p(\mathbf{n}), \quad (3.1)$$

where $p(\mathbf{y}|\mathbf{x}, \mathbf{n})$ is the likelihood according to our measurement model in (2.1) and $p(\mathbf{x})$ and $p(\mathbf{n})$ are prior distributions which can be modeled using score-based DMs with the objective in (2.2). The parallel posterior sampling (for both signal and structured noise components) is achieved by extension of (2.4), through substitution of the score with the score of the joint posterior distribution. This results in the coupled diffusion process described by the following reverse-time SDE:

$$d(\mathbf{x}_\tau, \mathbf{n}_\tau) = \left[f(\tau)(\mathbf{x}_\tau, \mathbf{n}_\tau) - g(\tau)^2 \nabla_{\mathbf{x}_\tau, \mathbf{n}_\tau} \log p(\mathbf{x}_\tau, \mathbf{n}_\tau|\mathbf{y}) \right] dt + g(t) \bar{\mathbf{w}}_t. \quad (3.2)$$

We can again factorize the joint posterior using Bayes' rule for scores as follows:

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau, \mathbf{n}_\tau|\mathbf{y}) = \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) + \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{n}_\tau), \quad (3.3)$$

$$\nabla_{\mathbf{n}_\tau} \log p(\mathbf{x}_\tau, \mathbf{n}_\tau|\mathbf{y}) = \nabla_{\mathbf{n}_\tau} \log p(\mathbf{n}_\tau) + \nabla_{\mathbf{n}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{n}_\tau). \quad (3.4)$$

From this factorization it follows that each separate reverse diffusion process (for both signal and structured noise) is entangled through the shared joint likelihood term $\log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{n}_\tau)$. The prior score of the signal component $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau)$ can be either learned using a DM or have an analytical prior such as sparsity. The structured noise score $\nabla_{\mathbf{n}_\tau} \log p(\mathbf{n}_\tau)$ is additionally learned using a separate DM due to its complex nature. For a detailed description of the full algorithm we refer the reader to [33]. In the following sections, we illustrate the practical application of these techniques addressing model mismatch, focusing on two major problems in the context of different sensing applications: ultrasound multipath scattering and radar interference.

Example 1: Ultrasound multipath scattering

Throughout this review, we will highlight several applications, demonstrating key techniques enabling the use of DGMs on sensory data. Each example is introduced with an information

box, explicitly specifying the forward model and how **DGM** is applied within the given context. Moreover, a short overview of data rates and sizes is provided to offer insight into the variability and characteristics of different types of sensory data.

Box 3.1 – Application: ultrasound imaging

Data characteristics: A typical ultrasound probe comprises hundreds of individual transducer elements, each of which operates at sampling frequencies at the Nyquist rate, typically in the range of tens of **Megahertz (MHz)**, leading to N fast-time samples per receive channel C . Depending on the transmit sequence (focused, diverging wave, plane wave) several hundreds slow-time sequences are acquired. In regular 2D brightness mode (B-Mode) imaging, at least several tens of frames per second can be expected, leading to raw data $\mathbf{y} \in \mathbb{R}^{C \times M \times N}$ rates that can quickly amount to several hundreds or thousands **Gigabits per second (Gbit/s)**. This problem is exaggerated in 3D ultrasound, where matrix probes consist of thousands of elements [34].

Forward model: \mathbf{y} is the observed (and hazy) beamformed **RF** data. \mathbf{x} is the clean beamformed **RF** data with the same dimensions, but only containing the direct path contributions. All clutter and multipath components are modeled through structured noise component \mathbf{n} using a separate **DGM**.

Application of DGMS: As both signal and haze contributions in the beamformed RF data are highly structured, **DGMs** can be fitted to both of these components.

The first application we discuss is ultrasound imaging, a widely used modality in medical diagnostics due to its non-invasive and real-time nature. See Box 3.1 for a general overview of the application. Through the transmittance of high-frequency sound waves into the body, internal tissue structures can be reconstructed from the backscattered echoes. However, ultrasound signals are subject to a range of different noise sources that clutter the image and limit interpretability. One of the major origins for loss in image quality is caused by multipath scattering amidst layers of skin, fat and muscle between the transducer and the tissue being examined. These multipath reflections amount to a haze-like appearance on the image, dubbed simply *haze*. Specifically, cardiac ultrasound is sensitive to haze due to the small transducer footprint and the addition of the ribs in line of sight of the probe. In order to suppress the multipath clutter \mathbf{n} and retrieve the direct path contribution \mathbf{x} from the measured ultrasound signals \mathbf{y} we consider the forward model in (2.1) and explicitly model both components separately with **DMs** [32,33]. For this purpose, we perform denoising score matching, see (2.2), on (unpaired) training data samples of clean ultrasound $\{\mathbf{x}^1, \dots, \mathbf{x}^L\} \sim p(\mathbf{x})$, and multipath haze recordings $\{\mathbf{n}^1, \dots, \mathbf{n}^L\} \sim p(\mathbf{n})$ to learn two separate score functions, conditioned on the diffusion time step τ :

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) \approx s_\theta(\mathbf{x}_\tau, \tau) \quad \text{and} \quad \nabla_{\mathbf{n}_\tau} \log p(\mathbf{n}_\tau) \approx s_\phi(\mathbf{n}_\tau, \tau). \quad (3.5)$$

Note that paired data of clean and hazy samples is not required to train these two generative models. This greatly reduces the difficulty of creating suitable datasets, as the structured noise can be acquired in isolation or simulated. See [32] for more details on the curation of the cardiac haze dataset. Moreover, learning each distribution with separate **DGMs** is more robust compared to a supervised method on paired data. The latter approach struggles with generalization due to the variability in paired samples and potential to overfit to specific instances of noise [32]. During inference, the two trained **DGMs** can be deployed within the joint-posterior sampling framework as seen in (3.3) and (3.4). The impact of a learned noise prior, compared to a traditional Gaussian prior on the problem of dehazing medical ultrasound data is illustrated in Fig. 2. The learned prior yields improved contrast and clearer structural details, whereas the Gaussian prior leaves residual noise with structured components, suggesting it inadvertently suppresses parts of the underlying signal.

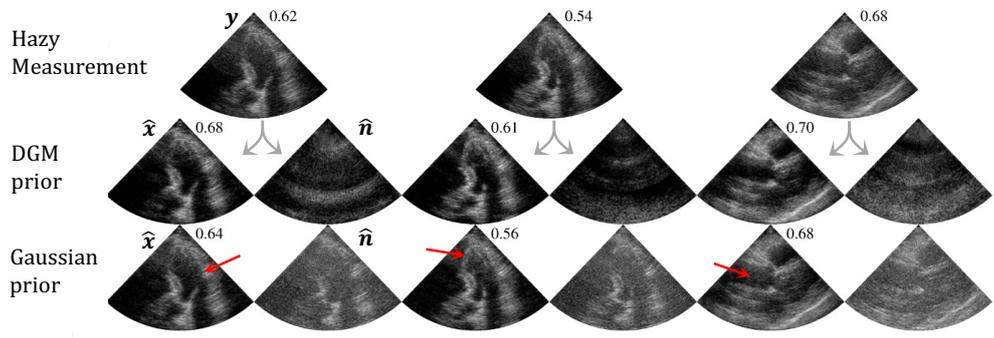


Figure 2: Comparison between a structured noise **DGM** prior and a Gaussian prior for the task of dehazing *in-vivo* medical ultrasound data. Posterior estimates of the signal $\hat{\mathbf{x}}$ and noise (haze) $\hat{\mathbf{n}}$ are shown for each method, alongside corresponding gCNR [35] (\uparrow) values, highlighting the improved performance of the structured noise prior. Figure adopted from [33].

(b) High dynamic range

Unlike natural images, which typically have a relatively narrow range of pixel intensities, raw **radio-frequency (RF)** data in sensor applications often exhibits a **high dynamic range (HDR)**, meaning that the signal amplitudes can vary drastically, see Fig 3. Besides this imposing constraints on the hardware side, requiring **HDR Analog-to-Digital Converters (ADCs)** [36], this also presents challenges when training generative models such as **DMs**. The wide range of intensities can lead to numerical instability, with gradients either exploding or vanishing, and can cause the network to focus disproportionately on the stronger signals while neglecting weaker, yet important, components. In [32], the **HDR** of ultrasound signals is addressed through transformation of the **RF** data using a technique known as *companding*. This is an invertible operation that can *compress* and *expand* the dynamic range of a signal as follows:

$$\text{compress: } C(\mathbf{x}_{\text{RF}}) = \text{sgn}(\mathbf{x}_{\text{RF}}) \frac{\ln(1 + \mu|\mathbf{x}_{\text{RF}}|)}{\ln(1 + \mu)}, \quad \text{expand: } C^{-1}(\mathbf{x}) = \text{sgn}(\mathbf{x}) \frac{(1 + \mu)^{|\mathbf{x}|} - 1}{\mu}, \quad (3.6)$$

with $-1 \leq \mathbf{x}_{\text{RF}} \leq 1$, $-1 \leq \mathbf{x} \leq 1$ and where μ is a parameter that determines the amount of compression applied. This ultimately leads to the following likelihood score:

$$\begin{aligned} \nabla_{\mathbf{x}_\tau, \mathbf{n}_\tau} \log p(\mathbf{y} | \mathbf{x}_\tau, \mathbf{n}_\tau) &\approx \nabla_{\mathbf{x}_\tau} \log p(\mathbf{y} | \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_\tau], \mathbb{E}[\mathbf{n}_0 | \mathbf{n}_\tau]) \\ &= \zeta \nabla_{\mathbf{x}_\tau, \mathbf{n}_\tau} \left\| \mathbf{y} - C(\mathbf{A}\mathbf{x}_{\text{RF},0|\tau} + \mathbf{n}_{\text{RF},0|\tau}) \right\|_2^2 \\ &= \zeta \nabla_{\mathbf{x}_\tau, \mathbf{n}_\tau} \left\| \mathbf{y} - C(C^{-1}(\mathbf{A}\mathbf{x}_{0|\tau}) + C^{-1}(\mathbf{n}_{0|\tau})) \right\|_2^2, \end{aligned} \quad (3.7)$$

which can be used in combination with the two priors in (3.5) modeled with diffusion networks $s_\theta(\mathbf{x}_\tau, \tau)$ and $s_\phi(\mathbf{n}_\tau, \tau)$ to perform joint posterior sampling according to the framework in (3.2). Note that we use Tweedie's formula from (2.7) to approximate (3.7) and introduce ζ to group the constants as a result of the derivation of the data consistency term (3.8). These exact steps, including the companding technique, were used to generate the dehazed ultrasound images shown in Fig. 2.

Example 2: Radar interference

As a second case study, we consider the growing problem of *mutual interference* in automotive radar, a domain where model mismatch arises from unpredictable signal interactions in increasingly congested environments. Mutual interference is becoming a major challenge in the

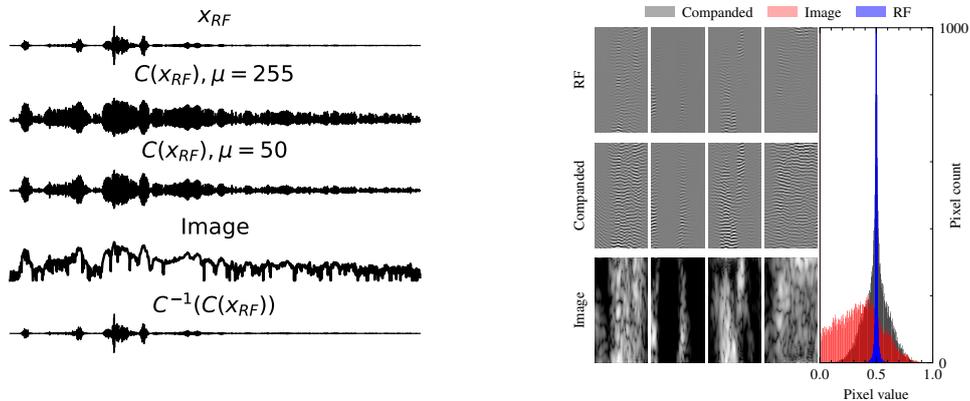


Figure 3: (Left) RF signals and their companded versions, where the μ value is adjusted to align the distribution with that of typical image pixel intensities. (Right) Histogram comparison of RF, companded RF, and ultrasound image data, illustrating the HDR nature of RF signals. Naturally, the companded RF data exhibits a distribution more closely resembling that of image-domain data, facilitating more stable and effective training of DGMs. Figure taken from [32].

automotive radar scene as more vehicles are being equipped with radar sensors, resulting in resource scarcity, i.e., time, frequency, and space [37]. Although the de-facto waveform currently implemented by the sensor manufacturers is **Frequency-Modulated Continuous Wave (FMCW)**, whose chirp-like signals linearly increase or decrease its carrier frequency over time, there is currently no standardization present, leading to a free-for-all situation. A broad range of transmission schemes have been developed and implemented over time, i.e., up-down chirps, stepped-frequency, and frequency-coded waveforms [38,39], resulting in a large variety of radar signals that can interfere with a victim radar. For more information see Box 3.2. This diversity burdens interference removal, causing simple existing mitigation strategies to become less effective. Model-based solutions [40], supervised neural networks [41–44], and model-based deep learning methods [45,46] have been proposed for interference removal. DGMs can learn a large variety of waveforms from training data to effectively remove interference signals; one implementation using posterior sampling is elaborated on below.

Generally, radar-to-radar interference mitigation occurs on the raw data directly, i.e., fast-time data, prior to any post-processing to avoid the interference from leaking into the other dimensions. Therefore, radar interference can be seen as structured noise, \mathbf{n} in (2.1), and leads to model mismatch with the interference being uncorrelated or correlated, ultimately reducing the radar’s sensitivity or creating false positive detections. The authors in [31] have used DGMs by applying score-based diffusion for solving (3.3) and (3.4) to separate the target reflections \mathbf{x} and interference signals \mathbf{n} from the observation \mathbf{y} .

(c) Model-based score function

In contrast to Section 3(a), where two score functions are approximated using deep neural networks (see (3.5)), specific signal properties, e.g. sparsity, can be inferred using model-based priors throughout the diffusion process to reduce complexity. As radar signals are known to be sparse in the range, Doppler and angular domain, the authors of [31] exploit this domain knowledge in a model-based prior. Instead of having a deep neural network for approximating the score, $\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) \approx s_\theta(\mathbf{x}_\tau, \tau)$, as defined in (3.5), an analytical model-based score function is calculated by rewriting (2.7) using Tweedie’s approach. Then, the score function for the target

Box 3.2 – Application: radar

Data characteristics: Every radar is required, as defined in ISO standards, to send updates to the car of the objects present in the sensing environment (typically every 40–100 ms, which includes the sensing time and the processing time of all downstream tasks). Therefore, real-time constraints are put on signal processing and deep learning solutions, such as interference mitigation, direction of arrival estimation, etc. Data rates in automotive radar can range from hundreds of Gbit/s to tens of Gbit/s, where a typical 3-D fast-time data cube $\mathbf{y} \in \mathbb{R}^{C \times M \times N}$ comprises C receive channels, M slow-time samples, and N fast-time samples, respectively. With the recent trend towards high-resolution automotive radars, data cubes can for example grow towards $32 \times 256 \times 1024$, putting more stringent requirements on data rates, memory and computational load in the post-processing stages.

Forward model: \mathbf{y} is the observed interfered signal, \mathbf{x} is the complex-valued sparse signal containing the range information to all object reflections, and \mathbf{n} and ϵ are the interference-only signal and thermal noise, respectively, in the same domain as \mathbf{y} . Therefore, we consider the following forward model: $\mathbf{y} = \mathbf{F}^H \mathbf{x} + \mathbf{n} + \epsilon$, where we are interested in separating \mathbf{x} and \mathbf{n} from \mathbf{y} using a DGM.

Application of DGMs: To mitigate interference, we use a model-based and data-driven score-based DGM to obtain estimates of \mathbf{x} and \mathbf{n} , respectively. It is applied using only N fast-time samples, for all C channels and M slow-time samples independently.

signals at time step τ can be analytically defined as

$$\nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau) = \frac{1}{\sigma_\tau^2} (\alpha_\tau \mathbf{x}_{0|\tau} - \mathbf{x}_\tau) , \quad (3.9)$$

where $\mathbf{x}_{0|\tau}$ denotes the posterior mean, which the authors opt to obtain using the following ℓ_1 -norm minimization by promoting sparsity in \mathbf{x} :

$$\mathbf{x}_{0|\tau} := \arg \min_{\mathbf{x}} \frac{1}{2\sigma_\tau^2} \|\mathbf{x}_\tau - \mathbf{x}\|_2^2 + \lambda_\tau \|\mathbf{x}\|_1 \quad (3.10)$$

$$= \text{prox}_{\lambda_\tau \|\cdot\|_1}(\mathbf{x}_\tau) = \frac{\mathbf{x}_\tau}{\|\mathbf{x}_\tau\|} (|\mathbf{x}_\tau| - \lambda_\tau)_+ . \quad (3.11)$$

Formally speaking, this is known as a denoising step that is readily implemented using soft thresholding as shown in (3.11) with λ_τ being time step-dependent. Additional benefits are that complex-valued score functions are avoided, which are generally hard to implement.

Furthermore, the authors have opted for a data-driven approach for approximating the interference score function $s_\phi(\mathbf{n}_\tau, \tau) \approx \nabla_{\mathbf{n}_\tau} \log p(\mathbf{n}_\tau)$, for which the network has been trained using the denoising score-matching objective of (2.2). As explained earlier, the structured noise of the interference signals is challenging to analytically model due to its large waveform diversity, hence the use of conditional DGMs is favorable due to its generative ability. The distribution of the interference signals is learned in the raw data format directly.

Under the guidance of the likelihood scores, using DPS, estimates of the targets and interference signals $\hat{\mathbf{x}}$ and $\hat{\mathbf{n}}$ are obtained using the joint posterior scores (3.3) and (3.4), respectively:

$$\nabla_{\mathbf{x}_\tau, \mathbf{n}_\tau} \log p(\mathbf{y}|\mathbf{x}_\tau, \mathbf{n}_\tau) \approx \zeta \nabla_{\mathbf{x}_\tau, \mathbf{n}_\tau} \left\| \mathbf{y} - \mathbf{F}^H \mathbf{x}_{0|\tau} - \mathbf{n}_{0|\tau} \right\|_2^2 . \quad (3.12)$$

In Fig. 4 the interference mitigation capabilities are shown for DGMs in a single target scenario for which a large part of the raw data is recovered from interference as shown in 4a, resulting in a reduced interference-induced noise-floor in Fig. 4b. Next, we explain how the methods of

Examples 1 and 2 can be accelerated to enable the application of **DGMs** for real-time ultrasound probing and radar sensing.

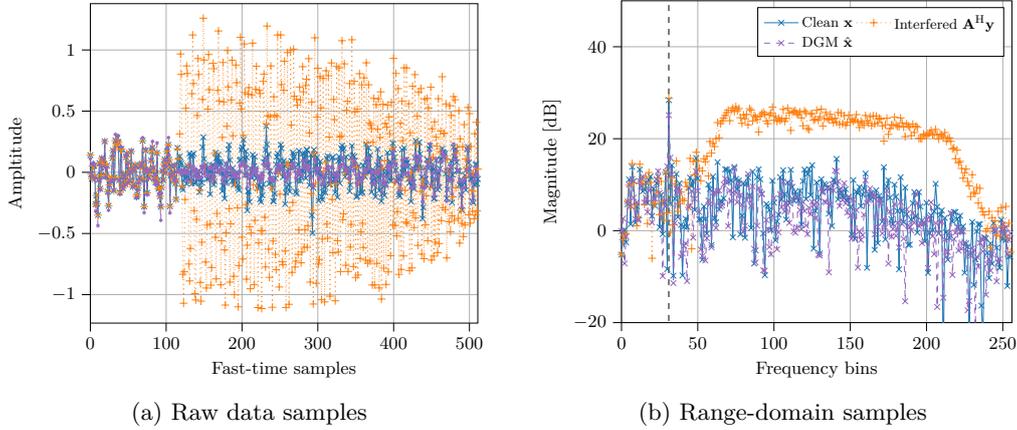


Figure 4: A semi-correlated interference scenario where a large portion of the raw data is contaminated, for which the capabilities of **DGMs** are shown. Figure taken from [31].

4. Real-time and high-data rates

Besides complex noise sources that cause model mismatches and impede image quality, excessive data rates and real-time conditions in sensor systems are another stringent requirement that incentivize efficient sensing and inference algorithms. Nevertheless, while extremely effective, **DGMs** are not generally known for their inference speed. In this section we discuss methods that either address this issue through development of accelerated methods that leverage **DGMs** in some way (Section 4(a)), or try to reduce data rates through active compressed sensing (Section 4(b)).

(a) Acceleration

To bridge the gap between real-time inference of **DGMs** and high data-rates of sensing applications, our initial focus will be on acceleration of current methods. In order to maximize throughput in applications with high data-rates, efficient inference using **DGMs** is essential. Specifically, we will discuss temporal inference (Section 4(a)-i), deep unfolding (Section 4(a)-ii), and knowledge distillation (Section 4(a)-iii).

(i) Temporal inference

The sequential and real-time nature of sensory applications can be both a blessing and curse. While indeed the high throughput of data requires low-latency inference, the temporal axis can be exploited to accelerate inference and even improve reconstruction of the raw sensory signals. One way to accelerate posterior sampling methods using **DMs** as discussed in Section 2(c), is to initialize a given diffusion trajectory conditioned on previous frames, effectively reducing the number of diffusion iterations necessary [47]. Formally, given a set of K diffusion posterior samples of previous frames $\mathbf{x}_0^{t-K:t} = \{\mathbf{x}_0^t, \mathbf{x}_0^{t-1}, \dots, \mathbf{x}_0^{t-K}\}$ we would like to estimate $p(\mathbf{x}^{t+1} | \mathbf{x}_0^{t-K:t})$ with some transition model (such as a **Convolutional LSTM (ConvLSTM)** or **Video Vision Transformer (ViViT)**) such that the number of diffusion steps necessary is minimized with $\tau' \ll \mathcal{T}$. Rather than starting each diffusion trajectory from scratch at $\tau = \mathcal{T}$ with a Gaussian sample $\mathbf{x}_{\mathcal{T}} \sim \mathcal{N}(0, \sigma_{\mathcal{T}}^2 \mathbf{I})$, we use an appropriate estimate $\tilde{\mathbf{x}}$ based on past observations which we diffuse forward up to $\tau = \tau'$ which leads to initialization of a shortened diffusion trajectory: $\mathbf{x}_{\tau'} \sim$

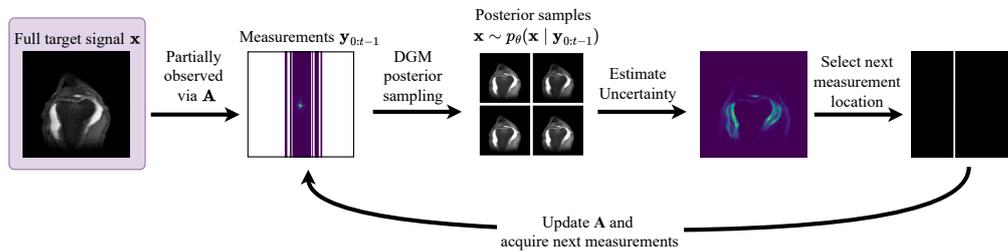


Figure 5: Illustrative example of an active compressed sensing step for MRI acceleration. The sensing matrix $\mathbf{A} = \mathbf{UF}$ consists of a subsampling matrix \mathbf{U} and DFT matrix \mathbf{F} . See *fastMRI* [1] for more information about MRI acceleration.

$\mathcal{N}(\alpha_T \tilde{\mathbf{x}}, \sigma_T^2 \mathbf{I})$. As shown in [47], this reduces inference times for cardiac ultrasound imaging using DMs by 25 \times . Although in this case \mathbf{x} represents B-mode images, applying sequential posterior sampling to raw sensory data could offer even greater advantages, which we consider an intriguing direction for future work.

(ii) Deep unfolding

Deep unfolding (or deep unrolling) is a method that utilizes iterative model-based algorithms, such as proximal-gradient methods, in combination with neural networks to solve inverse problems. By unrolling the iterative optimization algorithm as a feed forward network it takes the structure of the iterations and allows for learning the parameters of the algorithm in successive steps. This will essentially apply multiple iterations in a single forward pass, thus accelerating the iterative algorithm at the cost of higher memory usage. Deep unfolding has been used in many real-world inverse problems such as sparse-coding [48], sub-Nyquist sampling [49] and medical imaging [50] but typically rely on discriminative networks. While there are prior works that combine deep unfolding with generative models [51,52], there are none for high-data rate sensing applications, which could be a fruitful avenue to explore.

(iii) Knowledge distillation

Another powerful method to decrease inference time for efficient inference of DGMs is *knowledge distillation*, in which a new *student model* is trained to produce the same outputs as the original generative model using orders of magnitude fewer parameters. Knowledge distillation has been successfully applied to accelerate inference with both GANs [53] and DMs [54], among other architectures.

(b) Active compressed sensing

While DGMs have shown to produce excellent solutions to highly ill-posed CS problems in many domains, e.g., medical imaging [55,56], we focus here on *active CS* using generative models. Active CS [57] aims to sequentially design the sensing matrix \mathbf{A} in real-time as the measurements are acquired, further compressing the required measurement vector and therefore of particular interest in applications with high data rates. Active CS algorithms are considered ‘active’ in the sense that they iteratively choose which measurements \mathbf{y}_t to acquire next based on the measurements $\mathbf{y}_{0:t-1}$ they have observed so far. In the case of subsampling, for example, this involves choosing which measurement locations, e.g. pixels, time instances, antennas or other degrees-of-freedom, to sample in order to optimally reconstruct \mathbf{x} . Active CS has a long history in various sensory applications, with existing works commonly using supervised deep learning [58] and reinforcement learning [59,60] to choose sampling locations. In the following, we review a number of recently proposed approaches to active compressed sensing using DGMs to jointly

guide sampling and reconstruct target signals. In each case, the goal is to minimize uncertainty about \mathbf{x} , as measured by some uncertainty estimate derived from a set of posterior samples $p(\mathbf{x} | \mathbf{y}_{0:t-1})$ generated by the DGM. A challenging aspect in this paradigm is estimating the decrease in uncertainty that will result from choosing a particular sensing design, leading to a variety of distributional and parametric assumptions in the methods discussed.

Sanchez *et al.* [61] propose **Generative Adaptive Sampling (GAS)**, a method for active subsampling in which the regions of the measurement signal with highest predicted variance are sampled next. This variance is computed by generating posterior samples of the full target signal $\mathbf{x} \sim p(\mathbf{x} | \mathbf{y}_{0:t-1})$ and passing them through the measurement model $\mathbf{y}_t = \mathbf{A}\mathbf{x}$ to generate samples of the full measurement signal at time t , yielding $\mathbf{y}_t \sim p(\mathbf{y}_t | \mathbf{y}_{0:t-1})$. The sample variance of this distribution over measurements $\mathbb{V}[\mathbf{A}\mathbf{x} | \mathbf{y}_{0:t-1}]$ can thus be computed directly from the posterior samples \mathbf{x} . The pixel or set of pixels maximizing this variance is then sampled at time t , and the algorithm repeats. Note that because the measurement noise is independent of the sampling location, choosing to sample the measurement location with the highest entropy will result in minimizing uncertainty about \mathbf{x} [62]. Variance is however only proportional to the entropy under certain distributional assumptions, e.g. isotropic Gaussian. Despite this assumption, **GAS** proves to significantly outperform variable-density sampling in reconstructing MNIST images using a Wasserstein **GAN** [63] as the **DGM**.

Van de Camp *et al.* [62] follow a similar approach to active subsampling, instead proposing the use of a **Gaussian Mixture Model (GMM)** to approximate the measurement posterior, i.e. $p(\mathbf{y}_t | \mathbf{y}_{0:t-1}) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{N}(\mathbf{y}_t | \mathbf{y}_{0:t-1}, \mathbf{I}\sigma^2)$, where N_s is the number of posterior samples. The measurement locations with maximum entropy are then selected to be sampled next, estimating the **GMM** entropy via an approximation introduced by [64]. This entropy approximation is a function of the L2 distances between pixels across all pairs of posterior samples, leading to a regions with high 'disagreement' among samples being assigned high entropy. Van de Camp *et al.* validate their sampling pipeline using two combinations of generative modeling architecture and posterior sampling technique to sample from $p(\mathbf{y}_t | \mathbf{y}_{0:t-1})$. In particular, they use (i) a **VAE** trained on MNIST [65] images, with Markov Chain Monte Carlo used to produce posterior samples in the latent space, which are decoded to generate full measurement samples $\mathbf{y}_t | \mathbf{y}_{0:t-1}$, and (ii) a **GAN** trained on **MRI** images from the *fastMRI* [1] dataset, with annealed Langevin dynamics [23] for posterior sampling.

Elata *et al.* [66] propose *AdaSense*, an adaptive compressed sensing method using a **Denosing Diffusion Restoration Model (DDRM)** [67] for posterior sampling. They propose using the **Mean Squared Error (MSE)** attained by the linear **Minimum Mean Squared Error (MMSE)** predictor as a measure of uncertainty. In the case where the values in \mathbf{A} can be freely designed, *AdaSense* uses the principle components of the posterior covariance as the rows of \mathbf{A} , leveraging that the principle components of the data covariance produce the linear **MMSE** predictor, thus minimizing uncertainty about \mathbf{x} . The algorithm proceeds iteratively by acquiring measurements \mathbf{y} using \mathbf{A} , generating new posterior samples, and then adding the top r principle components as new rows to \mathbf{A} , and repeating. Low values for r then result in highly adaptive sampling, and vice versa for higher r values. In many real world applications, however, \mathbf{A} is constrained by the measurement process, and may not be freely designed. In these cases, *AdaSense* incorporates a new objective, aiming to minimize the linear **MMSE** when \mathbf{A} is constrained to a set of possible sensing matrices \mathcal{A} , leading to the objective $\arg \max_{\mathbf{A} \in \mathcal{A}} \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x} | \mathbf{y}_{0:t-1}])^\top \mathbf{A}^\dagger \mathbf{A} (\mathbf{x} - \mathbb{E}[\mathbf{x} | \mathbf{y}_{0:t-1}]) | \mathbf{y}_{0:t-1}]$. Here, \mathbf{A}^\dagger is the Moore-Penrose pseudo-inverse of \mathbf{A} . For example, in **MRI** acceleration, the set \mathcal{A} might consist of all possible next masks, where each next mask adds a new k-space line. *AdaSense* is validated on **MRI** acceleration and natural image reconstruction tasks.

A limiting factor that is faced when using **DMs** to perform active sampling is the number of **Neural Function Evaluations (NFEs)** necessary to perform posterior sampling, due to the iterative nature of the reverse diffusion process. This may be in the range of hundreds to thousands for high quality image generation. Running an entire reverse diffusion process for each active sampling step may thus be infeasible for applications with high sampling rates. Nolan *et*

al. [68] address this problem with **Active Diffusion Subsampling (ADS)**, which performs $K < T$ active sampling steps in a single reverse diffusion process consisting of T steps, resulting in a significant speedup for applications with large K . **ADS** uses **DPS** [28] as its posterior sampling engine, tracking an estimate of the posterior throughout the reverse diffusion process, and using it to select new measurement locations as it goes. This estimate of the posterior is computed using a set of N_s partially denoised samples $\{\mathbf{x}_\tau^{(i)} | \mathbf{y}_{0:t-1}\}_{i=0}^{N_s}$ at reverse diffusion step τ , from which fully-denoised samples $\{\mathbf{x}_{0|\tau}^{(i)} | \mathbf{y}_{0:t-1}\}_{i=0}^{N_s}$ are computed via Tweedie's formula. **ADS** uses the **GMM**-based approximation for $p(\mathbf{y}_t | \mathbf{y}_{0:t-1})$ proposed by Van de Camp *et al.* [62] to select maximum entropy sampling locations, validating the method on **MRI** acceleration as well as natural image subsampling.

Box 4.1 – Application: MRI

Data characteristics: Of particular relevance regarding the data rate in **MRI** is the **Repetition Time (TR)**, which measures the amount of time between successive pulse sequences on the same slice, and therefore determines the acquisition time for a single **MRI** slice. **TR** tends to range from hundreds to thousands of milliseconds. In the popular *fastMRI* [1], for example, knee slices are acquired using **TRs** in the range 2200-3000ms.

Forward model: A typical **MRI** scan images a 3D volume consisting of a set of stacked 2D *slices*. **MRI** measurements are taken in the Fourier-space representation of the image, referred to as the *k-space*, following the model $\mathbf{y} = \mathbf{F}\mathbf{x} + \epsilon$, where ϵ is measurement noise, \mathbf{x} is the target image, and \mathbf{y} are the *k-space* measurements. *k-space* measurements are typically acquired by a series of *pulse sequences* for each slice, each of which provides a single line of points in the *k-space*. Given a full set of these *k-space* lines, the image can be recovered by the inverse Fourier transform. In accelerated **MRI**, the *k-space* lines are *subsampled*, leading to the model $\mathbf{y}_{0:t} = \mathbf{U}\mathbf{F}\mathbf{x} + \epsilon$, where $\mathbf{y}_{0:t}$ are the measurements selected until time t by subsampling matrix $\mathbf{U} \in \mathbb{R}^{N \times M}$, i.e. $\mathbf{A} = \mathbf{U}\mathbf{F}$, with the goal of reconstructing the full image \mathbf{x} .

Application of DGMs: For **MRI** reconstruction **DGMs** are typically fit on fully-observed target images \mathbf{x} . The **DGM** may, for example, be fit on complex-valued images $\mathbf{x} \in \mathbb{C}^M$ [68], real images with a zeroed imaginary component [66], or other variants [58]. Then, given such a **DGM**, posterior sampling algorithms may be employed to recover full images $\mathbf{x} \sim p(\mathbf{x} | \mathbf{y})$ from a subsampled *k-space* \mathbf{y} .

Example 3: Accelerated MRI

As the final application, we consider accelerated **MRI**, a well-established use case for active **CS**. For more context, see Box 4.1. Due to the relatively slow acquisition time (**TR**) in **MRI**, inference time for popular **DGM** architectures falls within real-time ranges, particularly when leveraging estimates of the posterior as in **ADS** or fast sampling algorithms such as **DDRM** as in *AdaSense*. The active **CS** methods thus iteratively acquire *k-space* lines and update the subsampling matrix so as to select maximally informative next measurements, leading finally to a posterior distribution over fully reconstructed **MRI** images given the entire set of acquired measurements. This process is illustrated in Figure 5.

It remains an open challenge to accelerate such methods further to achieve real-time active **CS** in domains with shorter acquisition times, such as ultrasound imaging, which may require a posterior inference time on the order of tens of milliseconds. However, with a combination of algorithmic and hardware advancements, this may soon be achievable.

5. Conclusion

Deep generative models are increasingly used to tackle problems involving high dimensional data. However, their integration with active array sensing applications poses unique challenges due to the need for real-time processing of the complex and dynamic nature of sensory data. Despite these hurdles, the potential gains of using of **DGMs** to enhance signal reconstruction by accurately modeling the underlying sensory data is significant. In this work, we highlight several works that aspire to close the gap between current **DGM** capabilities and the demanding requirements of sensing applications, with the focus on two key areas: mitigating model mismatch through modeling of structured noise and addressing high-data rates and real-time processing through reduced inference times and active compressed sensing techniques. To this end, we showcase several illustrative applications that effectively apply **DGMs** for relevant problems in the domains of medical imaging and automotive radar. While the methods discussed in this review make substantial progress towards enabling real-time inference with **DGMs**, significant challenges remain in addressing the complexities of sensory data. Notably, applications involving extremely high data rates, such as 3D ultrasound and real-time radar systems, are not yet fully explored. To avoid a latency increase in the aforementioned real-time sensing applications, future work should optimize current methods by finding a balance between the powerful modeling capabilities of **DGMs** and the deployed acceleration techniques.

References

1. Zbontar J, Knoll F, Sriram A, Murrell T, Huang Z, Muckley MJ, Defazio A, Stern R, Johnson P, Bruno M et al.. 2018 fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*.
2. Van Sloun RJ, Cohen R, Eldar YC. 2019 Deep learning in ultrasound imaging. *Proceedings of the IEEE* **108**, 11–29.
3. Doris K, Filippi A, Jansen F. 2022 Reframing Fast-Chirp FMCW Transceivers for Future Automotive Radar: The pathway to higher resolution. *IEEE Solid-State Circuits Magazine* **14**, 44–55.
4. Chung H, Ye JC. 2022 Score-based diffusion models for accelerated MRI. *Medical image analysis* **80**, 102479.
5. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. 2022 High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 10684–10695.
6. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A et al.. 2024 The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
7. Candes EJ, Tao T. 2006 Near-optimal signal recovery from random projections: Universal encoding strategies?. *IEEE transactions on information theory* **52**, 5406–5425.
8. Eldar YC, Kutyniok G. 2012 *Compressed sensing: theory and applications*. Cambridge university press.
9. Eldar YC. 2015 *Sampling theory: Beyond bandlimited systems*. Cambridge University Press.
10. Rani M, Dhok SB, Deshmukh RB. 2018 A systematic review of compressive sensing: Concepts, implementations and applications. *IEEE access* **6**, 4875–4894.
11. Chernyakova T, Eldar YC. 2014 Fourier-domain beamforming: the path to compressed ultrasound imaging. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* **61**, 1252–1267.
12. Ramkumar A, Thittai AK. 2019 Strategic undersampling and recovery using compressed sensing for enhancing ultrasound image quality. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **67**, 547–556.
13. Huijben IA, Veeling BS, Janse K, Mischi M, van Sloun RJ. 2020 Learning sub-sampling and signal recovery with applications in ultrasound imaging. *IEEE Transactions on Medical Imaging* **39**, 3955–3966.
14. Nguyen LH, Tran TD. 2011 Robust recovery of synthetic aperture radar data from uniformly under-sampled measurements. In *2011 IEEE International Geoscience and Remote Sensing Symposium* pp. 3554–3557. IEEE.

15. Cohen D, Eldar YC. 2018 Sub-Nyquist radar systems: Temporal, spectral, and spatial compression. *IEEE Signal Processing Magazine* **35**, 35–58.
16. Ye JC. 2019 Compressed sensing MRI: a review from signal processing perspective. *BMC Biomedical Engineering* **1**, 8.
17. Shlezinger N, Whang J, Eldar YC, Dimakis AG. 2023 Model-Based Deep Learning. *Proceedings of the IEEE* **111**, 465–499.
18. Gorban AN, Tyukin IY. 2018 Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**, 20170237.
19. Kingma DP. 2013 Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014 Generative adversarial nets. *Advances in neural information processing systems* **27**.
21. Kingma DP, Dhariwal P. 2018 Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems* **31**.
22. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. 2020 Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
23. Song Y, Ermon S. 2019 Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32**.
24. Bendel M, Ahmad R, Schniter P. 2024 A regularized conditional GAN for posterior sampling in image recovery problems. *Advances in Neural Information Processing Systems* **36**.
25. Song J, Vahdat A, Mardani M, Kautz J. 2023 Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*.
26. Mardani M, Song J, Kautz J, Vahdat A. 2023 A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*.
27. Daras G, Chung H, Lai CH, Mitsufuji Y, Ye JC, Milanfar P, Dimakis AG, Delbracio M. 2024 A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*.
28. Chung H, Kim J, Mccann MT, Klasky ML, Ye JC. 2022 Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.
29. Efron B. 2011 Tweedie's formula and selection bias. *Journal of the American Statistical Association* **106**, 1602–1614.
30. Milanfar P, Delbracio M. 2024 Denoising: A powerful building-block for imaging, inverse problems, and machine learning. *arXiv preprint arXiv:2409.06219*.
31. Overdeest J, Wei X, van Gorp H, van Sloun RJG. 2024 Model-Based Diffusion for Mitigating Automotive Radar Interference. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* pp. 284–288.
32. Stevens TSW, Meral FC, Yu J, Apostolakis IZ, Robert JL, Van Sloun RJ. 2024 Dehazing Ultrasound using Diffusion Models. *IEEE Transactions on Medical Imaging* pp. 1–1.
33. Stevens TSW, van Gorp H, Meral FC, Shin J, Yu J, Robert JL, van Sloun RJ. 2025 Removing Structured Noise using Diffusion Models. *Transactions on Machine Learning Research*.
34. Drori O, Mamistvalov A, Solomon O, Eldar YC. 2021 Compressed ultrasound imaging: From sub-Nyquist rates to super resolution. *IEEE BITS the information theory magazine* **1**, 27–44.
35. Rodriguez-Molares A, Rindal OMH, D'hooge J, Måsøy SE, Austeng A, Bell MAL, Torp H. 2019 The generalized contrast-to-noise ratio: A formal definition for lesion detectability. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* **67**, 745–759.
36. Azar E, Mulleti S, Eldar YC. 2025 Unlimited sampling beyond modulo. *Applied and Computational Harmonic Analysis* **74**, 101715.
37. Waldschmidt C, Hasch J, Menzel W. 2021 Automotive Radar — From First Efforts to Future Systems. *IEEE Journal of Microwaves* **1**, 135–148.
38. Lulu A, Mobasser BG. 2016 Chirp diversity waveform design and detection by stretch processing. In *2016 IEEE Radar Conference (RadarConf)* pp. 1–6.
39. McGroary F, Lindell K. 1991 A stepped chirp technique for range resolution enhancement. In *NTC '91 - National Telesystems Conference Proceedings* pp. 121–126.
40. Bechter J, Roos F, Rahman M, Waldschmidt C. 2017 Automotive radar interference mitigation using a sparse sampling approach. In *2017 European Radar Conference (EURAD)* pp. 90–93.
41. Mun J, Ha S, Lee J. 2020 Automotive Radar Signal Interference Mitigation Using RNN with Self Attention. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 3802–3806.

42. Fuchs J, Dubey A, Lübke M, Weigel R, Lurz F. 2020 Automotive Radar Interference Mitigation using a Convolutional Autoencoder. In *2020 IEEE International Radar Conference (RADAR)* pp. 315–320.
43. Ristea NC, Anghel A, Ionescu RT. 2021 Estimating the Magnitude and Phase of Automotive Radar Signals Under Multiple Interference Sources With Fully Convolutional Networks. *IEEE Access* **9**, 153491–153507.
44. Fuchs A, Rock J, Toth M, Meissner P, Pernkopf F. 2021 Complex-valued convolutional neural networks for enhanced radar signal denoising and interference mitigation. In *2021 IEEE Radar Conference (RadarConf21)* pp. 1–6. IEEE.
45. Overdevest J, Koppelaar A, Bekooij M, Youn J, Sloun Rv. 2023 Signal Reconstruction for FMCW Radar Interference Mitigation Using Deep Unfolding. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 1–5.
46. Overdevest J, Koppelaar AGC, Youn J, Wei X, Sloun RJGv. 2024 Neurally Augmented Deep Unfolding for Automotive Radar Interference Mitigation. *IEEE Transactions on Radar Systems* **2**, 712–724.
47. Stevens TSW, Nolan O, Robert JL, Van Sloun RJ. 2025 Sequential Posterior Sampling with Diffusion Models. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* Hyderabad, India.
48. Gregor K, LeCun Y. 2010 Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning ICML'10* p. 399–406 Madison, WI, USA. Omnipress.
49. Mulleti S, Zhang H, Eldar YC. 2023 Learning to sample: Data-driven sampling and reconstruction of FRI signals. *IEEE Access* **11**, 71048–71062.
50. Li Y, Bar-Shira O, Monga V, Eldar YC. 2023 p. 53–86. In *Deep Algorithm Unrolling for Biomedical Imaging*, p. 53–86. Cambridge University Press.
51. Wei X, van Gorp H, Gonzalez-Carabarin L, Freedman D, Eldar YC, van Sloun RJG. 2022 Deep Unfolding With Normalizing Flow Priors for Inverse Problems. *IEEE Transactions on Signal Processing* **70**, 2962–2971.
52. Metz L, Poole B, Pfau D, Sohl-Dickstein J. 2017 Unrolled Generative Adversarial Networks. *arXiv preprint arXiv:1611.02163*.
53. Aguinaldo A, Chiang PY, Gain A, Patil A, Pearson K, Feizi S. 2019 Compressing gans using knowledge distillation. *arXiv preprint arXiv:1902.00159*.
54. Salimans T, Ho J. 2022 Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.
55. Song Y, Shen L, Xing L, Ermon S. 2021 Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*.
56. Narnhofer D, Hammernik K, Knoll F, Pock T. 2019 Inverse GANs for accelerated MRI reconstruction. In *Wavelets and Sparsity XVIII* vol. 11138 pp. 381–392. SPIE.
57. Braun G, Pokutta S, Xie Y. 2015 Info-greedy sequential adaptive compressed sensing. *IEEE Journal of selected topics in signal processing* **9**, 601–611.
58. Van Gorp H, Huijben I, Veeling BS, Pezzotti N, Van Sloun RJ. 2021 Active deep probabilistic subsampling. In *International Conference on Machine Learning* pp. 10509–10518. PMLR.
59. Bakker T, van Hoof H, Welling M. 2020 Experimental design for MRI by greedy policy search. *Advances in Neural Information Processing Systems* **33**, 18954–18966.
60. Stevens TSW, Chennakeshava N, de Bruijn FJ, Pekař M, van Sloun RJ. 2022 Accelerated intravascular ultrasound imaging using deep reinforcement learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 1216–1220. IEEE.
61. Sanchez T, Krawczuk I, Sun Z, Cevher V. 2020 Uncertainty-driven adaptive sampling via GANs. In *NeurIPS 2020 Workshop on Deep Learning and Inverse Problems*.
62. van de Camp KC, Joudeh H, Antunes DJ, van Sloun RJ. 2023 Active Subsampling Using Deep Generative Models by Maximizing Expected Information Gain. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 1–5. IEEE.
63. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. 2017 Improved training of wasserstein gans. *Advances in neural information processing systems* **30**.
64. Hershey JR, Olsen PA. 2007 Approximating the Kullback Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* vol. 4 pp. IV–317. IEEE.

65. LeCun Y, Bottou L, Bengio Y, Haffner P. 1998 Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324.
66. Elata N, Michaeli T, Elad M. 2024 Adaptive Compressed Sensing with Diffusion-Based Posterior Sampling. *arXiv preprint arXiv:2407.08256*.
67. Kawar B, Elad M, Ermon S, Song J. 2022 Denoising diffusion restoration models. *Advances in Neural Information Processing Systems* **35**, 23593–23606.
68. Nolan O, Stevens TSW, van Nierop WL, Sloun RV. 2025 Active Diffusion Subsampling. *Transactions on Machine Learning Research*.