Listening to the Noise: Blind Denoising with Gibbs Diffusion

David Heurtel-Depeiges [1] **Charles C. Margossian** [2] **Ruben Ohana** [2] **Bruno Régaldo-Saint Blancard** [2]

Abstract

In recent years, denoising problems have become intertwined with the development of deep generative models. In particular, diffusion models are trained like denoisers, and the distribution they model coincide with denoising priors in the Bayesian picture. However, denoising through diffusion-based posterior sampling requires the noise level and covariance to be known, preventing *blind denoising*. We overcome this limitation by introducing Gibbs Diffusion (GDiff), a general methodology addressing posterior sampling of both the signal and the noise parameters. Assuming arbitrary parametric Gaussian noise, we develop a Gibbs algorithm that alternates sampling steps from a conditional diffusion model trained to map the signal prior to the family of noise distributions, and a Monte Carlo sampler to infer the noise parameters. Our theoretical analysis highlights potential pitfalls, guides diagnostic usage, and quantifies errors in the Gibbs stationary distribution caused by the diffusion model. We showcase our method for 1) blind denoising of natural images involving colored noises with unknown amplitude and spectral index, and 2) a cosmology problem, namely the analysis of cosmic microwave background data, where Bayesian inference of "noise" parameters means constraining models of the evolution of the Universe.

1. Introduction

Denoising is an old problem in signal processing, which has experienced significant advancements in the last decade, propelled by the advent of deep learning (Tian et al., 2020; Elad et al., 2023). Convolutional neural networks have led to the emergence of state-of-the-art image denoisers (e.g., Zhang et al., 2017; 2022), by learning complex prior distributions of target signals implicitly. Denoisers were also found to be powerful tools for generative modeling, as strikingly demonstrated by diffusion models (Ho et al., 2020; Song et al., 2021; Saharia et al., 2022; Rombach et al., 2022). Considerable effort has been devoted to recovering highly accurate, noise-free versions of contaminated signals, often at the expense of fully addressing the noise's complexity. In numerous practical scenarios, both in industry and scientific research, accurately characterizing the noise itself is of paramount importance (e.g., medical imaging, astronomy, speech recognition, financial market analysis). This paper addresses the challenge of *blind denoising*, with the objective of simultaneously recovering both the signal and the noise characteristics.

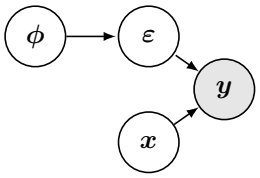

Figure 1. Graphical model: we observe $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}$ and aim to infer both $\boldsymbol{x}$ and $\boldsymbol{\phi}$ in a Bayesian framework.

We formalize the problem as follows. We observe a signal $\boldsymbol{y} \in \mathbb{R}^d$ that is an additive mixture of an arbitrary signal $\boldsymbol{x}$ and a Gaussian signal $\boldsymbol{\varepsilon}$ with covariance $\boldsymbol{\Sigma}_{\boldsymbol{\phi}} \in \mathbb{R}^{d\times d}$:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}}), \tag{1}$$

where $\boldsymbol{\phi} \in \mathbb{R}^K$ is an unknown vector of parameters (see Fig. 1 for a graphical model). The functional form of $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ can be arbitrary, although computational constraints would typically require that the number of parameters K remains reasonable. For example, the vector $\boldsymbol{\phi}$ would typically include a parameter $\sigma > 0$ controlling the overall noise amplitude, but could also encompass parameters describing the local variations or spectral properties of the noise (cf Sect. 3). We frame the problem in a Bayesian picture, where prior information on $\boldsymbol{x}$ and $\boldsymbol{\phi}$ is given, representing some pre-existing knowledge or assumptions on the target data. We will assume that the prior information over $\boldsymbol{\phi}$ takes the form of an analytical prior distribution $p(\boldsymbol{\phi})$. Crucially, although the prior distribution $p(\boldsymbol{x})$ on the signal $\boldsymbol{x}$ is typically analytically intractable, we assume access to a set of

[1]Ecole Polytechnique, Institut Polytechnique de Paris [2]Flatiron Institute. Correspondence to: David Heurtel-Depeiges <david.heurtel-depeiges@polytechnique.edu>, Bruno Régaldo-Saint Blancard <bregaldo@flatironinstitute.org>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

examples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ drawn from $p(\boldsymbol{x})$. In this Bayesian context, solving *blind denoising* amounts to sampling the posterior distribution $p(\boldsymbol{x}, \boldsymbol{\phi} \mid \boldsymbol{y})$.

Contributions. 1) We introduce Gibbs Diffusion (GDiff), a novel approach for blind denoising (Sect. 2). The method combines diffusion models with a Gibbs sampler, tackling simultaneously the challenges of modeling the prior distribution $p(\boldsymbol{x})$ based on the samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, and sampling of the posterior $p(\boldsymbol{x}, \boldsymbol{\phi} \mid \boldsymbol{y})$. 2) We establish conditions for the stationary distribution's existence and quantify inference error propagation. 3) We showcase our method in two areas (Sect. 3). First, we tackle blind denoising of natural images contaminated by arbitrary colored noise, outperforming standard baselines. Second, we demonstrate the interest of our method for cosmology, where Bayesian inference of "noise" parameters corresponds to constraining cosmological parameters of models describing the Universe's evolution. 4) We provide our code on GitHub.[1]

Related work. Blind denoising has garnered substantial attention in the literature, especially in image processing, where denoisers based on convolutional neural networks have become the standard (e.g., Zhang et al., 2017; Batson & Royer, 2019; El Helou & Süsstrunk, 2020). However, these techniques predominantly focus on white Gaussian noises and typically aim for point estimates such as the minimum mean square error or maximum a posteriori estimates, often neglecting the quantification of uncertainties in both the signal and the noise parameters. With the advent of diffusion models, innovative methods for removing structured noise have been developed (Stevens et al., 2023), and posterior sampling has become achievable (e.g., Heurtel-Depeiges et al., 2023; Xie et al., 2023). However, to the best of our knowledge, no existing work has yet fully addressed blind denoising using diffusion-based priors.

Denoising problems can be seen as a subset of the broader category of linear inverse problems, for which observations read $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$ with $\boldsymbol{A}$ a general linear operator. Diffusion models have proven to be of great interest for these problems, generating a wealth of literature adjacent to our problem. In various techniques, diffusion-based priors have been integrated as plug-and-play models (see e.g., Kadkhodaie & Simoncelli, 2021; Meng & Kabashima, 2022; Kawar et al., 2022; Chung et al., 2023; Song et al., 2023a; Zhu et al., 2023; Rout et al., 2023), relying on the decomposition of the conditional score $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{y})$. These methods typically involve approximations of the guiding term $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t)$. In Appendix E, we discuss limitations of some of these methods for our problem. Still employing the diffusion model as a plug-and-play prior but avoiding such approximations, Cardoso et al. (2024) introduced a sequential Monte Carlo method for Bayesian inverse problems. Interestingly, Murata et al. (2023) developed a Gibbs sampler to address the case of a linear operator $\boldsymbol{A}$ depending on unknown parameters $\boldsymbol{\phi}$. While the methodology bears similarities with the approach taken in this paper, it does not seem easily adaptable to the blind denoising problem we address in this work. Finally, let us mention Laroche et al. (2024), which addresses blind deconvolution problems by estimating the maximum a posteriori using an Expectation-Minimization algorithm combined with a guidance-based sampling method.

[1] https://github.com/rubenohana/Gibbs-Diffusion

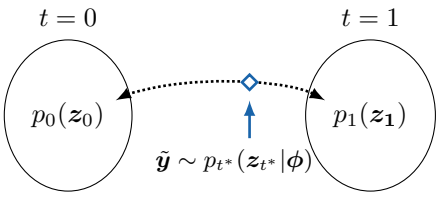

Figure 2. We train diffusion models to define a stochastic linear interpolation between the signal prior distribution $p(\boldsymbol{x})$ and the noise distribution $p(\boldsymbol{\varepsilon}') \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}})$.

2. Method

Our method builds on the ability of diffusion models to perform posterior sampling in denoising contexts. We elaborate on this aspect in Sect. 2.1, a topic we found to be not directly addressed in the existing denoising literature. Then, in Sect. 2.2, we introduce GDiff, a method that addresses posterior sampling of $p(\boldsymbol{x}, \boldsymbol{\phi} \mid \boldsymbol{y})$. Finally, we study theoretical properties of GDiff in Sect. 2.3.

2.1. Diffusion Models as Denoising Posterior Samplers

In a non-blind denoising setting, noise parameters $\boldsymbol{\phi}$ are known. In the Bayesian picture, solving a denoising problem means sampling the posterior distribution $p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})$ for a given prior $p(\boldsymbol{x})$. We show here, that a diffusion model can be naturally trained to both define a prior $p(\boldsymbol{x})$ and yield an efficient means to sample the posterior distribution $p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})$. This idea was leveraged in Heurtel-Depeiges et al. (2023) for a scientific application in cosmology. This was also the concern of Xie et al. (2023).

Diffusion models are a class of deep generative models that are trained to reverse a process consisting in gradually adding noise to the target data. These models are well described using the formalism of stochastic differential equations (SDE, Song et al., 2021).

The noising process is called the forward SDE, and we only

consider here noising Itô processes $(\boldsymbol{z}_t)_{t\in[0,1]}$ defined by:

$$\begin{cases} \mathrm{d}\boldsymbol{z}_t = -f(t)\boldsymbol{z}_t\,\mathrm{d}t + g(t)\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}^{\frac{1}{2}}\,\mathrm{d}\boldsymbol{w}_t, \\ \boldsymbol{z}_0 = \boldsymbol{x}, \end{cases} \tag{2}$$

where $f, g : [0,1] \to \mathbb{R}_+$ are two continuous functions, $(\boldsymbol{w})_{t\in[0,1]}$ is a standard d-dimensional Wiener process, and $\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}$ corresponds to a normalized version of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ introduced in Eq. (1) (e.g., $\boldsymbol{\Sigma}_{\boldsymbol{\varphi}} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}}/\|\boldsymbol{\Sigma}_{\boldsymbol{\phi}}\|$). Within the noise parameters $\boldsymbol{\phi}$, we make a distinction between the noise amplitude σ (i.e. typically $\|\boldsymbol{\Sigma}_{\boldsymbol{\phi}}\|^{1/2}$), and the rest of the parameters represented as $\boldsymbol{\varphi}$. These play distinct roles in the noising process: $\boldsymbol{\varphi}$ modulates the structure of the noise while σ is muddled with time t (see App. A). In this setting, the following proposition applies.

Proposition 2.1. *For a forward SDE $(\boldsymbol{z})_{t\in[0,1]}$ defined by Eq. (2), for all $t \in [0,1]$, $\boldsymbol{z}_t$ reads*

$$\boldsymbol{z}_t = a(t)\boldsymbol{x} + b(t)\boldsymbol{\varepsilon'}, \textit{with } \varepsilon' \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}}), \tag{3}$$

with $a : [0,1] \to \mathbb{R}_+$ a decreasing function with $a(0) = 1$, and $b : [0,1] \to \mathbb{R}_+$ an increasing function with $b(0) = 0$.[2]

Moreoever, for f and g suited so that $\sigma \leq b(1)/a(1)$, there exists $t^* \in [0,1]$ such that

$$\tilde{\boldsymbol{y}} := a(t^*)\boldsymbol{y} \overset{d}{=} \boldsymbol{z}_{t^*}. \tag{4}$$

Proof: See App. A.

Remark: Forward SDEs of denoising probabilistic diffusion models (DDPM, Sohl-Dickstein et al., 2015; Ho et al., 2020) are particular cases of Eq. (2) where $f(t) = \beta(t)/2$ and $g(t) = \sqrt{\beta(t)}$ for an increasing function $\beta : [0,1] \to \mathbb{R}_+$.

Denoting by $p_t(\boldsymbol{z}_t)$ the distribution of $\boldsymbol{z}_t$, Prop. (2.1) presents the forward process $(\boldsymbol{z})_{t\in[0,1]}$ as a stochastic linear interpolant between $p_0(\boldsymbol{z}_0) \sim p(\boldsymbol{x})$ and $p_1(\boldsymbol{z}_1)$ (for a broader perspective, see Albergo et al., 2023). In particular, provided that $\sigma \leq b(1)/a(1)$, our observation $\boldsymbol{y}$ can be viewed as a rescaled realization of $\boldsymbol{z}_{t^*}$ with $t^* \in [0,1]$ (see Fig. 2 for an illustration).

The existence of a reverse process associated with the forward SDE is proved in Anderson (1982). For the forward SDE (2), it takes the form of a process $(\bar{\boldsymbol{z}}_t)_{t\in[0,1]}$ defined by:

$$\begin{cases} \mathrm{d}\bar{\boldsymbol{z}}_t = \left[-f(t)\bar{\boldsymbol{z}}_t - g(t)^2\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}\nabla_{\bar{\boldsymbol{z}}_t}\log p_t(\bar{\boldsymbol{z}}_t)\right]\,\mathrm{d}t \\ \qquad\qquad + g(t)\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}^{\frac{1}{2}}\,\mathrm{d}\overline{\boldsymbol{w}}_t, \\ \bar{\boldsymbol{z}}_1 \overset{d}{=} \boldsymbol{z}_1 \end{cases} \tag{5}$$

where time flows backward and $\overline{\boldsymbol{w}}_t$ denotes a standard backward-time Wiener process. According to Anderson (1982), $(\bar{\boldsymbol{z}}_t)_{t\in[0,1]}$ and $(\boldsymbol{z}_t)_{t\in[0,1]}$ are equal in law. This implies that the marginal distributions of $(\bar{\boldsymbol{z}}_t)$ and $(\boldsymbol{z}_t)$ coincide for all t, but also that the joint distributions $p(\boldsymbol{z}_{t_1}, \boldsymbol{z}_{t_2})$ and $p(\bar{\boldsymbol{z}}_{t_1}, \bar{\boldsymbol{z}}_{t_2})$ coincide for all pair $(t_1, t_2) \in [0,1]^2$. Thanks to this latter fact, sampling $p(\boldsymbol{z}_0 \,|\, \boldsymbol{z}_{t^*})$ can be achieved by solving the reverse SDE (5) starting from time $t = t^*$ and initialization $\bar{\boldsymbol{y}}$. This sampling procedure naturally yields an (approximate) sample of $p(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\phi})$.

[2] See App. A for expressions of a and b as functions of f and g.

Algorithm 1 GDiff: Gibbs Diffusion for Blind Denoising

Input: observation $\boldsymbol{y}$, $p_{\text{init}}(\boldsymbol{\phi})$, M
Initialize $\boldsymbol{\phi}_0 \sim p_{\text{init}}(\boldsymbol{\phi})$.
Initialize $\boldsymbol{x}_0 \sim q(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\phi}_0)$ (Diffusion step)
for $k = 1$ **to** M **do**
  $\boldsymbol{\varepsilon}_{k-1} \leftarrow \boldsymbol{y} - \boldsymbol{x}_{k-1}$
  $\boldsymbol{\phi}_k \sim q(\boldsymbol{\phi} \,|\, \boldsymbol{\varepsilon}_{k-1})$ (HMC step)
  $\boldsymbol{x}_k \sim q(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\phi}_k)$ (Diffusion step)
end for
Output: samples $(\boldsymbol{x}_k, \boldsymbol{\phi}_k)_{1\leq k\leq M}$.

2.2. Blind Denoising with Gibbs Diffusion

We now address the problem of solving blind denoising, where the noise parameters $\boldsymbol{\phi}$ are to be inferred. In our Bayesian picture, the goal is to sample the joint posterior distribution $p(\boldsymbol{x}, \boldsymbol{\phi} \,|\, \boldsymbol{y})$, for given priors $p(\boldsymbol{x})$ and $p(\boldsymbol{\phi})$.

Gibbs Sampling. GDiff takes the form of a Gibbs algorithm that iteratively constructs a Markov chain $(\boldsymbol{x}_k, \boldsymbol{\phi}_k)_{0\leq k\leq M}$ by alternating the sampling between the conditional distributions $p(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\phi})$ and $p(\boldsymbol{\phi} \,|\, \boldsymbol{y}, \boldsymbol{x})$. For ideal sampling, after an initial warm-up phase allowing the system to reach its stationary regime, the iterations of the chains produce samples from the joint distribution $p(\boldsymbol{x}, \boldsymbol{\phi} \,|\, \boldsymbol{y})$.

1. **Sampling of** $p(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\phi})$. Sect. 2.1 has shown that a diffusion model can be trained to naturally address posterior sampling of $p(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{\phi})$ provided that the forward process meets certain conditions. In practice, our forward process must take the form of Eq. (2) and f and g have to be such that $\sigma \leq b(1)/a(1)$, which is always easily achievable (cf Sect. 3). Moreover, since $\boldsymbol{\phi}$ is a priori unknown, we train a conditional diffusion model where $\boldsymbol{\phi}$ is given an an input to the score network, using a range of values consistent with the prior distribution $p(\boldsymbol{\phi})$.

2. **Sampling of** $p(\boldsymbol{\phi} \,|\, \boldsymbol{y}, \boldsymbol{x})$. If $\boldsymbol{y}$ and $\boldsymbol{x}$ are known, then $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{x}$ is known, so that sampling $p(\boldsymbol{\phi} \,|\, \boldsymbol{y}, \boldsymbol{x})$ is equivalent to sampling $p(\boldsymbol{\phi} \,|\, \boldsymbol{\varepsilon})$. We address the sampling of $p(\boldsymbol{\phi} \,|\, \boldsymbol{\varepsilon})$ using a Hamiltonian Monte Carlo (HMC) sampler (see Neal, 2010; Betancourt, 2018,

for a review). Provided we can efficiently evaluate and differentiate the (unnormalized) target distribution $p(\boldsymbol{\phi} \mid \boldsymbol{\varepsilon})$ with respect to $\boldsymbol{\phi}$, a HMC sampler constructs a Markov chain that yields (weakly) correlated samples of the target distribution once the stationary regime is reached. In our case, the log posterior distribution $\log p(\boldsymbol{\phi} \mid \boldsymbol{\varepsilon}) = \log [p(\boldsymbol{\varepsilon} \mid \boldsymbol{\phi})] + p(\boldsymbol{\phi}) + C_{\boldsymbol{\varepsilon}}$ requires the ability to evaluate and differentiate efficiently the prior distribution (assumed to be analytically tractable) and the log likelihood $\log p(\boldsymbol{\varepsilon} \mid \boldsymbol{\phi})$, which reads:

$$\log [p(\boldsymbol{\varepsilon} \mid \boldsymbol{\phi})] = -\frac{1}{2} \log [\det(\boldsymbol{\Sigma}_{\boldsymbol{\phi}})] - \frac{1}{2} \boldsymbol{\varepsilon}^T \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\varepsilon} + C, \tag{6}$$

due to the Gaussianity of ε. This can be easily achieved for the applications considered in Sect. 3. It is important to note, however, that our Gibbs algorithm does not strictly require HMC for this step. Alternative sampling strategies (e.g., neural estimation with normalizing flows) could also be employed and may be more appropriate for other types of applications.

Algorithm and practical details. Algo 1 describes our algorithm. It requires specifying the number of Gibbs iterations M, which includes both the iterations needed for the warm-up phase and the desired number of output samples from the target distribution. Additionally, an initialization strategy for the chains must be defined, represented by the distribution p_{init}. In theory, any strategy of initialization of $\boldsymbol{\phi}$ is viable, as the stationary distribution is independent from p_{init}. A natural choice for this is the prior distribution $p(\boldsymbol{\phi})$. However, in practice, to aid in the convergence of the Gibbs sampler, initial guesses informed by simple heuristics or auxiliary inference strategies might be advantageous.

2.3. Properties of the Gibbs Sampler

Under certain regularity conditions, the ordinary Gibbs sampler converges to the posterior distribution (Geman & Geman, 1984). However, in our setting, as we do not draw samples from the exact conditionals of $\boldsymbol{x}$ and ϕ, additional care must be taken to understand our algorithm's behavior.

Let T be the *transition kernel* of the Gibbs sampler, that is the probability of evolving from a state $(\phi_{k-1}, \boldsymbol{x}_{k-1})$ to a state $(\phi_k, \boldsymbol{x}_k)$. Then the stationary distribution π is the distribution left invariant by T, meaning that if $(\phi_{k-1}, \boldsymbol{x}_{k-1}) \sim \pi$ and we take one step with the Gibbs sampler, then it is also the case that $(\phi_k, \boldsymbol{x}_k) \sim \pi$. Now, T is a composition of two transition kernels: $T_{\boldsymbol{\phi}}$, the HMC step, and $T_{\boldsymbol{x}}$, the diffusion model step.

Proposition 2.2. *The transition kernel $T_{\boldsymbol{\phi}}$ leaves invariant any joint distribution $\pi(\boldsymbol{\phi}, \boldsymbol{x} \mid \boldsymbol{y})$ satisfying $\pi(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{x}) = p(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{x})$, while $T_{\boldsymbol{x}}$ leaves invariant any joint distribution $\tilde{\pi}(\boldsymbol{\phi}, \boldsymbol{x} \mid \boldsymbol{y})$ satisfying $\tilde{\pi}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi}) = q(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})$.*

Proof. The proof is detailed in App. D.

T_{ϕ} has the same invariant distribution, whether we use an HMC step or exact sampling from $p(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{x})$. Therefore, the details of T_{ϕ}, such as the number of HMC steps we take per iteration and the tuning of HMC itself impact convergence rate but not the stationary distribution. On the other hand, using a diffusion model rather than sampling from $p(\boldsymbol{x} \mid \boldsymbol{y}, \phi)$ changes the stationary distribution.

Existence of the stationary distribution. In general, there may not exist a joint distribution over $\boldsymbol{x}$ and ϕ, which matches the conditionals $p(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{x})$ and $q(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})$, in which case we say the conditionals are *incompatible*. Incompatibility can notably arise if $q(\boldsymbol{x} \mid \boldsymbol{y}, \phi)$ poorly approximates the true conditional $p(\boldsymbol{x} \mid \boldsymbol{y}, \phi)$ (e.g., Hobert & Casella, 1998; Liu et al., 2012). We show in App. D that the main necessary condition for compatibility is,

$$\int_{\mathcal{X}} \frac{q(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})}{p(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{x})} \mathrm{d}\boldsymbol{x} < \infty. \tag{7}$$

We adapt this result from Arnold et al. (2001, Theorem 1) and emphasize in our formulation the relationship between the conditionals. Conceptually, Eq. (7) tells us that there can be no measurable set in which $\boldsymbol{x}$ is likely given ϕ but ϕ is unlikely given $\boldsymbol{x}$. If the compatibility condition is verified, then under mild conditions the stationary distribution is unique and we say the two conditionals verify *determinacy*; see Gourieroux & Montfort (1979) and Liu et al. (2021) for a more detailed discussion.

Error in the Gibbs sampler. Assume now compatibility and determinacy. Denote $\pi(\boldsymbol{\phi}, \boldsymbol{x} \mid \boldsymbol{y})$ the unique stationary distribution. How do the approximations in our Gibbs sampler impact our inference on $\boldsymbol{\phi}$? Let $\rho_M(\boldsymbol{\phi} \mid \boldsymbol{y})$ be the distribution of $\boldsymbol{\phi}$ after M sampling iterations and let D_{TV} denote the total variation distance. Then,

$$\begin{aligned} &D_{\text{TV}}[\rho_M(\boldsymbol{\phi} \mid \boldsymbol{y}), p(\boldsymbol{\phi} \mid \boldsymbol{y})] \\ &\leq D_{\text{TV}}[\rho_M(\boldsymbol{\phi} \mid \boldsymbol{y}), \pi(\boldsymbol{\phi} \mid \boldsymbol{y})] + D_{\text{TV}}[\pi(\boldsymbol{\phi} \mid \boldsymbol{y}), p(\boldsymbol{\phi} \mid \boldsymbol{y})]. \end{aligned} \tag{8}$$

The first term on the right-hand-side of Eq. (8) is due to non-convergence. The second term is the error at stationarity due to the diffusion model. We can formally show that, in general, $\pi(\boldsymbol{\phi} \mid \boldsymbol{y}) \neq p(\boldsymbol{\phi} \mid \boldsymbol{y})$ and quantify this error.

Theorem 2.3. *Suppose $p(\boldsymbol{\phi} \mid \boldsymbol{y}, \boldsymbol{x})$ and $q(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})$ are compatible and are the conditionals of the joint $\pi(\boldsymbol{\phi}, \boldsymbol{x} \mid \boldsymbol{y})$. Then the Kullback-Leibler divergence is*

$$KL[p(\boldsymbol{\phi} \mid \boldsymbol{y}) || \pi(\boldsymbol{\phi} \mid \boldsymbol{y})] = \mathbb{E}_{p(\boldsymbol{\phi} \mid \boldsymbol{y})} \log \mathbb{E}_{p(\boldsymbol{x} \mid \boldsymbol{y})} \frac{q(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})}{p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})}.$$

Proof: The proof is given in App. D.

Theorem 2.3 upper-bounds Eq. (8) via Pinsker's inequality,

$$D_{\mathrm{TV}}(p(\boldsymbol{\phi} \mid \boldsymbol{y}), \pi(\boldsymbol{\phi} \mid \boldsymbol{y})) \leq \frac{1}{2}\sqrt{\mathrm{KL}[p(\boldsymbol{\phi} \mid \boldsymbol{y}) || \pi(\boldsymbol{\phi} \mid \boldsymbol{y})]}.$$

Without detailed knowledge of $q(\boldsymbol{x} \mid \boldsymbol{y}, \phi)$ and $p(\boldsymbol{x}, \phi \mid \boldsymbol{y})$, we cannot say more about the error in our Gibbs sampler using theory alone. Compatibility between conditionals has been studied in restricted contexts, for instance when the conditionals belong to certain parametric families (Arnold et al., 2001; Liu et al., 2012), but this does not describe our setting where the conditionals are intractable. Motivated by the variational autoencoder, Liu et al. (2021) recently proposed a loss function (eq. 2 in reference) which, when minimized to 0, enforces compatibility; checking the viability of this strategy in our setting is left as future work. Ergodicity and convergence rates for the Gibbs and HMC samplers are discussed in references (Geman & Geman, 1984; Livingstone et al., 2019) but typically, these rates cannot be computed explicitly.

In practice, we recommend empirically checking convergence and the algorithm's calibration. Convergence diagnostics help us detect if the Gibbs sampler has been run for a sufficient number of iterations and whether a stationary distribution exists. If poor calibration persists after convergence has been detected, then by Theorem 2.3, the diffusion model introduces a non-negligible error in our inference.

3. Applications

We apply GDiff in two independent contexts: 1) blind denoising of natural images considering colored noises with unknown amplitude and spectral index, 2) a cosmological problem, namely the analysis of cosmic microwave background (CMB) data, for which the noise is the signal of cosmological interest and its parameters constrain models of the evolution of the Universe.

Common technical details. In this section, we work with diffusion models relying on the DDPM forward SDE. With the notations of Eq. (2), we take $f(t) = \beta(t)/2$ and $g(t) = \sqrt{\beta(t)}$ with $\beta : [0, 1] \to \mathbb{R}_+$ a linearly increasing function with $\beta(0) = 0$ and $\beta(1)$ calibrated to be higher than the maximum noise level considered in the prior $p(\sigma)$.

Moreover, in both cases, we consider a covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ that is diagonal in Fourier space. We write $\boldsymbol{\Sigma}_{\boldsymbol{\phi}} = \sigma^2 \boldsymbol{F}^T \boldsymbol{D}_{\boldsymbol{\varphi}} \boldsymbol{F}$, where $\boldsymbol{F}$ is the (orthonormal) discrete Fourier transform (DFT) matrix, $\sigma > 0$ controls the noise amplitude and $\boldsymbol{D}_{\boldsymbol{\varphi}}$ is a diagonal matrix parametrized by $\boldsymbol{\varphi} \in \mathbb{R}^{K-1}$. With these notations, $\boldsymbol{\phi}$ represents the pair $(\sigma, \boldsymbol{\varphi})$. We identify $\boldsymbol{D}_{\boldsymbol{\varphi}}$ to a power spectrum function $S_{\boldsymbol{\varphi}}$, in the sense that the diagonal coefficients of $\boldsymbol{D}_{\boldsymbol{\varphi}}$ correspond to the evaluation of $S_{\boldsymbol{\varphi}}$ on a discrete set of Fourier modes $\boldsymbol{k}$. In this context, the expression of the log likelihood $\log p(\boldsymbol{\varepsilon} \mid \boldsymbol{\phi})$ in Eq. (6) takes a simpler form:

$$\log\left[p(\boldsymbol{\varepsilon} \mid \boldsymbol{\phi})\right] = -\frac{1}{2}\sum_{\boldsymbol{k}} \log \sigma^2 S_{\boldsymbol{\varphi}}(\boldsymbol{k}) - \frac{1}{2}\sum_{\boldsymbol{k}} \frac{|\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{k}}|^2}{\sigma^2 S_{\boldsymbol{\varphi}}(\boldsymbol{k})} + C, \tag{9}$$

where $\hat{\boldsymbol{\varepsilon}}$ denotes the DFT of $\boldsymbol{\varepsilon}$. Technical details on the HMC algorithm employed in this section are given in App. B.

3.1. Denoising for Natural Images

Data and setting. We address blind denoising on natural images contaminated by *colored noises*, that it stationary noises with power spectrum following a power law $S_\varphi(\boldsymbol{k}) = |\boldsymbol{k}|^\varphi$ for $k \neq 0$ and $S_\varphi(\boldsymbol{0}) = 1$, with $\varphi \in \mathbb{R}$. We train our diffusion model on the ImageNet data set (Deng et al., 2009; Russakovsky et al., 2015), made of ~ 1.2 M images, generating noises with parameters drawn from a uniform prior $(\sigma, \varphi) \sim p(\boldsymbol{\phi}) = \mathcal{U}([0, 1] \times [-1, 1])$. For the validation of our method, we will also consider a heldout set of ImageNet images, as well as images from the CBSD68 dataset (Martin et al., 2001). The latter dataset will serve as a means to quantify the transfer properties of our model.

Architecture and training. Our diffusion model relies on a U-net architecture with attention layers. It is conditioned by both the time t and parameter φ. This architecture is suited to 256x256 RGB images. Input data is rescaled and cropped accordingly. The diffusion model is trained by minimizing the denoising score matching loss in a discrete time setting with 5,000 time steps. Further technical details are provided in App. C.

Initialization. Algo 1 requires to provide an initialization distribution $p_{\mathrm{init}}(\boldsymbol{\phi})$ to initialize the Markov chains. For the initialization of the spectral index φ, we simply use the prior distribution $\mathcal{U}([-1, 1])$. To initialize σ, we design a simple heuristic based on a linear regression of σ given observations $\boldsymbol{y}$ fitted on a heldout set.

Denoising results. We demonstrate the performance of our GDiff method on a noisy ImageNet test image (Fig. 3) characterized by $\sigma = 0.2$ and $\varphi = 0.4$. We run our Gibbs algorithm on $M = 40$ iterations for 16 chains in parallel. We reach approximate convergence in less than 20 iterations. For our analysis, we consider only the final 20 samples from each chain. The figure presents, on the left, the original noisy image $\boldsymbol{y}$ alongside the true image $\boldsymbol{x}$, a denoised sample $\hat{\boldsymbol{x}}$, and the estimate of the posterior mean $\mathbb{E}\left[\boldsymbol{x} \mid \boldsymbol{y}\right]$ obtained by averaging all retained samples[3]. On the right,

[3]Note that the posterior mean estimate could have been alternatively computed using Tweedie's formula (Robbins, 1956; Efron, 2011).

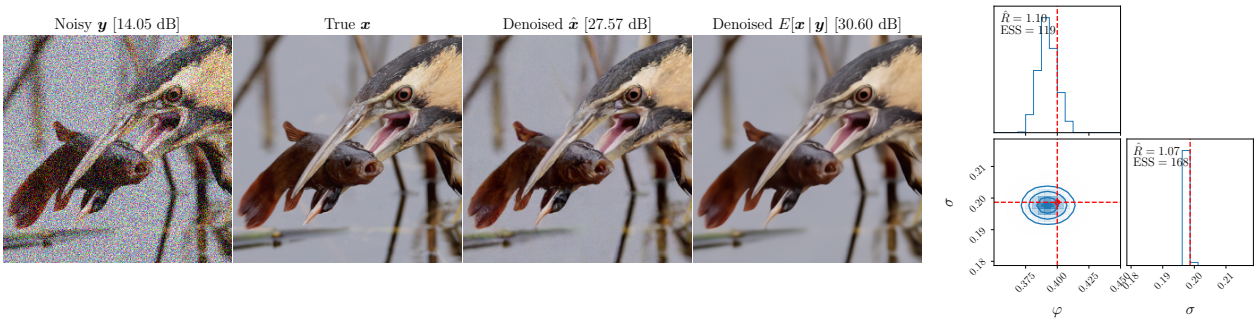

Figure 3. Example of blind denoising with GDiff on an ImageNet sample for $\sigma = 0.2$ and $\varphi = 0.4$. *Left:* Noisy example $\boldsymbol{y}$ next to the noise-free image $\boldsymbol{x}$, a denoised sample $\hat{\boldsymbol{x}}$ and an estimate of the posterior mean $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$ (with PSNR on top when relevant). *Right:* Inferred posterior distribution over the noise parameters.

Dataset	Noise Level σ	$\varphi = -1 \rightarrow$ Pink noise				$\varphi = 0 \rightarrow$ White noise				$\varphi = 1 \rightarrow$ Blue noise			
		BM3D	DnCNN	**GDiff** $\hat{\boldsymbol{x}}$	**GDiff** $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$	BM3D	DnCNN	**GDiff** $\hat{\boldsymbol{x}}$	**GDiff** $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$	BM3D	DnCNN	**GDiff** $\hat{\boldsymbol{x}}$	**GDiff** $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$
ImageNet	0.06	31.0$_{\pm 0.2}$	(30.2)$_{\pm 0.2}$	29.3$_{\pm 0.3}$	**32.2**$_{\pm 0.3}$	33.7$_{\pm 0.3}$	33.4$_{\pm 0.3}$	31.5$_{\pm 0.3}$	**34.4**$_{\pm 0.3}$	34.7$_{\pm 0.4}$	(33.8)$_{\pm 0.4}$	32.3$_{\pm 0.4}$	**35.3**$_{\pm 0.4}$
	0.1	27.8$_{\pm 0.2}$	(26.8)$_{\pm 0.1}$	26.7$_{\pm 0.2}$	**29.4**$_{\pm 0.2}$	31.8$_{\pm 0.3}$	31.8$_{\pm 0.4}$	29.7$_{\pm 0.4}$	**32.7**$_{\pm 0.4}$	32.1$_{\pm 0.4}$	(31.5)$_{\pm 0.3}$	29.9$_{\pm 0.4}$	**32.9**$_{\pm 0.3}$
	0.2	23.5$_{\pm 0.2}$	(21.7)$_{\pm 0.1}$	23.0$_{\pm 0.3}$	**25.7**$_{\pm 0.3}$	28.1$_{\pm 0.4}$	28.4$_{\pm 0.4}$	26.5$_{\pm 0.4}$	**29.3**$_{\pm 0.4}$	29.5$_{\pm 0.4}$	(28.6)$_{\pm 0.4}$	27.6$_{\pm 0.4}$	**30.5**$_{\pm 0.4}$
CBSD68	0.06	31.2$_{\pm 0.2}$	(30.6)$_{\pm 0.1}$	29.2$_{\pm 0.2}$	**32.2**$_{\pm 0.2}$	33.8$_{\pm 0.3}$	34.2$_{\pm 0.3}$	31.2$_{\pm 0.3}$	**34.4**$_{\pm 0.3}$	35.0$_{\pm 0.3}$	(34.8)$_{\pm 0.3}$	32.2$_{\pm 0.3}$	**35.5**$_{\pm 0.3}$
	0.1	27.9$_{\pm 0.2}$	(26.9)$_{\pm 0.1}$	26.2$_{\pm 0.3}$	**29.1**$_{\pm 0.3}$	31.3$_{\pm 0.3}$	31.7$_{\pm 0.3}$	28.6$_{\pm 0.3}$	**31.8**$_{\pm 0.3}$	33.0$_{\pm 0.3}$	(32.7)$_{\pm 0.3}$	30.6$_{\pm 0.4}$	**33.8**$_{\pm 0.4}$
	0.2	23.5$_{\pm 0.2}$	(21.7)$_{\pm 0.1}$	23.0$_{\pm 0.3}$	**25.6**$_{\pm 0.2}$	27.8$_{\pm 0.3}$	28.2$_{\pm 0.3}$	25.4$_{\pm 0.3}$	**28.5**$_{\pm 0.3}$	29.6$_{\pm 0.3}$	(28.9)$_{\pm 0.2}$	27.4$_{\pm 0.3}$	**30.6**$_{\pm 0.3}$

Table 1. Denoising performance in terms of PSNR (↑, in dB) for GDiff (blind) and baselines BM3D (non-blind) and DnCNN (blind). We report mean PSNR and standard error computed on batches of 50 images. For GDiff, we provide performance for both posterior samples $\boldsymbol{x}$ and estimates of the posterior mean $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$. We point out that DnCNN was trained with white noises only, hence results obtained for $\varphi \neq 0$ could be sub-optimal.

the inferred posterior distribution $q(\boldsymbol{\phi} \mid \boldsymbol{y})$ is shown, demonstrating tight constraints around the true parameters and indicative of effective inference. Notably, the denoised posterior sample $\hat{\boldsymbol{x}}$ significantly enhances the image's peak signal-to-noise ratio (PSNR). It is well known that optimal PSNR is attained for the posterior mean $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$ in the Bayesian picture. This is illustrated on this figure where the posterior mean estimate improves the PSNR of the sample by $\sim$3 dB. This is an interesting illustration of the antagonism between PSNR optimization and posterior sampling. The posterior sample is by construction a more realistic reconstruction in the light of the prior. Conversely, while the posterior mean yields a higher PSNR, indicating closer proximity to the true image, it may represent a less probable realization within the prior's distribution. We provide in Fig. F.1 additional denoising examples.

We now benchmark our blind denoiser against established methods, specifically focusing on the PSNR. We compare our performance with that of BM3D (Dabov et al., 2007; Mäkinen et al., 2020) and DnCNN (color-blind model, Zhang et al., 2017) and report our results in Table 1. We show mean performance on subsets of 50 images taken from the ImageNet validation set and CBSD68. It is important to note that the DnCNN model was initially trained for white noise conditions only, hence comparisons for $\varphi \neq 0$ should be interpreted with caution. Interestingly, posterior samples yielded by our method perform worse than BM3D and DnCNN for these metrics. On the contrary, the posterior mean estimates systematically outperform both BM3D and DnCNN. This is a clear success of our algorithm, which in a blind setting can outperform state-of-the-art denoisers. Moreover, we believe that these results could be further improved with additional finetuning of the diffusion model architecture. Finally, we report in Table 2 equivalent results for the structural similarity index measure (SSIM) metric (Wang et al., 2004). Similarly, for this metric, posterior mean estimates provided by GDiff systematically outperform BM3D and DnCNN.

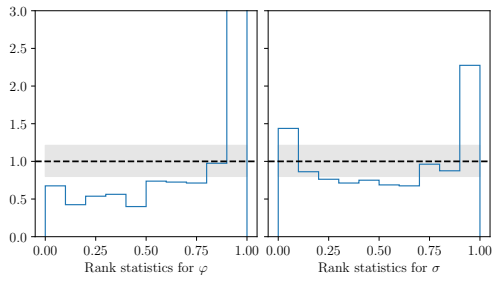

Figure 4. Posterior accuracy diagnostic of GDiff with simulation-based calibration on natural images blind denoising.

Validation. We validate our inference pipeline by focusing on the accuracy of the estimation of the noise parameters $\boldsymbol{\phi} = (\sigma, \varphi)$. We generate 800 noisy images using the ImageNet validation set, with noise parameters $\boldsymbol{\phi}$ sampled from their prior distribution $p(\boldsymbol{\phi})$. For each of the noisy image $\boldsymbol{y}$, we conduct inferences using 4 parallel chains, each running for $M = 60$ iterations. We discard the initial 30 samples as a warm-up phase. For each noisy observation, we compute the $\hat{R}$ convergence diagnostic and the effective sample size (ESS) (see e.g., Vehtari et al., 2021, and references therein). Fig. F.2 provides a scatter plot of these diagnostics across the parameter space. We find that the Markov chains achieve reasonable convergence, as indicated by $\hat{R} \lessapprox 1.1$, and that the ESS per chain is comparable to M.

$\hat{R}$ and ESS tell us how well the Markov chain converges to and explores its stationary distribution, but not how well the stationary distribution approximates the posterior. To check the latter, we implement the simulation-based calibration diagnostic (SBC, Talts et al., 2018). This entails computing the empirical distribution of the ranks of the true parameter values within the sampled values. For a well-calibrated inference pipeline, these empirical distributions should follow a uniform distribution. In Fig. 4, we present the resulting normalized rank statistics. The rank statistics for σ align well with a uniform distribution, suggesting reasonable accuracy. However, the rank statistics for φ reveal some biases in the predictions. As discussed in Sect.2.3, this discrepancy highlights potential limitations of the diffusion model in accurately modeling the true conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi})$.

3.2. Cosmological Inference from CMB Observations

The cosmic microwave background (CMB) is a pivotal cosmological observable for constraining models that describe the Universe's dynamical evolution over its nearly 14 billion-year history (Planck Collaboration I, 2020). Yet, CMB observations suffer from the contamination of various astrophysical signals, known as "foregrounds", necessitating their removal through precise component separation methods (e.g., Planck Collaboration IV, 2020). The quest for primordial B-modes in CMB polarization observations (Kamionkowski & Kovetz, 2016) underscores the challenge of accurately modeling the thermal emission from interstellar dust grains, or the "dust foreground," which obscures the CMB signal (BICEP2/Keck Array and Planck Collaborations, 2015).

The CMB closely approximates a Gaussian distribution, for which the covariance $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ relates to cosmological models and $\boldsymbol{\phi}$ denotes the target cosmological parameters. Interestingly, component separation for CMB analysis can be viewed as a blind denoising problem of the form of Eq. (1) where the "noise" $\boldsymbol{\varepsilon}$ corresponds to the CMB, and the signal $\boldsymbol{x}$ represents the foregrounds. Following a rich history of dust foreground modeling efforts (e.g., Planck Collaboration XI, 2014; Allys et al., 2019; Vansyngel et al., 2017; Aylor et al., 2021; Regaldo-Saint Blancard et al., 2020; 2021; Thorne et al., 2021; Krachmalnicoff & Puglisi, 2021; Régaldo-Saint Blancard et al., 2023; Mudur & Finkbeiner, 2022) and going beyond the work of Heurtel-Depeiges et al. (2023), we focus here on a diffusion-based dust prior trained on simulations and apply GDiff for the Bayesian inference of cosmological parameters $\boldsymbol{\phi}$ given a mixture of CMB $\boldsymbol{\varepsilon}$ and interstellar dust emission $\boldsymbol{x}$.

Data and setting. We introduce the simulated data in App. C.1. The CMB covariance $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ is parametrized by the CMB amplitude σ and cosmological parameters $\boldsymbol{\varphi}$. As in Heurtel-Depeiges et al. (2023), we only consider cosmological parameters $\boldsymbol{\varphi} = (H_0, \omega_b)$. We choose a broad prior $p(\boldsymbol{\varphi}) \sim \mathcal{U}([50, 90] \times [0.0075, 0.0567])$. We also consider $\sigma \in [\sigma_{\min}, 1.2]$ where $\sigma_{\min}$ is taken to be very close to 0.[4]

Architecture and training. The architecture of the diffusion model is a three-level U-net with ResBlocks, trained with a reweighted score matching loss as in Eq.(18). The model is conditioned by both t and parameters $\boldsymbol{\varphi}$. We encountered substantial training instabilities due to the large ($\sim 10^4$) condition number of $\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}$. By renormalizing the loss and model outputs, we were able to mitigate these issues. Further technical details are provided in App. C.

Initialization. For this cosmological application, the initialization of the Gibbs sampler plays a more critical role than in Sect. 3.1. Typically, chains initialized far from posterior typical set are likely to converge slowly, if not become stuck in local minima of the posterior distribution. In order to speed-up convergence and improve sampling efficiency, we initialize $\boldsymbol{\phi}$ using a CNN trained to estimate the posterior mean $\mathbb{E}[\boldsymbol{\phi} \mid \boldsymbol{y}]$ (i.e. *moment network* Jeffrey & Wandelt, 2020). We give further details on this model in App. B.

Cosmological inference. We showcase GDiff for this blind component separation context in Fig. 5 considering a mock mixture $\boldsymbol{y}$ with $\sigma = 0.5$ and $(H_0, \omega_b) = (70, 0.032)$. On the left, we show the true maps $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ composing $\boldsymbol{y}$, next to a pair of reconstructed samples $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{\varepsilon}}$. On the right, we show a corner plot of the posterior distribution on $\boldsymbol{\phi}$ (marginalized over $\boldsymbol{x}$). This inferred posterior distribution imposes tight constraints on the cosmological parameters, accurately encompassing the true parameters (marked in red). We observe that the large scales of the dust map are well reconstructed, as anticipated, due to their lower levels of perturbation. Conversely, at smaller scales,

[4] For stability reasons in the diffusion model, $\sigma_{\min}$ must be strictly positive.

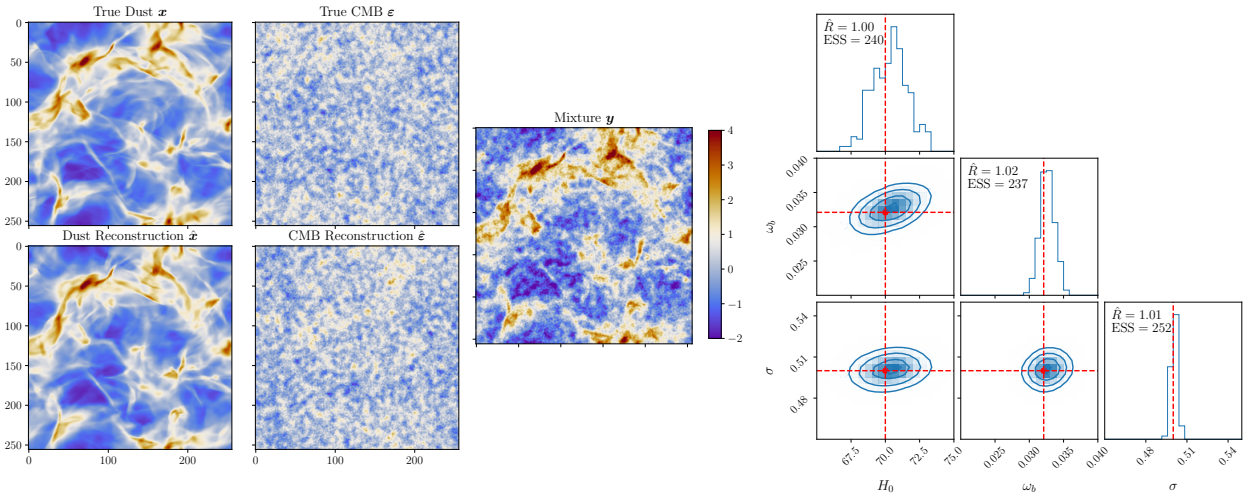

Figure 5. Left: Observed maps and maps reconstructed with GDiff. The true dust $\boldsymbol{x}$ and CMB $\boldsymbol{\varepsilon}$ maps compose the observed mixture $\boldsymbol{y}$. We reconstruct the dust $\hat{\boldsymbol{x}}$ and CMB $\hat{\boldsymbol{\varepsilon}}$ with our diffusion model. The global unit is arbitrary. *Right:* Inferred cosmological parameters.

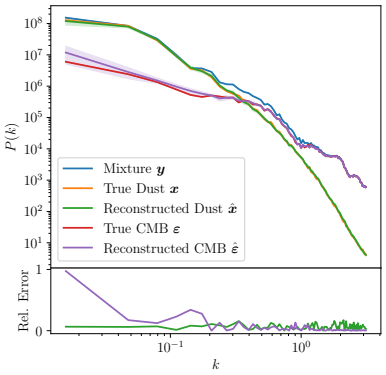

Figure 6. Power spectra of the mixture and the true and reconstructed dust and CMB maps.

the reconstruction process exhibits more stochastic behavior. In the case of the CMB, the situation is reversed: smaller scales are accurately reconstructed, while the reconstruction of larger scales tends to be more stochastic. We complement this visual assessment, by a power spectrum analysis, presented in Fig. 6. This figure displays the power spectra of both the true and reconstructed dust and CMB maps. The agreement between the true and reconstructed statistics across all scales quantitatively demonstrates the success of our reconstruction process.

Validation. We validate our inference pipeline similarly to Sect. 3.1. The $\hat{R}$ and ESS statistics (see Figs. F.3) for test observations generated across the entire parameter space, indicate that our sampler converges and mixes well. In addition, we conduct SBC for the parameters $\boldsymbol{\phi}$ and show in Fig. 7 the resulting rank statistics. These rank distributions are mostly compatible with the uniform distribution, indicating a well calibrated pipeline in particular for the inference of H_0 and ω_b. However, for σ, the distribution is slanted, indicating bias for some of the chains. Here again, this must be the consequence of errors in the diffusion model, which impact the reconstructed power of the CMB estimates.

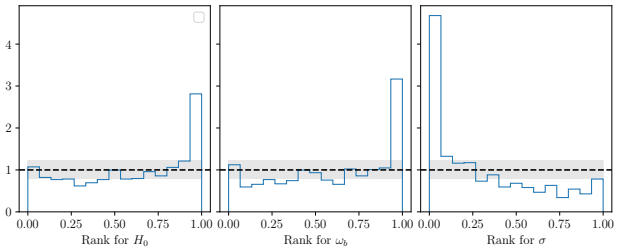

Figure 7. Posterior accuracy diagnostic of GDiff with simulation-based calibration on the cosmological application.

4. Conclusion and Perspectives

We introduced Gibbs Diffusion (GDiff), a new method to address blind denoising in a Bayesian framework. GDiff takes the form of a Gibbs sampler that alternates sampling of the target signal conditioned on the noise parameters and the observation, and sampling of the noise parameters given the estimated noise. We employed diffusion models for the first step, showing that they can be naturally trained to both address signal prior modeling and posterior sampling of the target signal. For the second step, we opted for a Hamiltonian Monte Carlo sampler, although other alternatives could have been considered. We also derived theoretical properties of the Gibbs sampler, pertaining to existence and invariance properties of the stationary distribution, as well

as a quantification of the propagation of errors.

We showcased our method's versatility by applying it to two distinct problems: denoising natural images affected by colored noises with unknown amplitude and spectral index, and performing cosmological inference from simulated cosmic microwave background (CMB) data, which consists of an additive mixture of CMB and a Galactic foreground. Interestingly, while the primary goal in natural image denoising is to recover the noise-free image, the cosmological problem shifts focus to cosmological parameters that characterize the CMB covariance. These parameters assume the role of noise parameters when viewing this task as a blind denoising problem. In both cases, we have shown that our Gibbs sampler converges very well in most situations and constitutes an efficient sampler. Nevertheless, further validation indicated a slight bias due to approximation errors in the diffusion model, suggesting areas for future refinement.

Although diffusion models enable the definition of highly refined prior models, the computational cost of sampling remains significant. Further improvement of our method to accelerate inference would need to take advantage of the recent literature on more efficient ways to address diffusion-based sampling (e.g., Rombach et al., 2022; Song et al., 2023b). Additionally, our method's applicability is currently limited to scenarios where noise follows a Gaussian distribution. Going beyond this Gaussian assumption while maintaining a rigorous and methodologically sound approach remains an open question, with potentially important applications in signal processing and natural sciences.

Acknowledgements

This work benefited from discussions with a number of colleagues whom we wish to acknowledge. In particular, it is a pleasure to thank Yuling Yao for pointing us to a Gibbs sampling approach, as well as Chirag Modi for providing us with a HMC code baseline. We also wish to thank Bob Carpenter and Loucas Pillaud-Vivien for valuable discussions that enriched this work in various ways. Finally, we gratefully acknowledge the Flatiron Institute and the Scientific Computing Core for their support.

Impact Statement

This paper presents work whose goal is to advance the fields of Denoising and Machine Learning for Science. There are potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Allys, E., Levrier, F., Zhang, S., Colling, C., Regaldo-Saint Blancard, B., Boulanger, F., Hennebelle, P., and Mallat, S. The RWST, a comprehensive statistical description of the non-Gaussian structures in the ISM. *Astronomy & Astrophysics*, 629:A115, 2019. doi: 10.1051/0004-6361/201834975.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Arnold, B. C., Castillo, E., and Sarabia, J. M. Conditionally specified distributions: an introduction. *Statistical Science*, 16:249–274, 2001.

Aylor, K., Haq, M., Knox, L., et al. Cleaning our own dust: simulating and separating galactic dust foregrounds with neural networks. *Monthly Notices of the Royal Astronomical Society*, 500(3):3889–3897, 2021. doi: 10.1093/mnras/staa3344.

Batson, J. and Royer, L. Noise2Self: Blind Denoising by Self-Supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533. PMLR, 2019.

Betancourt, M. Nested Sampling with Constrained Hamiltonian Monte Carlo. *AIP Conference Proceedings*, 1305(1):165–172, 03 2011. ISSN 0094-243X. doi: 10.1063/1.3573613.

Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2018.

BICEP2/Keck Array and Planck Collaborations. Joint Analysis of BICEP2/Keck Array and Planck Data. *Physical Review Letters*, 114(10):101301, 2015. doi: 10.1103/PhysRevLett.114.101301.

Burkhart, B., Appel, S., Bialy, S., Cho, J., Christensen, A., Collins, D., Federrath, C., Fielding, D., Finkbeiner, D., Hill, A., et al. The Catalogue for Astrophysical Turbulence Simulations (CATS). *The Astrophysical Journal*, 905(1):14, 2020.

Cardoso, G., el idrissi, Y. J., Moulines, E., and Corff, S. L. Monte Carlo guided Denoising Diffusion models for Bayesian linear inverse problems. In *The Twelfth International Conference on Learning Representations*, 2024.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*, 2023.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8): 2080 – 2095, 2007. ISSN 22195491.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Efron, B. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

El Helou, M. and Süsstrunk, S. Blind Universal Bayesian Image Denoising With Gaussian Noise Level Learning. *IEEE Transactions on Image Processing*, 29:4885–4897, 2020. doi: 10.1109/TIP.2020.2976814.

Elad, M., Kawar, B., and Vaksman, G. Image Denoising: The Deep Learning Revolution and Beyond—A Survey Paper. *SIAM Journal on Imaging Sciences*, 16(3):1594–1654, 2023. doi: 10.1137/23M1545859.

Geman, S. and Geman, D. Stochastic Relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

Gourieroux, C. and Montfort, A. On the characterization of a joint probability distribution by conditional distributions. *Journal of Econometrics*, 10:115–118, 1979.

Heurtel-Depeiges, D., Burkhart, B., Ohana, R., and Régaldo-Saint Blancard, B. Removing Dust from CMB Observations with Diffusion Models. *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2023. doi: 10.48550/arXiv.2310.16285.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hobert, J. and Casella, G. Functional compatibility, Markov chains and Gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics*, 7: 42–60, 1998.

Jeffrey, N. and Wandelt, B. D. Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2020.

Kadkhodaie, Z. and Simoncelli, E. P. Solving Linear Inverse Problems Using the Prior Implicit in a Denoiser. *arXiv preprint arXiv:2007.13640*, 2021.

Kamionkowski, M. and Kovetz, E. D. The quest for b modes from inflationary gravitational waves. *Annual Review of Astronomy and Astrophysics*, 54(1):227–269, 2016. doi: 10.1146/annurev-astro-081915-023433.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Kingma, D. P., Salimans, T., Poole, B., and Ho, J. On Density Estimation with Diffusion Models. In *Advances in Neural Information Processing Systems*, 2021.

Krachmalnicoff, N. and Puglisi, G. ForSE: A GAN-based Algorithm for Extending CMB Foreground Models to Subdegree Angular Scales. *The Astrophysical Journal*, 911(1):42, 2021. doi: 10.3847/1538-4357/abe71c.

Laroche, C., Almansa, A., and Coupeté, E. Fast Diffusion EM: a diffusion model for blind inverse problems with application to deconvolution. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5271–5281, 2024.

Lewis, A., Challinor, A., and Lasenby, A. Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models. *The Astrophysical Journal*, 538(2):473, 2000. doi: 10.1086/309179.

Liu, C., Tang, H., Qin, T., Wang, J., and Liu, T.-Y. On the genervative utility of cyclic conditionals. *Advances in Neural Information Processing Systems*, 2021.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., and Kropko, J. On the Stationary Distribution of Iterative Imputations. *Biometrika*, 101:155–173, 2012.

Livingstone, S., Betancourt, M., Byrne, S., and Girolami, M. On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli*, 25:3109–3138, 2019.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 416–423. IEEE, 2001.

Meng, X. and Kabashima, Y. Diffusion Model Based Posterior Samplng for Noisy Linear Inverse Problems. *arXiv preprint arXiv:2211.12343*, 2022.

Mudur, N. and Finkbeiner, D. P. Can denoising diffusion probabilistic models generate realistic astrophysical fields? *arXiv preprint arXiv:2211.12444*, 2022.

Murata, N., Saito, K., Lai, C.-H., Takida, Y., Uesaka, T., Mitsufuji, Y., and Ermon, S. GibbsDDRM: A partially collapsed Gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25501–25522. PMLR, 2023.

Mäkinen, Y., Azzari, L., and Foi, A. Collaborative Filtering of Correlated Noise: Exact Transform-Domain Variance for Improved Shrinkage and Patch Matching. *IEEE Transactions on Image Processing*, 29:8339–8354, 2020. doi: 10.1109/TIP.2020.3014721.

Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009. doi: 10.1007/s10107-007-0149-x.

Planck Collaboration XI. *Planck* 2013 results. XI. All-sky model of thermal dust emission. *Astronomy & Astrophysics*, 571:A11, 2014. doi: 10.1051/0004-6361/201323195.

Planck Collaboration I. *Planck* 2018 results. I. Overview, and the cosmological legacy of *Planck*. *Astronomy & Astrophysics*, 641:A1, 2020. doi: 10.1051/0004-6361/201833880.

Planck Collaboration IV. *Planck* 2018 results. IV. Diffuse component separation. *Astronomy & Astrophysics*, 641: A4, 2020. doi: 10.1051/0004-6361/201833881.

Regaldo-Saint Blancard, B., Levrier, F., Allys, E., Bellomi, E., and Boulanger, F. Statistical description of dust polarized emission from the diffuse interstellar medium. A RWST approach. *Astronomy & Astrophysics*, 642:A217, 2020. doi: 10.1051/0004-6361/202038044.

Regaldo-Saint Blancard, B., Allys, E., Boulanger, F., Levrier, F., and Jeffrey, N. A new approach for the statistical denoising of Planck interstellar dust polarization data. *Astronomy & Astrophysics*, 649:L18, 2021. doi: 10.1051/0004-6361/202140503.

Régaldo-Saint Blancard, B., Allys, E., Auclair, C., Boulanger, F., Eickenberg, M., Levrier, F., Vacher, L., and Zhang, S. Generative Models of Multichannel Data from a Single Example-Application to Dust Emission. *The Astrophysical Journal*, 943(1):9, 2023. doi: 10.3847/1538-4357/aca538.

Robbins, H. E. An Empirical Bayes Approach to Statistics. 1956.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., and Shakkottai, S. Solving Linear Inverse Problems Provably via Posterior Sampling with Latent Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Gontijo-Lopes, R., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015.

Song, J., Vahdat, A., Mardani, M., and Kautz, J. Pseudoinverse-Guided Diffusion Models for Inverse Problems. In *International Conference on Learning Representations*, 2023a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Stevens, T. S. W., van Gorp, H., Meral, F. C., Shin, J., Yu, J., Robert, J.-L., and van Sloun, R. J. G. Removing Structured Noise with Diffusion Models. *arXiv preprint arXiv:2302.05290*, 2023.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Thorne, B., Knox, L., and Prabhu, K. A generative model of galactic dust emission using variational autoencoders. *Monthly Notices of the Royal Astronomical Society*, 504 (2):2603–2613, 2021. doi: 10.1093/mnras/stab1011.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2020.07.025.

Vansyngel, F., Boulanger, F., Ghosh, T., et al. Statistical simulations of the dust foreground to cosmic microwave background polarization. *Astronomy & Astrophysics*, 603: A62, 2017. doi: 10.1051/0004-6361/201629992.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667 – 718, 2021. doi: 10.1214/20-BA1221.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xie, Y., Yuan, M., Dong, B., and Li, Q. Diffusion Model for Generative Image Denoising. *arXiv preprint arXiv:2302.02398*, 2023.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., and Timofte, R. Plug-and-Play Image Restoration With Deep Denoiser Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2022. doi: 10.1109/TPAMI.2021.3088914.

Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., and Gool, L. V. Denoising Diffusion Models for Plug-and-Play Image Restoration. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (NTIRE)*, 2023.

A. Proof of Prop. 2.1

Let f and g be two positive real valued functions of time t on the interval $[0,1]$. We consider the following SDE:

$$\mathrm{d}\boldsymbol{z}_t = f(t)\boldsymbol{z}_t\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{w}_t, \tag{10}$$

Proposition A.1. *For a forward SDE $(\boldsymbol{z})_{t\in[0,1]}$ defined by Eq. (2), for all $t \in [0,1]$, $\boldsymbol{z}_t$ reads*

$$\boldsymbol{z}_t = a(t)\boldsymbol{x} + b(t)\boldsymbol{\varepsilon'}, \text{with } \varepsilon' \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}}), \tag{11}$$

with $a : [0,1] \to \mathbb{R}_+$ a decreasing function with $a(0) = 1$, and $b : [0,1] \to \mathbb{R}_+$ an increasing function with $b(0) = 0$.[5]

Moreover, for f and g suited so that $\sigma \leq b(1)/a(1)$, there exists $t^* \in [0,1]$ such that

$$\tilde{\boldsymbol{y}} := a(t^*)\boldsymbol{y} \overset{d}{=} \boldsymbol{z}_{t^*}. \tag{12}$$

Proof: We write here the proof when $\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}$ is the identity but the proofs holds for any matrix.

If f can be integrated, let $F(t)$ be a primitive of f with value 0 at time 0. If $\boldsymbol{z}_t$ is an Ito process then so is $\boldsymbol{y}_t = \exp(F(t))\boldsymbol{z}_t$ and we have:

$$\begin{aligned}
\mathrm{d}\boldsymbol{y}_t &= \exp(F(t))\mathrm{d}F(t)\boldsymbol{z}_t + \exp(F(t))\mathrm{d}\boldsymbol{z}_t \\
&= \exp(F(t))f(t)\boldsymbol{z}_t, \mathrm{d}t + -\exp(F(t))f(t)\boldsymbol{z}_t\,\mathrm{d}t + \exp(F(t))g(t)\,\mathrm{d}\boldsymbol{w}_t \\
&= \exp(F(t))g(t)\,\mathrm{d}\boldsymbol{w}_t.
\end{aligned}$$

Therefore, integrating this equation from 0 to t yields:

$$\boldsymbol{y}_t = \boldsymbol{y}_0 + \int_0^t \exp(F(u))g(u)\,\mathrm{d}\boldsymbol{w}_u.$$

And since $\boldsymbol{z}_t = \exp(-F(t))\boldsymbol{y}_t$, we have:

$$\boldsymbol{z}_t = \exp(-F(t))\boldsymbol{y}_0 + \int_0^t \exp(F(u) - F(t))g(u)\,\mathrm{d}\boldsymbol{w}_u.$$

Using the fact that $F(0) = 0$, we obtain:

$$\boldsymbol{z}_t = \exp(-F(t))\boldsymbol{z}_0 + \int_0^t \exp(F(u) - F(t))g(u)\,\mathrm{d}\boldsymbol{w}_u. \tag{13}$$

Finally, we have:

$$\int_0^t \exp(F(u) - F(t))g(u)\,\mathrm{d}\boldsymbol{w}_u \sim \mathcal{N}\left(0, \exp(-2F(t))\int_0^t \exp(2F(u))g(u)^2\,\mathrm{d}u\right) \tag{14}$$

$$\boldsymbol{z}_t \sim \mathcal{N}\left(\exp(-F(t))\boldsymbol{z}_0, \exp(-2F(t))\int_0^t \exp(2F(u))g(u)^2\,\mathrm{d}u\right). \tag{15}$$

Denoting $a(t) = \exp -F(t)$ and $b(t)^2 = \exp(-2F(t))\int_0^t \exp(2F(u))g(u)^2\,\mathrm{d}u$, we have:

$$\boldsymbol{z}_t = a(t)\boldsymbol{x} + b(t)\boldsymbol{\varepsilon'}, \text{with } \varepsilon' \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}}). \tag{16}$$

Furthermore, because f is postive, F is increasing and a is a decreasing function of time with $a(0) = 1$. If g is not too small[6] then b is increasing. As a consequence, the amount of noise in the mixture $\boldsymbol{z}_t$ is $b(t)/a(t)$, itself an increasing function of time and $\forall \sigma \leq b(1)/a(1), \exists t^* \in [0,1], a(t^*)\boldsymbol{y} \sim z_{t^*}$.

This is why, for reasonable functions f and g, we can identity any mixture of the form $y = z_0 + \sigma\varepsilon$ with a realisation of $\boldsymbol{z}_t$ at time t for some t and initial condition $\boldsymbol{z}_0$ (this is a stopped forward process, up to a time-dependent rescaling constant). The corresponding time t is often unique for common choices of f and g (both VPSDE, VESDE and all Ornstein-Uhlenbeck like processes used in DM). When tractable, we implement a function $t(\sigma)$ that yields t^* corresponding to σ and otherwise, the equation $\sigma = b(t)/a(t)$ is solved via an iterative method.

[5] Expressions of a and b as functions of f and g are in App. A.

[6] $\forall t, g(t) \geq b(t)\sqrt{2f(t)}$

B. Additional Details on HMC

We give details on the practical implementation of the HMC sampler used for the applications of this paper.

Integrator. During sampling, Hamilton's equations of motion are solved using a leapfrog integrator. Moreover, to improve calibration of the HMC algorithm, for each HMC step, we solve Hamilton's equations over a random number of time steps drawn from a uniform distribution $\mathcal{U}(\{5, \ldots, 15\})$. Since both applications of this paper involve compact prior distributions $p(\boldsymbol{\phi})$, we also had to implement domain constraints in the integrator. We take into account the prior boundaries by implementing elastic collisions in the leapfrog scheme as described in Betancourt (2011).

Initialization Heuristic. For the cosmological application, initialization of the parameters $\boldsymbol{\phi}$ of the Gibbs sampler was found to be more critical than in Sect.3.1. As explained in Sect.3.2, we intialize $\boldsymbol{\phi} = (\boldsymbol{\varphi}, \sigma)$ using a CNN with ResBlocks trained to estimate the posterior mean $\mathbb{E}[\boldsymbol{\phi} \,|\, \boldsymbol{y}]$. In practice, we sample $\boldsymbol{\varphi}, \sigma$ according to their prior distribution and then sample $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}$ with $\boldsymbol{x} \sim p_{\text{data}}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\phi}})$. The network is then trained to minimize the MSE loss $\mathbb{E}_{\boldsymbol{y}}[||\hat{\phi}(\boldsymbol{y}) - \phi||^2]$, thus yielding a posterior mean estimator. We found that this moment network was less accurate on the boundary of our (compact) prior. To avoid initializing chains outside the prior, which would get stuck, all posterior mean estimates are projected within the prior domain.

Warm-up. In the first iteration of the Gibbs sampler (see Algo. 1), we first follow a warm-up phase of $P = 300$ iterations allowing to adapt the step size and estimate a mass matrix. We adapt the step size using a dual averaging procedure (Nesterov, 2009) parametrized by a target acceptance rate of 0.65, a regularization scale $\gamma = 0.05$, an iteration offset $T_0 = 10$, and a relaxation exponent $\kappa = 0.75$. The inverse of the mass matrix is estimated at iteration $\lfloor 3P/4 \rfloor$ using the (unbiased) sample covariance matrix of the warm-up samples after having discarded samples from iterations up to $\lfloor P/4 \rfloor$.

C. Models and Training Details

C.1. Datasets and Training

Natural images application. We train our diffusion model on the ImageNet 2012 dataset, which consists of 1,281,167 training images. Images are resized to $3 \times 256 \times 256$ and centered crop. During training, we augment the data with random horizontal flips and rescale the images to the $[0, 1]$ range.

We train over 100 epochs on a node of 8 H100 GPUs 80GB with a batch size per GPU of 128 images. Training takes about 41 hours using Data Parallelism. We use the AdamW (Loshchilov & Hutter, 2019) optimizer and report good stability of training with respect to learning rate (tested on 0.0001, 0.001, 0.005), with no learning rate schedule. Observing no impact from weight decay, we set its value to 0.

Cosmological application. We use the same data as in Heurtel-Depeiges et al. (2023). We construct simulated dust emission maps in total intensity from a turbulent hydrodynamic simulation of the diffuse interstellar medium taken from the CATS database (Burkhart et al., 2020). We assume here that dust emission is proportional to the gas density of the simulation. Simulated maps then simply correspond to gas column density maps. Our dataset consists of 991 dust emission maps of size 256×256 (for an example, see Fig. 5 top left panel). We withheld 10% of the 991 images to validate the denoising and inference pipeline based on our diffusion model trained on the other 90%.

If we neglect secondary anisotropies, CMB anisotropies are extremely well described by an isotropic Gaussian random field on the sphere (Planck Collaboration I, 2020), entirely characterized by its covariance matrix (or power spectrum). As a consequence, we can obtain CMB maps by sampling Gaussian draws with covariance $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ parametrized by cosmological parameters $\boldsymbol{\varphi}$ (for an example, see Fig. 5 top middle panel). CMB power spectra are computed using `CAMB` (Lewis et al., 2000), and Gaussian draws on the sphere are projected on 256×256 patches using `pixell` [7] with a pixel size of $8'$. In this work, we consider a standard cosmological model. We only vary the cosmological parameters H_0 and ω_b and choose for the remaining ones fiducial values consistent with *Planck* 2018 analysis (Planck Collaboration I, 2020).[8] The resulting simulated maps cover an effective angular surface of approximately $34 \times 34\,\text{deg}^2$.

Training examples are generated by combining dust samples and CMB realizations for cosmological parameters randomly

[7] https://github.com/simonsobs/pixell

[8] We choose $\Omega_K = 0$, $\omega_c = 0.12$, $\tau = 0.0544$, $n_s = 0.9649$, $\ln(10^{10} A_s) = 3.044$, $m_\nu = 0.08$.

drawn from the prior $p(\boldsymbol{\phi})$. Since the computation of a single covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ takes a few seconds, we train neural emulator to approximate $\boldsymbol{\phi} \to \boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ as described in the Appendix of Heurtel-Depeiges et al. (2023). This emulator is differentiable, thus enabling the differentiation of Eq. (9) for HMC.

We train over 100,000 epochs on a single A100 80GB using a batch size of 64 images. Training takes about 96 hours. We use the AdamW optimizer with an inverse square root scheduler with warm-up. We observed that continuing training after the loss had plateaued significantly improved the diffusion model's precision, as evidenced by the simulation-based calibration (SBC) diagnostic.

C.2. Architectures

Natural images application. Our diffusion model is a U-Net, that takes images of size 256x256 as an input. It is composed of 5 Downsampling layers, a bottleneck of size 16x16 and 5 Upsampling layers. We add 2 self-Attention layers before the bottleneck and one after the bottleneck. The total number of parameters is $\sim 70,000,000$.

As it is standard in a U-Net architecture, each Upsampling layer is concatenated with the corresponding Downsampling layer. A sinusoidal time embedding is added to each DownConv and UpConv. A 2 layer neural network of inner dimension 100 and SiLU activations are used to embed the exponent φ of the colored noise to add this information into the network. We choose a variance preserving DDPM with a discrete time composed of $t_r = 5,000$ time-steps with the schedule $\beta(t) = \frac{0.1}{t_r} + \frac{\beta_{max}-\beta_{min}}{t_r}.t$, with $\beta_{min} = 0.1$ and $\beta_{max} = 20$.

Cosmological application. For the cosmological application, our score network is a UNet with ResBlocks and a bottleneck of size 32×32 or 16×16. Each ResBlock has three convolution, with GroupNorm as normalization, SiLU for activation and rescaled skip-connection. Long skip connections are concatenated, contrary to internal skip connections who are added. We do not use attention, due to the small size of the dataset. Since both our data and noise have periodic boundary conditions, we use 'circular' padding in the convolutions. As in (Ho et al., 2020; Song et al., 2021), time is also seen as an input. We use the continuous time framework from (Song et al., 2021) with Fourier embedding.

In order to manage a parameterized family of SDE (Eq. (2)) we train a diffusion model to reverse each SDE independently by also providing φ as an argument to the model. The dependency on φ is managed similarly to that of the time t with the exception that the embedding is a linear layer with activation function of the parameters φ. In each ResBlock, the time embedding and the parameters embedding are then transformed by a small MLP with a depth of 1, prior to being added to the input of the respective ResBlock.

In addition, for the cosmological component separation task only, we trained a model exclusively on low noise levels by reducing the value of $\beta_{\max}$ to 2, as opposed to 20. Consequently, our network is limited to denoising images for σ values within the specified prior range. This adjustment restricts the model's utility as a generative prior but enhances its precision in reconstructing summary statistics and calibrating the inference method. Furthermore, we found that at lower noise-levels, feeding the network Fourier features (as in Kingma et al., 2021) improved calibration of the posterior distribution.

C.3. Loss Functions

Our models are trained by minimizing a modified denoising score-matching loss whose original formulation would be:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\phi}\left[\lambda(t)\mathbb{E}_{\boldsymbol{z}_0,\,\boldsymbol{z}_t|(\boldsymbol{z}_0,\boldsymbol{\phi})}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}(\boldsymbol{z}_t,t) + \boldsymbol{\Sigma}_{\boldsymbol{\phi}}{}^{-1}\left(\boldsymbol{z}_t - \sqrt{\bar{\alpha}(t)}\boldsymbol{z}_0\right)/\left(1-\bar{\alpha}(t)\right)\right\|_2^2\right]\right], \tag{17}$$

with $\bar{\alpha}(t) = \exp(-\int_0^t \beta(u)\mathrm{d}u)$. As in Heurtel-Depeiges et al. (2023), we re-normalize all vectors in this loss by multiplying them by $-\sqrt{1-\bar{\alpha}(t)}\boldsymbol{\Sigma}_{\phi}$ to stabilize the learning process and reduce numerical errors. Consequently, the adjusted loss function read:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\phi}\left\{\lambda(t)\mathbb{E}_{\boldsymbol{z}_0,\,\boldsymbol{z}_t|(\boldsymbol{z}_0,\phi)}\left[\left\|\mathbf{m}_{\boldsymbol{\theta}}(\boldsymbol{z}_t,t,\phi) - \frac{1}{\sqrt{1-\bar{\alpha}(t)}}\left(\boldsymbol{z}_t - \sqrt{\bar{\alpha}(t)}\boldsymbol{z}_0\right)\right\|_2^2\right]\right\} \tag{18}$$

This loss in turns corresponds to the standard loss from Ho et al. (2020) where one tries to estimate the added noise (up to a rescaling constant).

In order to verify that the model is not overfitting, we check for mode collapse and dataset copying with a specific L_2

distance taking into account the periodic boundary conditions. This was particularly important as our models seem more prone to overfitting than model trained under white noise.

D. Properties of the Gibbs sampler

In this section, we derive theoretical properties of the proposed Gibbs sampler. Specifically, we work out the sampler's stationary distribution, review conditions for its existence, and quantify its error.

In what follows, all distributions are conditioned on $\boldsymbol{y}$, and so we drop this dependency to alleviate the notation.

D.1. Formal notation

Any Markov chain Monte Carlo (MCMC) sampler defines a transition kernel between a current state $(\phi, \boldsymbol{x})$ and a new state $(\phi', \boldsymbol{x}')$. We denote this transition $T((\phi, \boldsymbol{x}) \to (\phi', \boldsymbol{x}'))$. The *invariant distribution* π is the distribution that verifies

$$\int T((\phi, \boldsymbol{x}) \to (\phi', \boldsymbol{x}'))\pi(\phi, \boldsymbol{x})\, d\phi \mathrm{d}\boldsymbol{x} = \pi(\phi', \boldsymbol{x}'). \tag{19}$$

In words, if we apply the transition kernel T to a state $(\phi, \boldsymbol{x})$ drawn from π, then the new state $(\phi', \boldsymbol{x}')$ is also distributed according to π.

The transition kernel T of the Gibbs sampler decomposes into two kernels $T = T_\phi \circ T_{\boldsymbol{x}}$. The first transition kernel T_ϕ updates ϕ but maintains $\boldsymbol{x}$ fixed, while $T_{\boldsymbol{x}}$ updates $\boldsymbol{x}$ and leaves ϕ unchanged.

T_ϕ draws ϕ from an MCMC sampler that admits $p(\phi \mid \boldsymbol{x}, \boldsymbol{y})$ as its stationary distribution. In our paper, we use Hamiltonian Monte Carlo but the results we derive here would hold for any MCMC algorithm with the right stationary distribution. The MCMC sampler defines a transition density $\tau(\phi \to \phi' \mid \boldsymbol{x})$, and moreover

$$T_\phi = \tau(\phi \to \phi' \mid \boldsymbol{x})\delta(\boldsymbol{x}' = \boldsymbol{x}), \tag{20}$$

where δ is the dirac delta function. τ may correspond to the transition obtained after applying either a single or multiple iterations of the MCMC sampler.

Next, $T_{\boldsymbol{x}}$ draws samples from a diffusion model that approximates the data generating process $p(\boldsymbol{x} \mid \phi, \boldsymbol{y})$. It what follows, we assume that the distribution of the samples produced by the diffusion model admits a probability density $q(\boldsymbol{x} \mid \phi, \boldsymbol{y})$. Then

$$T_{\boldsymbol{x}} = q(\boldsymbol{x}' \mid \phi)\delta(\phi' = \phi). \tag{21}$$

D.2. Proof of Proposition 2.2: invariant distributions

Proof. Let $\pi(\phi, \boldsymbol{x})$ be any joint distribution, such that $\pi(\phi \mid \boldsymbol{x}) = p(\phi \mid \boldsymbol{x})$. Then

$$\begin{aligned}
\int T_\phi((\phi, \boldsymbol{x}) \to (\phi', \boldsymbol{x}'))\pi(\phi, \boldsymbol{x})\mathrm{d}\phi\mathrm{d}\boldsymbol{x} &= \int \tau(\phi \to \phi' \mid \boldsymbol{x})\delta(\boldsymbol{x}' = \boldsymbol{x})\pi(\phi, \boldsymbol{x})\mathrm{d}\phi\mathrm{d}\boldsymbol{x} \\
&= \int \tau(\phi \to \phi' \mid \boldsymbol{x}')p(\phi \mid \boldsymbol{x}')\pi(\boldsymbol{x}')\mathrm{d}\phi \\
&= \pi(\boldsymbol{x}') \int \tau(\phi \to \phi' \mid \boldsymbol{x}')p(\phi \mid \boldsymbol{x}')\mathrm{d}\phi \\
&= \pi(\boldsymbol{x}')p(\phi \mid \boldsymbol{x}') \\
&= \pi(\phi', \boldsymbol{x}'),
\end{aligned} \tag{22}$$

where for the penultimate line we used the assumption that τ's stationary distribution is $p(\phi \mid \boldsymbol{x})$. From 19, we then have that $\pi(\phi, \boldsymbol{x})$ is an invariant distribution of T_ϕ.

Similarly let $\tilde{\pi}(\phi, \boldsymbol{x})$ be any joint distribution such that $\tilde{\pi}(\boldsymbol{x} \mid \phi) = q(\boldsymbol{x} \mid \phi)$. Then

$$\begin{aligned} \int T_{\boldsymbol{x}}((\phi, \boldsymbol{x}) \to (\phi', \boldsymbol{x}'))\tilde{\pi}(\phi, \boldsymbol{x})\mathrm{d}\phi\mathrm{d}\boldsymbol{x} &= \int q(\boldsymbol{x}' \mid \phi)\delta(\phi' = \phi)\tilde{\pi}(\phi, \boldsymbol{x})\mathrm{d}\phi\mathrm{d}\boldsymbol{x} \\ &= q(\boldsymbol{x}' \mid \phi') \int \tilde{\pi}(\phi', \boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= q(\boldsymbol{x}' \mid \phi')\tilde{\pi}(\phi') \\ &= \tilde{\pi}(\phi', \boldsymbol{x}'). \end{aligned} \tag{23}$$

Hence $\tilde{\pi}(\phi, \boldsymbol{x})$ is invariant under $T_{\boldsymbol{x}}$. □

D.3. Existence of Stationary Distribution

Definition D.1. Consider two families of conditional distributions, $f(\phi \mid \boldsymbol{x})$ and $g(\phi \mid \boldsymbol{x})$. We say f and g are *compatible* if there exists a joint distribution $\pi(\phi, \boldsymbol{x})$ such that $\pi(\phi \mid \boldsymbol{x}) = f(\phi \mid \boldsymbol{x})$ and $\pi(\boldsymbol{x} \mid \phi) = g(\boldsymbol{x} \mid \phi)$. Else, we say that f and g are *incompatible*.

More general definitions of compatibility can be found in references (Arnold et al., 2001; Liu et al., 2012). Sufficient conditions for compatibility are provided by Arnold et al. (2001, Theorem 1), restated here for convenience (and adapted to our notation).[9]

Theorem D.2. *$f(\phi \mid \boldsymbol{x})$ and $g(\boldsymbol{x} \mid \phi)$ are compatible if*

1. $\{(\phi, \boldsymbol{x}) : f(\phi \mid \boldsymbol{x}) > 0\} = \{(\phi, \boldsymbol{x}) : g(\boldsymbol{x} \mid \phi) > 0\} := S$,
2. *There exist functions $u(\boldsymbol{x})$ and $v(\phi)$ such that for every $(\phi, \boldsymbol{x}) \in S$,*

$$\frac{g(\boldsymbol{x} \mid \phi)}{f(\phi \mid \boldsymbol{x})} = u(\boldsymbol{x})v(\phi), \tag{24}$$

with $u(\boldsymbol{x})$ integrable, that is

$$\int_{\mathcal{X}} u(\boldsymbol{x})d\boldsymbol{x} < \infty. \tag{25}$$

We find it enlightening to rewrite the last condition as an integral over the left-hand-side of Eq. (25). Applying this condition to our problem we have

$$\int_{\mathcal{X}} \frac{q(\boldsymbol{x} \mid \phi, \boldsymbol{y})}{p(\phi \mid \boldsymbol{x}, \boldsymbol{y})}\mathrm{d}\boldsymbol{x} < \infty,$$

which is Eq. (7).

D.4. Error in the Stationary Distribution

Moving forward, we assume that $p(\phi \mid \boldsymbol{x})$ and $q(\boldsymbol{x} \mid \phi)$ are compatible, and so our Gibbs sampler admits a stationary distribution $\pi(\boldsymbol{x}, \phi)$. We now prove Theorem 2.3, which tells us the KL-divergence from the true (posterior) distribution $p(\phi)$ to the marginal stationary distribution $\pi(\phi)$ of our Gibbs sampler.

Proof. Standard application of the probability chain rule gives

$$p(\boldsymbol{x}) = p(\phi)\frac{p(\boldsymbol{x} \mid \phi)}{p(\phi \mid \boldsymbol{x})}. \tag{26}$$

Clearly, $p(\phi)$ must act as a normalizing constant and so

$$p(\phi) = \frac{1}{\int_{\mathcal{X}} \frac{p(\boldsymbol{x}|\phi)}{p(\phi|\boldsymbol{x})}\mathrm{d}\boldsymbol{x}}. \tag{27}$$

[9] As argued by Liu et al. (2021, Theorem 2.3), the conditions by Arnold et al. (2001) are not necessary and item 1 in Theorem D.2 can be replaced with a weaker condition.

Recalling that $\pi(\phi \mid \boldsymbol{x}) = p(\phi \mid \boldsymbol{x})$, a similar argument yields

$$\pi(\phi) = \frac{1}{\int_{\mathcal{X}} \frac{q(\boldsymbol{x}|\phi)}{p(\phi|\boldsymbol{x})} \mathrm{d}\boldsymbol{x}}. \tag{28}$$

Then

$$\begin{aligned}
& \tfrac{1}{p(\phi)} - \tfrac{1}{\pi(\phi)} &&= \int_{\mathcal{X}} \frac{p(\boldsymbol{x} \mid \phi) - q(\boldsymbol{x} \mid \phi)}{p(\phi \mid \boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\
\iff \quad & \tfrac{\pi(\phi) - p(\phi)}{p(\phi)\pi(\phi)} &&= \int_{\mathcal{X}} \frac{p(\boldsymbol{x} \mid \phi) - q(\boldsymbol{x} \mid \phi)}{p(\phi \mid \boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\
\iff \quad & \tfrac{\pi(\phi) - p(\phi)}{\pi(\phi)} &&= \int_{\mathcal{X}} (p(\boldsymbol{x} \mid \phi) - q(\boldsymbol{x} \mid \phi)) \frac{p(\boldsymbol{x})}{p(\boldsymbol{x} \mid \phi)} \mathrm{d}\boldsymbol{x} \\
\iff \quad & 1 - \tfrac{p(\phi)}{\pi(\phi)} &&= 1 - \int_{\mathcal{X}} \frac{q(\boldsymbol{x} \mid \phi)}{p(\boldsymbol{x} \mid \phi)} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
\iff \quad & \tfrac{p(\phi)}{\pi(\phi)} &&= \int_{\mathcal{X}} \frac{q(\boldsymbol{x} \mid \phi)}{p(\boldsymbol{x} \mid \phi)} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{29}$$

Finally recall that

$$\mathrm{KL}(p(\phi) || \pi(\phi)) = \int_{\Phi} \log \frac{p(\phi)}{\pi(\phi)} p(\phi) \mathrm{d}\phi, \tag{30}$$

and so taking on both sides of Eq. (29) the logarithm and the expectation value with respect to $p(\phi)$, we obtain the wanted result. □

Remark. Eq. (29), which applies to the density ratio at any point, provides a more general result than the result on averaged density ratios stated in Theorem 2.3.

E. Limitations of Guidance-based Posterior Sampling Methods

By examining the applicability of existing methods DPS (Chung et al., 2023) and Π-GDM (Song et al., 2023a) in the cosmology application presented in Sect. 3.2, we identified limitations in these guidance-based methods, which further legitimize the interest in our approach. We report our investigation in this appendix.

First, it is important to note that, since the exact conditional guidance term is typically intractable, guidance-based posterior sampling methods require methodological approximations. These approximations may impact the estimation of the posterior distribution in unpredictable ways. For scientific applications like our problem of cosmological inference from CMB data, this can be a significant downside. Furthermore, we encountered significant numerical difficulties with DPS and Π-GDM, which we consider strong evidence of the limitations of these methods for our cosmology application.

Let us first rewrite Eq. (1) as a standard linear inverse problem equation (i.e., involving white Gaussian noise). Starting from Eq. (1):

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\varepsilon}, \quad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}}), \tag{31}$$

and assuming $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ is invertible, we define $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1/2} \boldsymbol{\varepsilon}$, so that $\tilde{\boldsymbol{\varepsilon}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Then, introducing $\tilde{\boldsymbol{y}}_{\boldsymbol{\phi}} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1/2} \boldsymbol{y}$ and $\boldsymbol{A}_{\boldsymbol{\phi}} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1/2}$, we get:

$$\tilde{\boldsymbol{y}}_{\boldsymbol{\phi}} = \boldsymbol{A}_{\boldsymbol{\phi}} \boldsymbol{x} + \tilde{\boldsymbol{\varepsilon}}, \tag{32}$$

which resembles a standard linear inverse problem where the linear operator is $\boldsymbol{A}_{\boldsymbol{\phi}} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1/2}$. Also, note that the observation $\tilde{\boldsymbol{y}}_{\boldsymbol{\phi}}$ depends on the parameters $\boldsymbol{\phi}$, which adds a layer of complexity when the goal is to infer $\boldsymbol{\phi}$.

We implemented off-the-shelf DPS (Chung et al., 2023) and Π-GDM (Song et al., 2023a) algorithms in the context of our cosmology application[10]. The standard regularization (guidance strength) schemes initially yielded numerical float overflows. To avoid these numerical limitations, we had to decrease the guidance strength to a point where it was clear that

[10]Note that, in our case, the Π-GDM guidance term is straightforward to compute since the operator $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1/2}$ is diagonal in Fourier space.

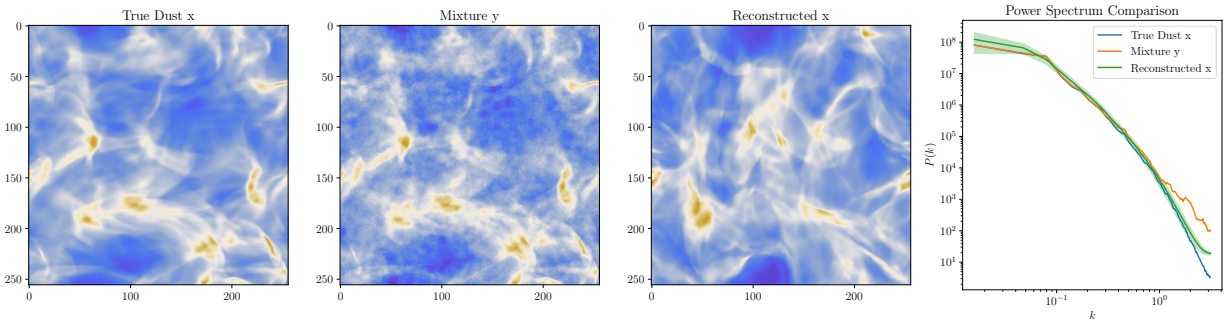

Figure E.1. Best results when optimizing the guidance strength for the standard off-the shelf Π-GDM algorithm.

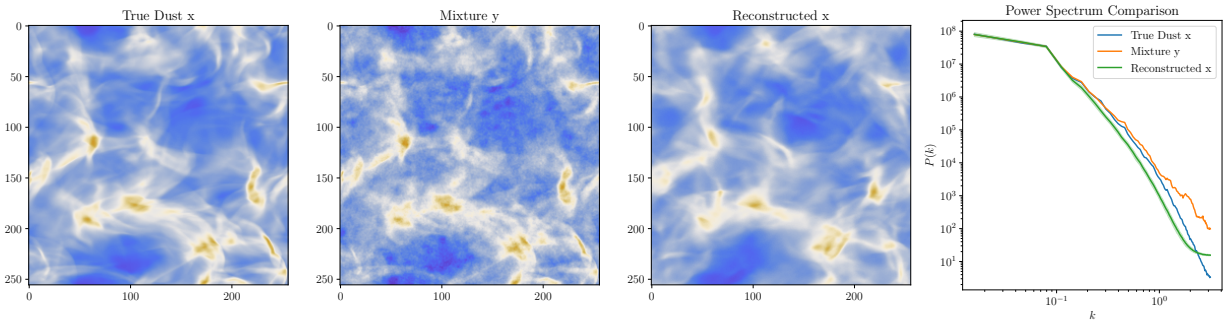

Figure E.2. Best results when optimizing a regularization in a modified Π-GDM.

we were not sampling the posterior distribution (large scales structure were off, as shown in Fig. E.1). Our most satisfying approach turned out to be a variation of Π-GDM involving a time-dependent regularization term (Song et al., 2023a, App. A.3) way outside the normal range for any inverse problems. Fig. E.2 depicts our results with associated summary statistics, proving that the posterior distribution is very poorly estimated. Furthermore, when trying to use this last method in our Gibbs pipeline, we could not reach convergence to any posterior distribution over ϕ.

We interpret these numerical challenges as the result of the significant dynamic range of eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$. For our cosmology application, the condition number of $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$ is approximately 6×10^5, and the largest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1/2}$ is 1.6×10^2. In comparison, standard inverse problems such as inpainting, deblurring, or super-resolution problems, typically involve operators with eigenvalues bounded between 0 and 1. Since guidance-based methods rely on estimates of $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t | \boldsymbol{y}, \boldsymbol{\phi})$, these large eigenvalues lead to numerical instabilities along their corresponding eigenvectors (i.e., for our problem, at high frequencies). If we attempt to mitigate this instability using regularization, the guidance along eigenvectors linked to the smallest eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1/2}$ becomes insufficient, resulting in impaired low frequencies/large-scale structures.

F. Additional Results

Dataset	Noise Level σ	$\varphi = -1 \rightarrow$ Pink noise				$\varphi = 0 \rightarrow$ White noise				$\varphi = 1 \rightarrow$ Blue noise			
		BM3D	DnCNN	**GDiff** $\hat{\boldsymbol{x}}$	**GDiff** $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$	BM3D	DnCNN	**GDiff** $\hat{\boldsymbol{x}}$	**GDiff** $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$	BM3D	DnCNN	**GDiff** $\hat{\boldsymbol{x}}$	**GDiff** $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$
ImageNet	0.06	0.90$_{\pm 0.01}$	(0.88)$_{\pm 0.00}$	0.86$_{\pm 0.01}$	**0.92**$_{\pm 0.00}$	0.92$_{\pm 0.00}$	0.92$_{\pm 0.00}$	0.88$_{\pm 0.01}$	**0.93**$_{\pm 0.00}$	0.94$_{\pm 0.00}$	(0.92)$_{\pm 0.00}$	0.90$_{\pm 0.00}$	**0.95**$_{\pm 0.00}$
	0.1	0.81$_{\pm 0.01}$	(0.76)$_{\pm 0.01}$	0.77$_{\pm 0.01}$	**0.86**$_{\pm 0.01}$	0.90$_{\pm 0.01}$	0.90$_{\pm 0.01}$	0.84$_{\pm 0.01}$	**0.91**$_{\pm 0.00}$	0.90$_{\pm 0.01}$	(0.88)$_{\pm 0.01}$	0.85$_{\pm 0.01}$	**0.92**$_{\pm 0.00}$
	0.2	0.63$_{\pm 0.01}$	(0.53)$_{\pm 0.01}$	0.62$_{\pm 0.02}$	**0.74**$_{\pm 0.02}$	0.79$_{\pm 0.01}$	0.80$_{\pm 0.01}$	0.74$_{\pm 0.01}$	**0.83**$_{\pm 0.01}$	0.83$_{\pm 0.01}$	(0.79)$_{\pm 0.01}$	0.78$_{\pm 0.01}$	**0.87**$_{\pm 0.01}$
CBSD68	0.06	0.90$_{\pm 0.01}$	(0.88)$_{\pm 0.00}$	0.84$_{\pm 0.01}$	**0.92**$_{\pm 0.00}$	0.93$_{\pm 0.00}$	**0.94**$_{\pm 0.00}$	0.88$_{\pm 0.00}$	**0.94**$_{\pm 0.00}$	0.94$_{\pm 0.00}$	(0.94)$_{\pm 0.00}$	0.90$_{\pm 0.00}$	**0.95**$_{\pm 0.00}$
	0.1	0.82$_{\pm 0.01}$	(0.78)$_{\pm 0.01}$	0.75$_{\pm 0.01}$	**0.85**$_{\pm 0.01}$	0.89$_{\pm 0.00}$	**0.90**$_{\pm 0.00}$	0.81$_{\pm 0.01}$	**0.90**$_{\pm 0.00}$	0.91$_{\pm 0.00}$	(0.90)$_{\pm 0.00}$	0.86$_{\pm 0.00}$	**0.93**$_{\pm 0.00}$
	0.2	0.65$_{\pm 0.01}$	(0.55)$_{\pm 0.01}$	0.61$_{\pm 0.01}$	**0.74**$_{\pm 0.01}$	0.79$_{\pm 0.01}$	0.80$_{\pm 0.01}$	0.69$_{\pm 0.01}$	**0.81**$_{\pm 0.01}$	0.83$_{\pm 0.01}$	(0.79)$_{\pm 0.01}$	0.76$_{\pm 0.01}$	**0.86**$_{\pm 0.01}$

Table 2. Image quality in terms of SSIM (↑) after denoising with GDiff (blind) and baselines BM3D (non-blind) and DnCNN (blind). We report mean PSNR and standard error computed on batches of 50 images. For GDiff, we provide performance for both posterior samples $\boldsymbol{x}$ and estimates of the posterior mean $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$. We point out that DnCNN was trained with white noises only, hence results obtained for $\varphi \neq 0$ could be sub-optimal.

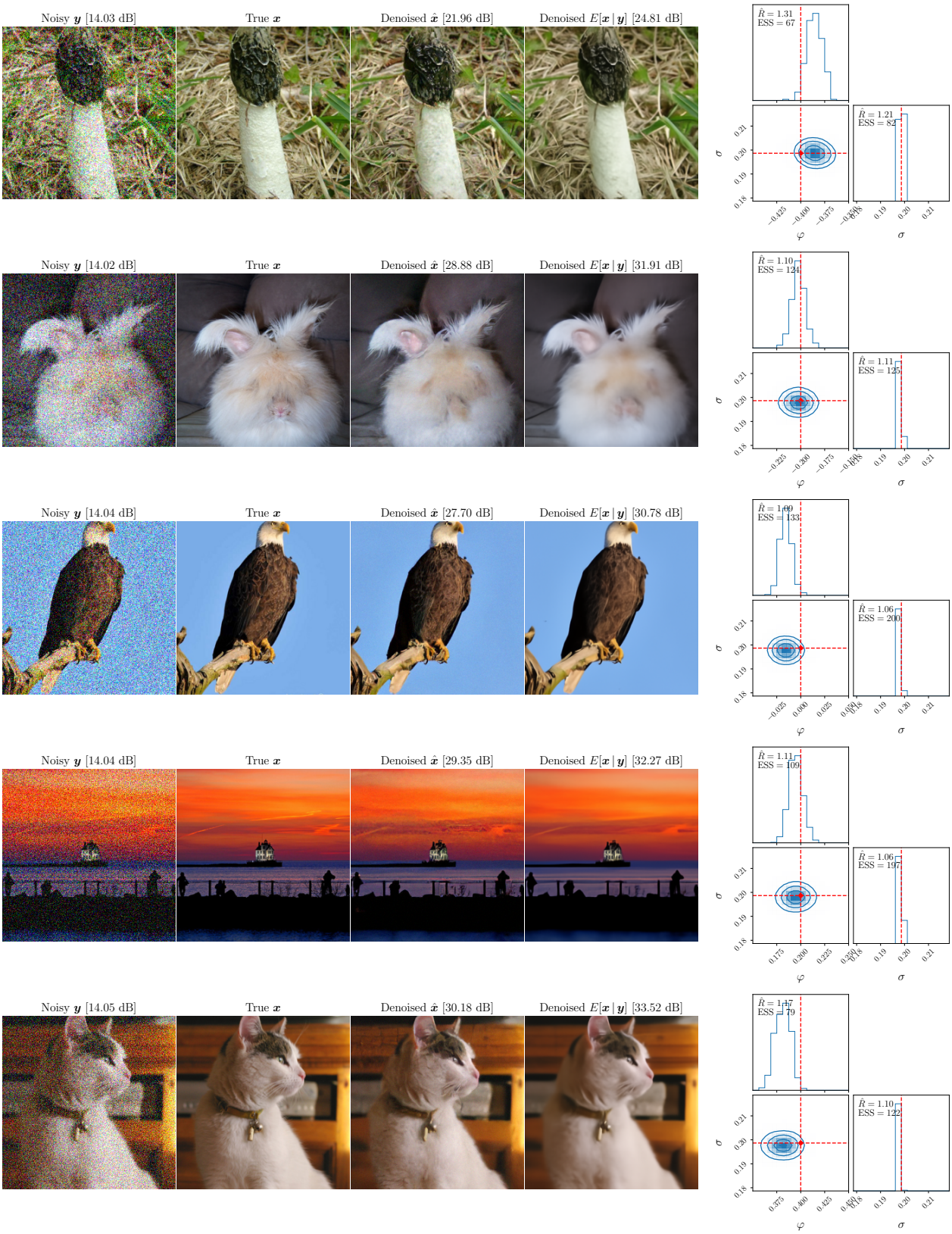

Figure F.1. Examples of blind denoising with GDiff on noisy ImageNet mocks with $\sigma = 0.2$ and $\varphi \in \{-0.4, -0.2, 0.0, 0.2, 0.4\}$ (from top to bottom). *Left:* Noisy examples $\boldsymbol{y}$ next to the noise-free images $\boldsymbol{x}$, denoised samples $\hat{\boldsymbol{x}}$ and estimates of the posterior mean $\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}]$. *Right:* Inferred posterior distributions over the noise parameters.

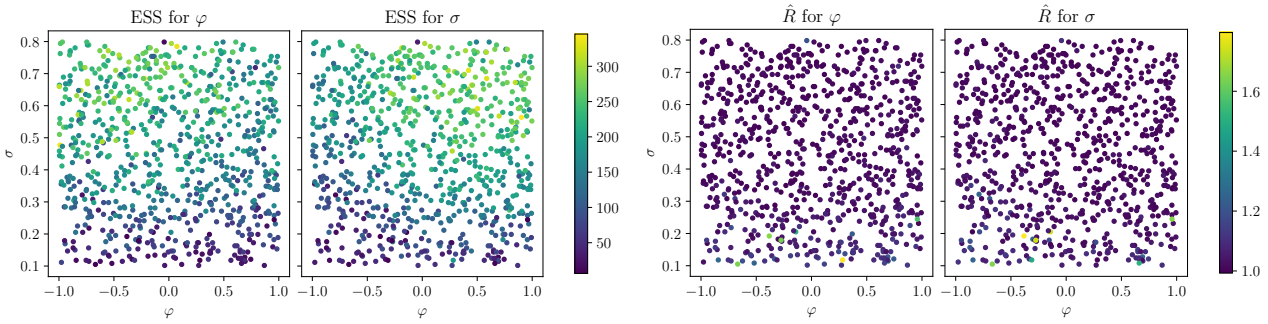

Figure F.2. ESS (left) and $\hat{R}$ (right) statistics for inferences on noisy natural images across the parameter space.

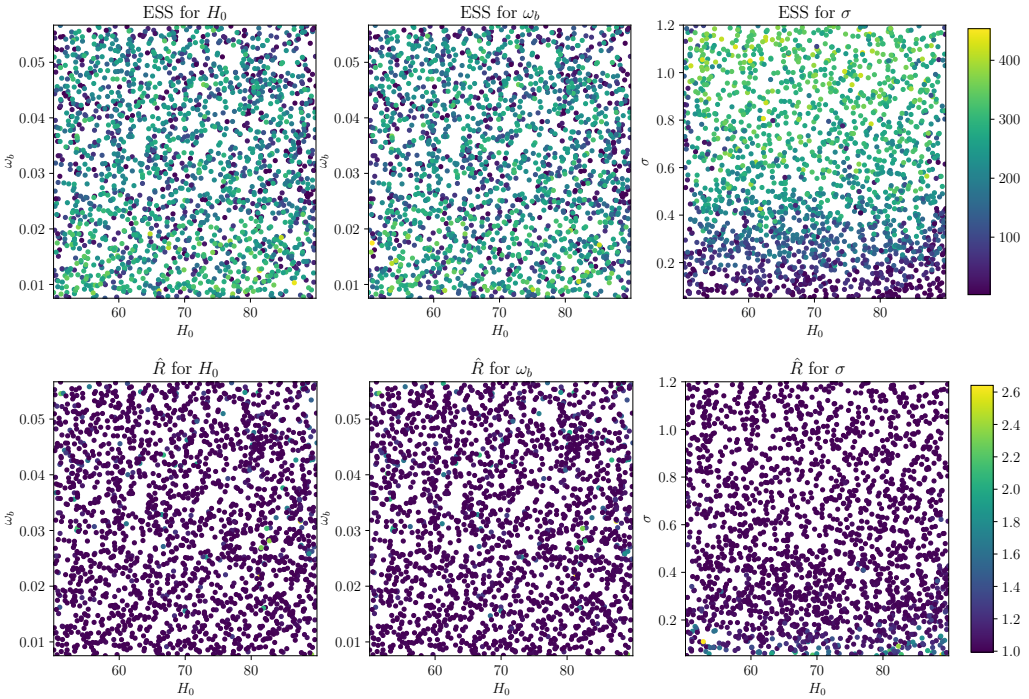

Figure F.3. ESS (top) and $\hat{R}$ (bottom) statistics for inferences on dust/CMB mixtures across the parameter space.