On the Asymptotic Mean Square Error Optimality of Diffusion Models

Benedikt Fesl
benedikt.fesl@tum.de

Benedikt Böck
benedikt.boeck@tum.de

Florian Strasser
f.strasser@tum.de

Michael Baur
mi.baur@tum.de

Michael Joham
joham@tum.de

Wolfgang Utschick
utschick@tum.de

Chair of Signal Processing, Technical University of Munich

Abstract

Diffusion models (DMs) as generative priors have recently shown great potential for denoising tasks but lack theoretical understanding with respect to their mean square error (MSE) optimality. This paper proposes a novel denoising strategy inspired by the structure of the MSE-optimal conditional mean estimator (CME). The resulting DM-based denoiser can be conveniently employed using a pre-trained DM, being particularly fast by truncating reverse diffusion steps and not requiring stochastic re-sampling. We present a comprehensive (non-)asymptotic optimality analysis of the proposed diffusion-based denoiser, demonstrating polynomial-time convergence to the CME under mild conditions. Our analysis also derives a novel Lipschitz constant that depends solely on the DM's hyperparameters. Further, we offer a new perspective on DMs, showing that they inherently combine an asymptotically optimal denoiser with a powerful generator, modifiable by switching re-sampling in the reverse process on or off. The theoretical findings are thoroughly validated with experiments based on various benchmark datasets.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

DMs (Sohl-Dickstein et al., 2015; Ho et al., 2020) and score-based models (Song and Ermon, 2019; Song et al., 2021b) are the backbone of foundation models such as Stable Diffusion (Rombach et al., 2022). Their success over different generative models, e.g., generative adversarial networks (GANs) (Goodfellow et al., 2014) or variational autoencoders (VAEs) (Kingma and Welling, 2014), is built upon their improved generative ability (Dhariwal and Nichol, 2021) and their theoretical understanding, particularly in terms of the convergence to the prior distribution of deterministic-time (Li et al., 2024; Liang et al., 2025) and continuous-time DMs (Bortoli et al., 2021; Chen et al., 2023c; Lee et al., 2023; Chen et al., 2023a; Benton et al., 2024; Pedrotti et al., 2024; Conforti et al., 2025; Chen et al., 2023b; Bortoli, 2022; Block et al., 2022; Chen et al., 2023d) with polynomial-time convergence guarantees. This advancement has led to the design of novel algorithms that utilize DMs, e.g., for inverse problems (Chung et al., 2022; Meng and Kabashima, 2024) and image denoising (Carlini et al., 2023; Xie et al., 2023; Fabian et al., 2024), super-resolution (Saharia et al., 2023), and restoration (Zhu et al., 2023). Moreover, DMs as generative priors have been successfully applied in various domains (Yang et al., 2023; Kong et al., 2021; Arvinte and Tamir, 2023; Huy and Minh Quan, 2023).

When considering denoising, the CME is the Bayesian-optimal solution in terms of the MSE (Lehmann and Casella, 2006), which is closely related to the peak signal-to-noise ratio (PSNR). It is of great importance in estimation and information theory (Kay, 1993; Asoodeh et al., 2016), statistical signal processing (Scharf and Demeure, 1991), and a pivotal benchmark for the analysis of denoising problems (Elad et al., 2023). Through Tweedie's formula (Efron, 2011), the CME

can be directly related to the score, which was, e.g., exploited in Noise2Score (Kim and Ye, 2021). Moreover, the minimum mean square error (MMSE) denoiser is frequently utilized in plug-and-play approaches (Venkatakrishnan et al., 2013; Romano et al., 2017) for which DMs are particularly viable (Zhu et al., 2023). In image processing or related fields, the MMSE is an imperative metric to assess the perception-distortion trade-off (Elad et al., 2023; Blau and Michaeli, 2018; Om and Biswas, 2014), which states that optimizing for perceptual quality often conflicts with minimizing distortion measures, e.g., the MSE. In addition, in applications where the perceptual quality is not directly measurable, the CME is a desirable performance bound, e.g., in speech enhancement (Loizou, 2013), wireless communications (Koller et al., 2022; Fesl et al., 2023), biomedical applications (Gargiulo and McEwan, 2011), or wavelet domain processing (Kazerouni et al., 2013).

Despite the prevailing usage of DMs in estimation tasks (Meng and Kabashima, 2024; Saharia et al., 2023; Whang et al., 2022; Huang et al., 2024; Tai et al., 2023), the solid theoretical guarantees for the convergence to the prior distribution (Li et al., 2024; Liang et al., 2025; Bortoli et al., 2021; Chen et al., 2023c; Lee et al., 2023; Chen et al., 2023a; Benton et al., 2024; Pedrotti et al., 2024; Conforti et al., 2025; Chen et al., 2023b; Bortoli, 2022; Block et al., 2022; Chen et al., 2023d), and the importance of the CME in various fields, there is a notable gap in the theoretical understanding of the connection of DM-based denoising and the CME. In view of the pivotal role of DMs for foundational models and their excellent denoising capabilities, a natural question to ask is: *How can a pre-trained DM be utilized in a fast and efficient manner for MSE-optimal estimation with provable convergence guarantees?* In short, we show the following main result.

Theorem 1.1 (**Main result (informal)**). *Let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n} \in \mathbb{R}^N$ be a noisy observation with additive white Gaussian noise (AWGN) $\boldsymbol{n}$ and the stepwise denoising error of the DM's reverse process be bounded by Δ. Then, the distance of the proposed denoiser $f_{\boldsymbol{\theta}}(\boldsymbol{y})$ that utilizes a pre-trained DM with T timesteps to the MSE-optimal CME $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]$ is bounded as*

$$\|\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] - f_{\boldsymbol{\theta}}(\boldsymbol{y})\| \leq \mathcal{O}(T^{-\gamma}\log\hat{t}) + \mathcal{O}(\hat{t}\log\hat{t})\Delta \quad (1)$$

with $\gamma > 0$ and $\hat{t} < T$ being the number of inference steps depending on the observation's signal-to-noise ratio (SNR) and the hyperparameters of the DM.

The error bound in (1) depends solely on the number of discretization steps T and the bounded stepwise denoising error Δ of the DM. Moreover, the error bound approaches zero in polynomial time as the number of diffusion steps T increases, making the proposed DM-based denoiser asymptotically optimal.

The paper's contributions are summarized as follows.

1) We motivate a novel denoising strategy that utilizes a pre-trained DM but only forwards the stepwise conditional mean in the inference phase without drawing stochastic samples in the reverse process by showing a connection to the ground-truth CME. This results in a fast and efficient inference procedure that lays the foundation for the theoretical analysis.

2) We derive a new closed-form expression for the DM's stepwise Lipschitz constant, which depends solely on the DM's hyperparameters and holds under mild assumptions; in particular, it remains valid regardless of the stepwise estimation error of the DM. This novel Lipschitz constant is pivotal in deriving the (non-)asymptotic convergence results. Additionally, it serves as the foundation to demonstrate that the proposed DM-based denoiser inherently optimizes the bias-variance trade-off, a fundamental principle in estimation theory.

3) We connect the optimality analysis of the proposed DM-based denoiser to the convergence analysis of the prior distribution. In our main result, we rigorously prove that the proposed DM-based denoising strategy is close to the ground-truth CME with polynomial-time convergence guarantee based on the derived Lipschitz constant under mild assumptions, i.e., we only require the common assumption of a bounded stepwise error of the DM, without assumptions about the convergence to the prior distribution or otherwise restrictive assumptions about the data distribution.

4) We reveal a novel perspective that DMs are comprised of a powerful generative model and an asymptotically MSE-optimal denoiser at the same time by switching the stochastic re-sampling in the reverse process on and off. In addition, we thoroughly validate the theoretical findings by experiments based on various benchmark datasets and perform ablation studies, demonstrating that the proposed DM-based denoising strategy is robust and asymptotically optimal.

2 RELATED WORK

Deterministic Sampling with DMs. Designing a DM with a deterministic reverse process is of great interest in current research since it allows for accelerated sampling and simplifies the design of a model consistent with the data manifold. The most prominent work investigating deterministic sampling is from Song et al. (2021a), where a non-Markovian forward process is designed to lead to a Markovian reverse process. This allows for a deterministic generative process where no re-sampling is necessary in each step. Song et al. (2023) designed a consistency model with a one-step

network that maps a noise sample to its corresponding datapoint. In contrast, our work presents a deterministic adaptation of the DM architecture presented by Ho et al. (2020), focusing on denoising rather than generation and allowing the use of pre-trained DMs.

Song et al. (2021b) were the first to interpret deterministic DMs as solutions to the corresponding probability flow ordinary differential equation (ODE). Building on this perspective, subsequent research has introduced new findings, e.g., (Lu et al., 2022; Chen et al., 2023b; Kim et al., 2024; Zhou et al., 2024). The ODE formulation of DMs can also be viewed as a neural ODE (Chen et al., 2018), highlighting its similarities with continuous normalizing flows (Papamakarios et al., 2021). Lipman et al. (2023) proposed an architecture that integrates both methodologies for image generation. While our work focuses on the DM architecture proposed by Ho et al. (2020), our theoretical results may be extendable to these alternative approaches.

Diffusion-based Denoising Approaches. Denoising has a large history in machine learning (Elad et al., 2023), and utilizing powerful generative priors for denoising tasks is a well-known and promising strategy (Kadkhodaie and Simoncelli, 2021). Building on DMs, stochastic denoising, i.e., posterior sampling, was proposed (Saharia et al., 2023; Meng and Kabashima, 2024; Whang et al., 2022; Huang et al., 2024; Tai et al., 2023).

Alternatively, conditional DMs were designed where the DM network is trained conditioned on a noisy version of the true sample (Saharia et al., 2023; Whang et al., 2022). Carlini et al. (2023) and Xie et al. (2023) proposed a new denoising strategy by starting the reverse sampling at an intermediate timestep that corresponds to the noise level of the observation rather than from pure noise to reduce the sampling time. This procedure is possible if the DM is consistent with the noise model. Nichol and Dhariwal (2021) and Xiao et al. (2023) also studied sub-sampling of DMs in the context of generation rather than denoising to reduce sampling time. Our work utilizes a similar strategy, i.e., the reverse sampling is started from the noise level of the observation using a pre-trained DM since we assume that the observations are corrupted with AWGN. However, we afterward use a deterministic reverse process and study the convergence to the CME rather than employing stochastic denoising.

Another related work that was recently devised, sharing a similar idea of forwarding the conditional mean in each step, is from Delbracio and Milanfar (2023). A fundamental difference is that they do not train a generative model but rather a nested regression process, entailing an entirely different convergence behavior.

In contempt of these advancements, the connection between the DM's inference process and the MSE-optimal CME remains unexplored, and the role of the stochastic re-sampling in the reverse process is not fully understood. Beyond that, an effective inference process to achieve the MSE-optimal solution utilizing a pre-trained DM is absent, being a notable gap in the study of DMs that is addressed in this work.

Convergence to the Prior Distribution. Many recent works have been devoted to the analysis of the DM's convergence to the prior distribution under various conditions on the data distribution and technical assumptions of the DM, resulting in an extensive collection of different convergence results (Li et al., 2024; Liang et al., 2025; Bortoli et al., 2021; Chen et al., 2023c; Lee et al., 2023; Chen et al., 2023a; Benton et al., 2024; Pedrotti et al., 2024; Conforti et al., 2025; Chen et al., 2023b; Bortoli, 2022; Block et al., 2022). The works from Li et al. (2024) and Liang et al. (2025) are most related to our setup since they also consider discrete-time DMs. Shah et al. (2023) showed that the DM can provably recover the ground-truth parameters of a Gaussian mixture model (GMM) distribution. This result is of particular interest in our context since the MSE-optimal CME can be computed in closed-form for a GMM distribution (Koller et al., 2022; Yang et al., 2015). Similar to our main result, the DM's hyperparameters and the stepwise error are assumed to entail an asymptotic behavior over the number of diffusion steps by Liang et al. (2025). Several works discussed theoretical justifications of this natural assumption of a bounded stepwise DM error (Chen et al., 2023c,a; Block et al., 2022). However, although these strong results about the convergence to the prior distribution exist, no result discusses the theoretical capability of DMs to approximate the MSE-optimal CME.

3 PRELIMINARIES

Problem Formulation. We consider a denoising task where the true sample is corrupted by AWGN, yielding an observation

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n} \in \mathbb{R}^N \tag{2}$$

where $\boldsymbol{x} \sim p(\boldsymbol{x})$ follows an unknown distribution and $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \eta^2 \mathbf{I})$ with known variance η^2. Minimizing the MSE (or maximizing the PSNR) yields the optimization problem

$$g^*(\boldsymbol{y}) = \underset{g:\mathbb{R}^N \to \mathbb{R}^N}{\arg\min}\ \mathbb{E}[\|\boldsymbol{x} - g(\boldsymbol{y})\|_2^2] \tag{3}$$

where g^* is known as the CME, computed as

$$g^*(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) \mathrm{d}\boldsymbol{x} = \int \boldsymbol{x} \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}. \tag{4}$$

However, for an unknown prior $p(\boldsymbol{x})$, the CME is intractable to compute due to the high-dimensional integral, and reasonable approximations must be found. A naive approach is to parameterize a one-step regression neural network (NN) $g_{\boldsymbol{\theta}}$ trained on the MSE loss (3) to provide an estimate $g_{\boldsymbol{\theta}}(\boldsymbol{y}) \approx g^*(\boldsymbol{y})$ at the output of the NN (Dong et al., 2016; Ongie et al., 2020). However, in contrast to a regression network, the proposed DM-based denoiser has several significant advantages.

First, one-step regression networks are "task-specific" and have to be trained from scratch on paired examples (commonly noisy and clean samples) (Elad et al., 2023). In contrast, a pre-trained DM can be considered a "general purpose" denoiser leveraging powerful foundational DMs such as Stable Diffusion (Rombach et al., 2022). Second, one-step networks are generally prone to generalization issues when being trained on a large range of noise levels, necessitating either re-training or a meticulous design of the architecture and loss function (Zhao et al., 2017; Mohan et al., 2020). In contrast, DMs are expected to generalize better as they are inherently trained on the full SNR range. Third, one-step networks are not very well understood theoretically, as the convergence to the CME is solely attributed to the universal approximation property (Diaz et al., 2021). With this paper, we aim to address this limitation by rigorously proving the (non-)asymptotic convergence of the proposed DM-based denoising strategy to the CME under mild assumptions.

Diffusion Models. We briefly review the DM formulations from Ho et al. (2020). Given a data distribution $\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0)$, the *forward process* which produces latents $\boldsymbol{x}_1$ through $\boldsymbol{x}_T$ by adding Gaussian noise at time t with the hyperparameters $\alpha_t, \beta_t \in (0,1)$ with $\beta_t = 1 - \alpha_t$ for all $t = 1, \dots, T$ is a Markov chain that is defined via the transition

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right). \tag{5}$$

Iteratively applying the reparameterization trick yields

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_{t-1} = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \tag{6}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The joint distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ is called the *reverse process* and is defined as a Markov chain via the parameterized Gaussian transitions

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \sigma_t^2 \mathbf{I}). \tag{7}$$

As Ho et al. (2020), we set the variances of the reverse process to untrained time-dependent constants. The choice in (7) is motivated by the fact that the forward and reverse process of a Markov chain have the same functional form when α_t is close to one for all $t = 1, \dots, T$ (Feller, 1949; Sohl-Dickstein et al., 2015).

However, the transitions in (7) are generally intractable and are thus learned via the forward posteriors, which are tractable when conditioned on $\boldsymbol{x}_0$, i.e.,

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0), \sigma_t^2 \mathbf{I}), \tag{8}$$

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t, \tag{9}$$

$$\sigma_t^2 = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}. \tag{10}$$

Training details of the DM are given in Appendix A.

In the remainder of this work, we consider $\mathbb{E}[\boldsymbol{x}_0] = \mathbf{0}$ and $\mathbb{E}[\|\boldsymbol{x}_0\|_2^2] = N$, which is ensured by a pre-processing step, such that the observation's SNR can be defined as $\mathrm{SNR}(\boldsymbol{y}) = 1/\eta^2$. Then, the DM timesteps can be equivalently interpreted as different SNR steps (Kingma et al., 2021; Luo, 2022) by defining the DM's SNR of step t, which is monotonically decreasing for increasing t, cf. (6), as

$$\mathrm{SNR}_{\mathrm{DM}}(t) = \frac{\mathbb{E}[\|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\|_2^2]}{\mathbb{E}[\|\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0\|_2^2]} = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}. \tag{11}$$

4 MSE-OPTIMAL DENOISING

4.1 Discussion of the DM-based Denoiser

Proposed Denoiser. We consider a DM that is consistent with the noise model, i.e., the latent variables in the forward process and the observation (2) are corrupted by AWGN. In that case, it is possible to initialize the reverse process at timestep $\hat{t} < T$ that corresponds to the SNR of the observation rather than from pure noise (Carlini et al., 2023; Xie et al., 2023). This results in a considerable reduction of reverse steps for denoising, effectively reducing the computational complexity and allowing the utilization of a pre-trained DM. Thus, after normalizing the observation from (2), i.e., $\tilde{\boldsymbol{y}} = (1+\eta^2)^{-\frac{1}{2}}\boldsymbol{y}$ such that $\mathbb{E}[\|\tilde{\boldsymbol{y}}\|_2^2] = N$, which is necessary since the DM is variance-preserving, we find the timestep of the DM that best matches the SNR of the observation, which is assumed to be known (we refer to Section 5 and Appendix J for an ablation study of a mismatch in the SNR information), via

$$\begin{aligned} \hat{t} &= \arg\min_t \left|\mathrm{SNR}(\boldsymbol{y}) - \mathrm{SNR}_{\mathrm{DM}}(t)\right| \\ &= \arg\min_t \left|\frac{1}{\eta^2} - \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right|. \end{aligned} \tag{12}$$

We note that $\hat{t} < T$ solely depends on the hyperparameter choice of the DM and can be either computed analytically or by a simple search over the DM's timesteps. Consequently, we initialize the reverse process with $\boldsymbol{x}_{\hat{t}} = \tilde{\boldsymbol{y}}$ and denote the ground-truth CME

Algorithm 1 Deterministic DM-based denoising.

Require: Observation $\boldsymbol{y}$, pre-trained DM $\{f_{\boldsymbol{\theta},t}\}_{t=1}^{T}$, noise variance η^2
1: Normalize observation's variance $\tilde{\boldsymbol{y}} = (1+\eta^2)^{-\frac{1}{2}}\boldsymbol{y}$
2: $\hat{t} = \arg\min_t |\mathrm{SNR}(\boldsymbol{y}) - \mathrm{SNR}_{\mathrm{DM}}(t)|$
3: Initialize $\hat{\boldsymbol{x}}_{\hat{t}} = \tilde{\boldsymbol{y}}$
4: **for** $t = \hat{t}$ **down to** 1 **do**
5: $\quad \hat{\boldsymbol{x}}_{t-1} = f_{\boldsymbol{\theta},t}(\hat{\boldsymbol{x}}_t)$
6: **end for**

Algorithm 2 Reverse process from Ho et al. (2020).

Require: Pre-trained DM $\{f_{\boldsymbol{\theta},t}\}_{t=1}^{T}$
1: Initialize $\hat{\boldsymbol{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T$ **down to** 1 **do**
3: $\quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
4: $\quad \hat{\boldsymbol{x}}_{t-1} = f_{\boldsymbol{\theta},t}(\hat{\boldsymbol{x}}_t) + \sigma_t \boldsymbol{z}$
5: **end for**

as $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}]$. The CME is a deterministic point estimate, motivating to avoid re-sampling in the reverse process during inference if convergence to the CME is desired, as opposed to stochastic denoising (Whang et al., 2022; Xie et al., 2023; Saharia et al., 2023; Meng and Kabashima, 2024). Thus, instead of stochastic sampling from the parameterized reverse process distribution as $\boldsymbol{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \sigma_t^2 \mathbf{I})$, cf. (7), we propose to solely forward its conditional mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ in each step. By denoting $f_{\boldsymbol{\theta},t}(\boldsymbol{x}_t) := \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, cf. (7), the proposed DM-based denoiser is given as

$$f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}}) = f_{\boldsymbol{\theta},1}(f_{\boldsymbol{\theta},2}(\cdots f_{\boldsymbol{\theta},\hat{t}}(\boldsymbol{x}_{\hat{t}})\cdots)). \tag{13}$$

The proposed denoising procedure is summarized in Algorithm 1. In comparison to the reverse sampling process of the DM by Ho et al. (2020), outlined in Algorithm 2 for convenience, the proposed DM-based denoiser avoids stochastic re-sampling and skips the reverse steps $t = T, T-1, \ldots, \hat{t}+1$ that correspond to a lower SNR value than that of the observation. Additionally, in Figure 1, the full Markov chain of the pre-trained DM is visualized, where the reverse steps shaded in gray are skipped in the proposed denoiser. The denoiser has a direct connection to the CME, laying the foundation for the convergence analysis as discussed in the following.

Connection to the CME. Let us exemplarily assume $\hat{t} = 2$ and utilize the law of total expectation to rewrite the ground-truth CME $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}} = \boldsymbol{x}_2]$ in terms of the latent variable $\boldsymbol{x}_1$ as

$$\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_2] = \mathbb{E}[\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_1, \boldsymbol{x}_2]|\boldsymbol{x}_2] \overset{(i)}{=} \mathbb{E}[\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_1]|\boldsymbol{x}_2] \tag{14}$$

$$\overset{(ii)}{=} \mathbb{E}[g_1(\boldsymbol{x}_1)|\boldsymbol{x}_2] \tag{15}$$

where in (i) the Markov property of the time reversal (Chung and Walsh, 2005, Ch. 10) is utilized, and in (ii) we defined the inner conditional expectation as a function $g_t(\boldsymbol{x}_t) := \mathbb{E}[\boldsymbol{x}_{t-1}|\boldsymbol{x}_t]$ that is ideally well-approximated by the DM, i.e., $f_{\boldsymbol{\theta},t} \approx g_t$ for all $t = 1, \ldots, T$. At this point, we aim to reformulate the ground-truth CME as a nested function $g_{1:2}(\boldsymbol{x}_t)$ to have the same functional form as the DM estimator (13). Since g_t is generally nonlinear (linearity only holds if the prior is Gaussian (Akyol et al., 2012))

$$\mathbb{E}[g_1(\boldsymbol{x}_1)|\boldsymbol{x}_2] \approx g_1(\mathbb{E}[\boldsymbol{x}_1|\boldsymbol{x}_2]) = g_{1:2}(\boldsymbol{x}_2) \tag{16}$$

remains an approximation, and the difference is called the Jensen gap (Chow and Teicher, 1997). The generalization to $\hat{t} > 2$ is straightforward and explicitly discussed in the later section. When the number of DM timesteps T approaches infinity, the SNR values of the subsequent timesteps get closer and, consequently, the approximation in (16) is better justified. An additional motivation is the linearity of $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0)$ that is approximated by $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ through the DM's loss function (35) since for a linear function the approximation in (16) is fulfilled with equality. However, when increasing the number of DM timesteps T, the number of error terms due to the Jensen gaps is also growing. In combination with a possible approximation error of the DM's NN, the overall convergence of the accumulated error terms necessitates a careful investigation.

4.2 A Novel Lipschitz Constant

Before stating our main results, we derive a novel expression for the Lipschitz constant of the DM network under mild assumptions, laying the foundation for the optimality analysis.

Lemma 4.1. *Let $f_{\boldsymbol{\theta},t}(\boldsymbol{x}_t) := \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ be continuous and differentiable almost everywhere for all $t = 2, \ldots, T$ with $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \sigma_t^2 \mathbf{I})$. Then,*

$$\|f_{\boldsymbol{\theta},t}(\boldsymbol{a}) - f_{\boldsymbol{\theta},t}(\boldsymbol{b})\| \leq L_t \|\boldsymbol{a} - \boldsymbol{b}\|, \tag{17}$$

$$L_t = \sqrt{\alpha_t}\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} < 1, \tag{18}$$

for all $t = 2, \ldots, T$, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$. Further, let $\tau_1, \tau_2 \in \mathbb{N}, 1 < \tau_1 \leq \tau_2 \leq T$. Then, for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$

$$\|f_{\boldsymbol{\theta},\tau_1:\tau_2}(\boldsymbol{a}) - f_{\boldsymbol{\theta},\tau_1:\tau_2}(\boldsymbol{b})\| \leq L_{\tau_1:\tau_2} \|\boldsymbol{a} - \boldsymbol{b}\|, \tag{19}$$

$$L_{\tau_1:\tau_2} = \frac{(1-\bar{\alpha}_{\tau_1-1})\sqrt{\bar{\alpha}_{\tau_2}}}{(1-\bar{\alpha}_{\tau_2})\sqrt{\bar{\alpha}_{\tau_1-1}}} < 1. \tag{20}$$

Proof sketch: The result is a consequence of the derivative identity for the CME from Dytso et al. (2020), revealing that the Lipschitz constant is independent of the realization $\boldsymbol{x}_t$ but solely depends on the DM's hyperparameters. For a rigorous proof, see Appendix B.

Discussion: Notably, $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0)$ has precisely the same Lipschitz constant L_t, which can be immediately seen

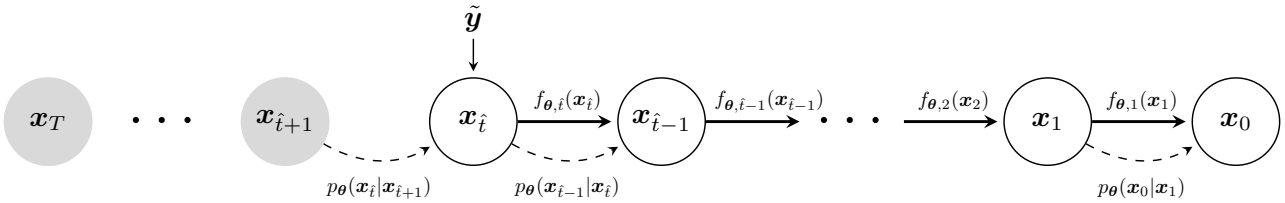

Figure 1: Markov chain of the DM's full reverse process with T steps and visualization of the proposed denoising procedure, where DM steps $t > \hat{t}$ (shaded in gray) are omitted for the estimation.

as it is linear in $\boldsymbol{x}_t$, cf. (9). The reason for this is the Gaussianity of the noise in the forward process and that $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ have the same covariance matrix that is a time-dependent constant (only depending on the DM's hyperparameters), resulting in the same derivative of the conditional means (Dytso et al., 2020). Interestingly, the result in Lemma 4.1 is independent of the mismatch between $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0)$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ and thus requires no assumptions about the error between the conditional first moments. Note that although different parameterizations of the DM exist, i.e., learning the noise rather than the conditional mean (Ho et al., 2020), we can always reparameterize such that all results are straightforwardly applicable to different DM parameterizations, see Appendix C.

Since the Lipschitz constant is related to the impact of the observation onto the estimated value, the SNR-dependency of the DM-based channel estimator's Lipschitz constant resembles a well-known property of the CME, i.e., optimizing the bias-variance trade-off, see Appendix D. This well-interpretable property of the proposed DM-based estimator motivates further investigation of its convergence to the CME. In Appendix J.4, we further analyze the Lipschitz constant and its behavior over the reverse steps empirically, validating the aforementioned theoretical findings.

The Lipschitz continuity of $f_{\boldsymbol{\theta},1}$ cannot be proven analytically since the ground-truth transition (8) is not available in this case, which is the reason for the *reconstruction term* in the evidence lower bound (ELBO) (34). In a practical deployment, the proposed adaptations of Yang et al. (2024) can be utilized to ensure a bounded and small Lipschitz constant for the last step.

4.3 Analysis under Prior Convergence

We aim to analyze the distance between the proposed DM-based estimator (13) and the ground-truth CME, i.e., $\|\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\|$. A natural question arising is whether the convergence results to the prior distribution in the literature, cf. Section 2, can be directly utilized to analyze the DM's convergence to the CME. By defining the CME parameterized through the DM's prior $p_{\boldsymbol{\theta}}(\boldsymbol{x}_0)$, i.e.,

$$\mathbb{E}^{\boldsymbol{\theta}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] := \int \boldsymbol{x}_0 \frac{p(\boldsymbol{x}_{\hat{t}}|\boldsymbol{x}_0)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0)}{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\hat{t}})}\,\mathrm{d}\boldsymbol{x}_0, \tag{21}$$

we can split the error term into two parts

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| \leq \underbrace{\|\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - \mathbb{E}^{\boldsymbol{\theta}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}]\|}_{\text{prior distribution}} \\ + \underbrace{\|\mathbb{E}^{\boldsymbol{\theta}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\|}_{\text{denoising procedure}}.\end{aligned} \tag{22}$$

The first term solely depends on the convergence of the prior distribution, whereas the second term only depends on the chosen denoising procedure. As shown in (Koller et al., 2022, Theorem 2), a sufficient condition for the convergence of the first term is uniform convergence of the prior distribution, being a strong assumption that is not guaranteed by the weaker statements about the Kullback-Leibler divergence, total variation or Wasserstein distance in existing works, see Section 2. Moreover, the second term is generally nonzero, containing the Jensen gaps, as discussed in the preceding section, that have to be bounded. In the following, we state our results when splitting the error term as in (22).

Theorem 4.2. *Let the DM's hyperparameters be given as (see Ho et al. (2020); Liang et al. (2025); Meng and Kabashima (2024); Nichol and Dhariwal (2021) for similar choices)*

$$\left\{\beta_t \;\middle|\; \gamma > 0, \beta_t = \mathcal{O}\left(T^{-\gamma}\right), \beta_1 \leq \beta_2 \leq \cdots \leq \beta_T\right\} \tag{23}$$

with $\beta_t = 1 - \alpha_t$ and let (7) hold for all $t = 1, \ldots, T$. Further, let $f_{\boldsymbol{\theta},1}$ be Lipschitz continuous with constant $L_1 < \infty$ and the score of the observation be bounded as

$$\|\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{y}) - \nabla \log p(\boldsymbol{y})\| \leq \Xi. \tag{24}$$

Then, the distance of the DM estimator $f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})$ to the ground-truth CME $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}]$ is bounded as

$$\begin{aligned}&\|\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| \\ &\leq 2NL_1\mathcal{O}(T^{-\gamma/2})(1 + \log\hat{t}) + \Xi\eta^2.\end{aligned} \tag{25}$$

Further, if $\lim_{T\to\infty}\Xi = 0$ *or* $\lim_{T\to\infty}\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_0) - p(\boldsymbol{x}_0)\|_\infty = 0$*, then the distance is vanishing in the asymptotic regime as*

$$\lim_{T\to\infty}\|\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| = 0. \tag{26}$$

Proof sketch: Utilizing a reformulation as in (15) allows to express the parameterized CME in terms of the DM estimator and a residual error term reflecting the Jensen gaps as in (16). Using the Lipschitz constant from (18) and the definition of the hyperparameters (23) together with some algebraic reformulations yields the first term in the bound (25). The second term in (25) is a direct consequence of Tweedie's formula (Efron, 2011). For a rigorous proof, see Appendix E.

Discussion: Theorem 4.2 connects the convergence to the prior distribution with the convergence to the CME. The first term in (25) stems from the denoising procedure and is vanishing in polynomial time, being the first convergence result to the CME for DMs. However, the assumptions of a bounded and vanishing score (Hyvärinen, 2005) or the uniform convergence of the prior (attributed to the universal approximation property (Nguyen et al., 2020)) are strong and neglect that the reverse process may have much fewer steps than the DM, i.e., $\hat{t} < T$. Thus, we study the convergence to the CME without assuming the convergence to the prior, which is considered the main result of this work.

4.4 Analysis under Stepwise Error Bound

Assumption 4.3. Let us define the DM's NN function of each step as the ground-truth conditional mean (9) conditioned on $\boldsymbol{x}_0$ plus an error term that is bounded for all $t = 1,\dots,\hat{t}$, $\boldsymbol{x}_t \in \mathbb{R}^N$ as

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0) = f_{\boldsymbol{\theta},t}(\boldsymbol{x}_t) + \boldsymbol{\delta}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tag{27}$$

$$\mathbb{E}_{q(\boldsymbol{x}_0|\boldsymbol{x}_t)}[\|\boldsymbol{\delta}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)\||\boldsymbol{x}_t] \leq \Delta. \tag{28}$$

The assumption of a bounded stepwise error is natural and common in many related works (Li et al., 2024; Liang et al., 2025; Bortoli et al., 2021; Chen et al., 2023c; Lee et al., 2023; Chen et al., 2023a,b; Block et al., 2022).

Theorem 4.4 (Main Result). *Let the hyperparameters of the DM be given as in* (23) *and let* (7) *hold. Further, let* $f_{\boldsymbol{\theta},1}$ *be Lipschitz continuous with constant* $L_1 < \infty$*. Under Assumption 4.3, the distance of the DM estimator* $f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})$ *to the ground-truth CME* $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}]$ *is bounded as*

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| &\leq 2NL_1(1+\log\hat{t})\mathcal{O}(T^{-\gamma/2}) \\ &\quad + \left(4L_1\hat{t}\log\hat{t} + (4L_1+2)\hat{t} + 2L_1\log\hat{t} + 2L_1 - 1\right)\Delta \qquad (29)\\ &= \mathcal{O}(T^{-\gamma/2}\log\hat{t}) + \mathcal{O}(\hat{t}\log\hat{t})\Delta. \qquad (30)\end{aligned}$$

Proof sketch: The proof has a similar structure as the one for Theorem 4.2 but with some additional error terms occurring due to the less restrictive assumption of an imperfect NN rather than assuming convergence to the prior. See Appendix F for a rigorous proof.

Assumption 4.5. The error bound (28) entails the asymptotic behavior (similar to Liang et al. (2025))

$$\Delta = \mathcal{O}(\hat{t}^{-\omega}),\ \omega > 1. \tag{31}$$

Corollary 4.6. *Under the same assumptions as in Theorem 4.4 and Assumption 4.5, the distance between the DM estimator and the ground-truth CME is asymptotically vanishing as*

$$\lim_{T\to\infty}\|\mathbb{E}_{0|\hat{t}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| = 0. \tag{32}$$

Proof: The result immediately follows from (30) together with Assumption 4.5.

Discussion: Our result of utilizing a pre-trained DM with a non-stochastic reverse process and an SNR-dependent number of inference steps having provable convergence to the MSE-optimal solution in polynomial time is the first result in this direction. The necessary assumptions are minimal and widely accepted in related literature. Notably, Theorem 4.4 improves over the result in Theorem 4.2 since no requirements about the convergence to the prior distribution are necessary. Moreover, in contrast to (25), the bound in (29) does not depend on the quality of the reverse steps $\hat{t}+1, \hat{t}+2,\dots,T$ and can thus be expected to be much tighter. Further, $\hat{t} = \mathcal{O}(T)$ trivially holds; however, $\hat{t}$ can also have a slower growth rate depending on the scheduling as exemplarily shown in Appendix G.

4.5 Denoising to Generation via Re-sampling

We highlight the novel observation that the DM is comprised of a denoiser that is asymptotically converging to the ground-truth CME in AWGN and a powerful generative model at the same time. The only difference between the two operation modes is switching the re-sampling in the reverse process on and off. However, since the variance of the re-sampling step for generating samples also converges to zero in the limit of $T \to \infty$, one may ask whether a stochastic denoising procedure via re-sampling in the reverse process collapses to a deterministic point estimate. As we show in Appendix H, this is not the case, and the two operation modes differ fundamentally. This novel perspective has an important connection to the work on foundational generative models such as Stable Diffusion (Rombach et al., 2022) by showing that an asymptotically optimal denoiser is inherently provided by a well-trained DM.

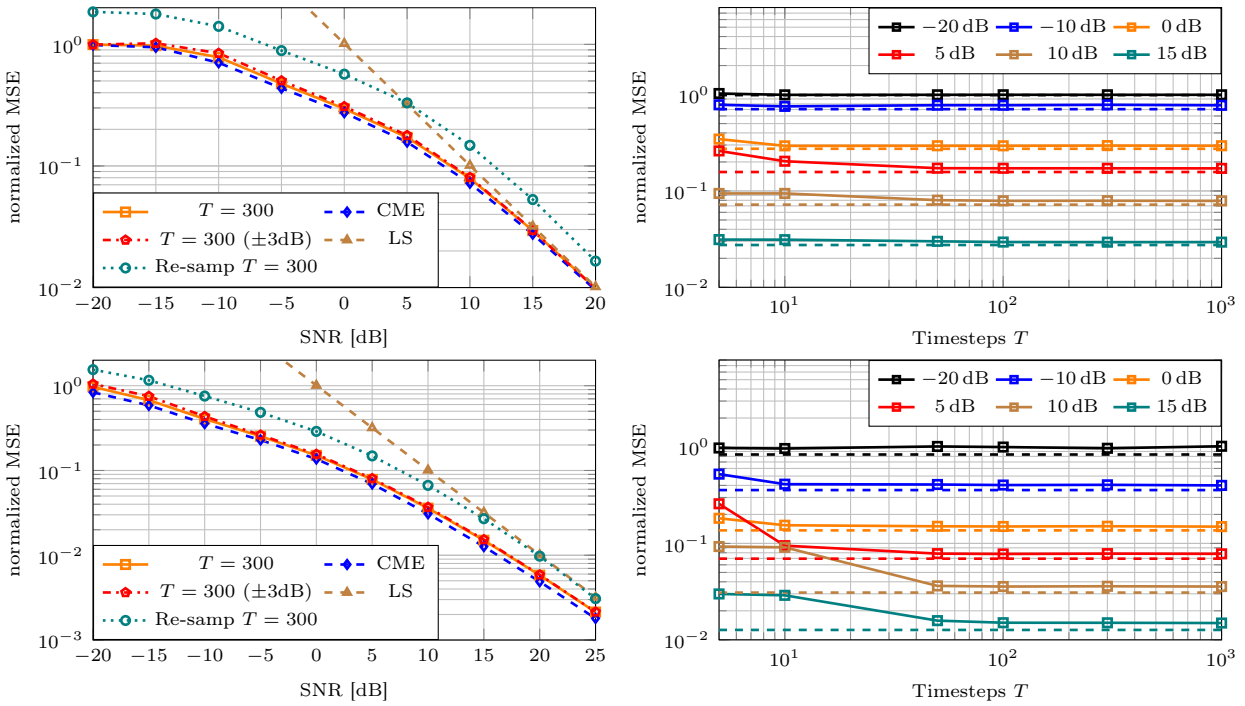

Figure 2: Evaluation of the **random** GMM with $N = 64$ dimensions (top row) and a pre-trained GMM based on **MNIST** data (bottom row) with $K = 128$ components as ground-truth distribution.

5 EXPERIMENTS

To validate our theoretical findings, we aim to choose a ground-truth distribution that is non-trivial but theoretically possible to learn via a DM and simultaneously has a closed-form CME expression. A well-known model that exhibits both properties is the GMM, cf. Shah et al. (2023), with a density $p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{C}_k)$ where $\{p(k), \boldsymbol{\mu}_k, \boldsymbol{C}_k\}_{k=1}^{K}$ are the mixing coefficients, means, and covariances. The corresponding MMSE estimator (Koller et al., 2022; Yang et al., 2015) is

$$g^*(\boldsymbol{y}) = \sum_{k=1}^{K} p(k|\boldsymbol{y}) \left(\boldsymbol{\mu}_k + \boldsymbol{C}_k \boldsymbol{C}_{\boldsymbol{y}|k}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_k)\right), \quad (33)$$

where $p(k|\boldsymbol{y}) \propto p(k)\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{C}_{\boldsymbol{y}|k})$ are the responsibilities, and $\boldsymbol{C}_{\boldsymbol{y}|k} = \boldsymbol{C}_k + \eta^2 \mathbf{I}$. As a baseline, we evaluate the least squares (LS) estimator $\hat{\boldsymbol{x}}_{\text{LS}} = \boldsymbol{y}$. We also study how a mismatch in the SNR information affects the DM-based denoiser where the ground-truth SNR is artificially corrupted with a uniformly distributed offset in the range $[-3, 3]$dB and perform a runtime analysis (see Appendix J). In Appendix J.5, we additionally present qualitative results of the proposed denoiser on GMM-generated and original image datasets, reinforcing the theoretical findings of this work and demonstrating its strong denoising performance.

Network Architecture. We employ a similar DM architecture as Ho et al. (2020) with only minor adaptations, i.e., we utilize a U-net (Ronneberger et al., 2015) backbone based on wide ResNet (Zagoruyko and Komodakis, 2017) and time-sharing of the NN parameters with a sinusoidal position embedding of the time information (Vaswani et al., 2017). In contrast to Ho et al. (2020), we leave out attention modules as this showed no improvement for the denoising task. We choose a linear schedule of β_t between constants β_1 and β_T. For each experiment, we use a training dataset consisting of 100,000 samples and evaluate the normalized MSE $\sum_{i=1}^{M_{\text{test}}} \|\boldsymbol{x}_{0,i} - \hat{\boldsymbol{x}}_{0,i}\|_2^2 / \sum_{i=1}^{M_{\text{test}}} \|\boldsymbol{x}_{0,i}\|_2^2$ with $M_{\text{test}} = 10{,}000$ test data. Details of the network architecture and the hyperparameter choice can be found in Appendix I. The simulation code is publicly available.[1]

Random GMM. We take a randomly initialized GMM with $K = 128$ components and $N = 64$ dimensions as ground-truth distribution, cf. Appendix J for details. In Figure 2 (top left), the DM denoiser with $T = 300$ timesteps is evaluated against the ground-truth CME that has complete knowledge of the GMM's parameters and computes the CME (33). The investigated denoising procedure via the pre-trained DM indeed achieves an estimation performance that is almost on par with that of the CME, validating the strong MSE performance that was theoretically analyzed, already for a reasonable number of timesteps.

[1] https://github.com/benediktfesl/Diffusion_MSE

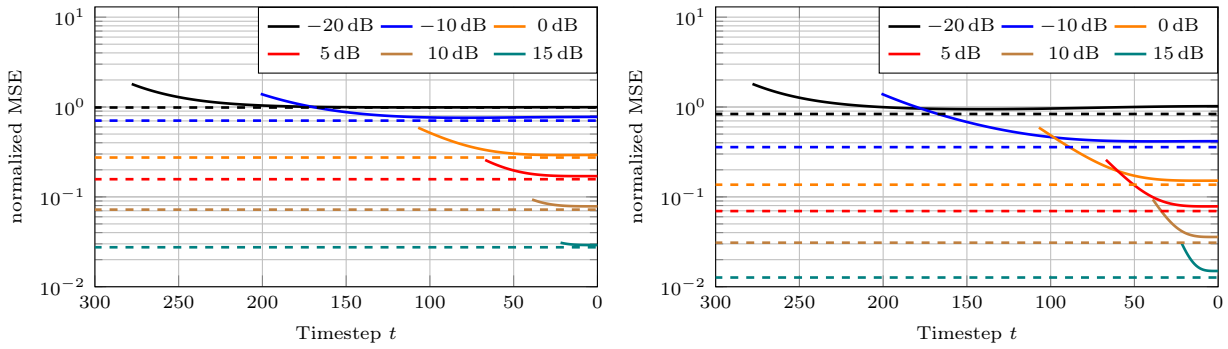

Figure 3: Comparison of the DM (solid) with the CME (dashed) after each timestep t with $T = 300$ for the **random** (left) and **pre-trained** GMM based on **MNIST** (right) with $K = 128$ components.

Astonishingly, a mismatch in the SNR information up to ±3dB has negligible impact on the performance, highlighting the excellent robustness of the proposed DM-based denoiser. In contrast, the re-sampling based DM procedure is far off the CME, which, in turn, validates Proposition H.1. Figure 2 (top right) shows that already a low number T is sufficient for the DM-based denoiser (solid curves) to achieve a performance close to the CME (dashed curves), especially in the low and high SNR regime. In addition, Figure 3 (left) shows that the MSE after each timestep is smoothly decreasing for increasing t, converging to the CME for all SNR values. In high SNR, the convergence is faster in accordance with the theoretical results of Theorem 4.4.

Pre-trained GMM. To ensure that the convergence behavior is also valid for distributions of natural signals, we pre-train a GMM using the implementation of Pedregosa et al. (2011) on image data, i.e., MNIST (LeCun et al., 2010) and Fashion-MNIST (Xiao et al., 2017), and on audio data, i.e., the Librispeech dataset (Panayotov et al., 2015); afterward, we utilize this GMM as ground-truth distribution. Since GMMs are universal approximators (Nguyen et al., 2020), it is reasonable to assume that the important features and the structure of the data are captured by the GMM, yielding a more practically relevant ground-truth distribution for which the CME can still be computed in closed form via (33). We thus train a GMM on the vectorized $28 \times 28 = 748$-dimensional image data with $K = 128$ components and subsequently sample the training and test data from this GMM. Further details and the results of the Fashion-MNIST and Librispeech data as well as qualitative results can be found in Appendix J.2 and Appendix J.5, respectively.

Figure 2 (bottom left) validates that the proposed DM-based denoiser with $T = 300$ is also very close to the ground-truth CME for the MNIST data, validating the theoretical results also for distributions of natural signals. Further, the impact of a mismatch in the SNR information (±3 dB) is negligible. The results regarding the convergence behavior over the number of DM timesteps T in Figure 2 (bottom right) are in alignment with the ones for the random GMM. However, a larger number of timesteps T is necessary until saturation, which can be reasoned by the increased dimensionality and the more structured data. The same argumentation holds for the results of the DM's intermediate estimates in Figure 3 (right), where the fast convergence in the high SNR is particularly noticeable.

6 CONCLUSION

This work has proposed a novel denoising algorithm utilizing pre-trained DMs, inspired by the structure of the Bayesian-optimal estimator, with provable polynomial-time convergence to the MSE-optimal CME. As part of the optimality analysis, a new Lipschitz constant of the DM network under mild assumptions was derived. Further, we connected the convergence to the prior distribution with the convergence to the CME and offered a novel perspective on the concurrent generation and denoising abilities of DMs. A thorough experimental evaluation on various benchmark datasets (1D, 2D, real-/complex-valued, image, speech, and generic data) validated the theoretical results. Furthermore, the proposed denoiser was successfully applied to wireless channel estimation by Fesl et al. (2024).

Limitations. So far, the theoretical analysis is concentrated on the AWGN case, outlining extensions to different noise models (Xie et al., 2023) and general inverse problems (Fabian et al., 2024; Chung et al., 2022; Meng and Kabashima, 2024). In addition, we note that the proposed denoiser is not optimized for perceptual quality due to the perception-distortion trade-off (Blau and Michaeli, 2018) but can serve the purpose of a better understanding of the same.

References

Akyol, E., Viswanatha, K., and Rose, K. (2012). On conditions for linearity of optimal estimation. *IEEE Transactions on Information Theory*, 58(6):3497–3508.

Arvinte, M. and Tamir, J. I. (2023). MIMO channel estimation using score-based generative models. *IEEE Transactions on Wireless Communications*, 22(6):3698–3713.

Asoodeh, S., Diaz, M., Alajaji, F., and Linder, T. (2016). Information extraction under privacy constraints. *Information*, 7(1).

Benton, J., Bortoli, V. D., Doucet, A., and Deligiannidis, G. (2024). Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*.

Blau, Y. and Michaeli, T. (2018). The perception-distortion tradeoff. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237.

Block, A., Mroueh, Y., and Rakhlin, A. (2022). Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint arXiv:2002.0010*.

Bortoli, V. D. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*.

Bortoli, V. D., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*.

Callahan, J. J. (2010). *Advanced Calculus: A Geometric View*. Springer New York, NY.

Carlini, N., Tramer, F., Dvijotham, K. D., Rice, L., Sun, M., and Kolter, J. Z. (2023). (certified!!) adversarial robustness for free! In *The Eleventh International Conference on Learning Representations*.

Chen, H., Lee, H., and Lu, J. (2023a). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2023b). The probability flow ODE is provably fast. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. (2023c). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*.

Chen, S., Daras, G., and Dimakis, A. G. (2023d). Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. In *Proceedings of the 40 th International Conference on Machine Learning*.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory: Independence, Interchangeability, Martingales*. Springer NY, 3rd edition.

Chung, H., Sim, B., and Ye, J. C. (2022). Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12403–12412.

Chung, K. and Walsh, J. (2005). *Markov processes, Brownian motion, and time symmetry*. Springer New York, 2nd edition.

Conforti, G., Durmus, A., and Silveri, M. G. (2025). KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109.

Delbracio, M. and Milanfar, P. (2023). Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *Transactions on Machine Learning Research*.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794.

Diaz, M., Kairouz, P., Liao, J., and Sankar, L. (2021). Neural network-based estimation of the MMSE. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1023–1028.

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307.

Dytso, A., Poor, H. V., and Shamai Shitz, S. (2020). A general derivative identity for the conditional mean estimator in Gaussian noise and some applications. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1183–1188.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

Elad, M., Kawar, B., and Vaksman, G. (2023). Image denoising: The deep learning revolution and beyond—

a survey paper. *SIAM Journal on Imaging Sciences*, 16(3):1594–1654.

Fabian, Z., Tinaz, B., and Soltanolkotabi, M. (2024). DiracDiffusion: Denoising and incremental reconstruction with assured data-consistency. In *Proceedings of Machine Learning Research*, volume 235, pages 12754–12783.

Feller, W. (1949). On the theory of stochastic processes, with particular reference to applications. In *Proceedings of the [First] Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 403–433.

Fesl, B., Baur, M., Strasser, F., Joham, M., and Utschick, W. (2024). Diffusion-based generative prior for low-complexity MIMO channel estimation. *IEEE Wireless Communications Letters*, 13(12):3493–3497.

Fesl, B., Koller, M., and Utschick, W. (2023). On the mean square error optimal estimator in one-bit quantized systems. *IEEE Transactions on Signal Processing*, 71:1968–1980.

Gargiulo, G. D. and McEwan, A. (2011). *Applied Biomedical Engineering*. IntechOpen, Rijeka.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, page 6840–6851.

Huang, Y., Huang, J., Liu, J., Yan, M., Dong, Y., Lv, J., Chen, C., and Chen, S. (2024). WaveDM: Wavelet-based diffusion models for image restoration. *IEEE Transactions on Multimedia*, 26:7058–7073.

Huy, P. N. and Minh Quan, T. (2023). Denoising diffusion medical models. In *IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.

Kadkhodaie, Z. and Simoncelli, E. (2021). Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In *Advances in Neural Information Processing Systems*, volume 34, pages 13242–13254.

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Kazerouni, A., Kamilov, U. S., Bostan, E., and Unser, M. (2013). Bayesian denoising: From MAP to MMSE using consistent cycle spinning. *IEEE Signal Processing Letters*, 20(3):249–252.

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. (2024). Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*.

Kim, K. and Ye, J. C. (2021). Noise2score: Tweedie's approach to self-supervised image denoising without clean images. In *Advances in Neural Information Processing Systems*.

Kingma, D., Salimans, T., Poole, B., and Ho, J. (2021). Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pages 21696–21707.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Koller, M., Fesl, B., Turan, N., and Utschick, W. (2022). An asymptotically MSE-optimal estimator based on Gaussian mixture models. *IEEE Transactions on Signal Processing*, 70:4109–4123.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021). Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*.

LeCun, Y., Cortes, C., and Burges, C. (2010). MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2. CC-BY-SA 3.0 license.

Lee, H., Lu, J., and Tan, Y. (2023). Convergence of score-based generative modeling for general data distributions. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pages 946–985.

Lehmann, E. and Casella, G. (2006). *Theory of Point Estimation*. Springer New York, 2nd edition.

Li, G., Wei, Y., Chen, Y., and Chi, Y. (2024). Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*.

Liang, Y., Ju, P., Liang, Y., and Shroff, N. (2025). Broadening target distributions for accelerated diffusion models via a novel analysis approach. In *The Thirteenth International Conference on Learning Representations*.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice*. CRC Press, Inc., USA, 2nd edition.

Lu, C., Zhou, Y., Bao, F., Chen, J., LI, C., and Zhu, J. (2022). DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pages 5775–5787.

Luo, C. (2022). Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*.

Meng, X. and Kabashima, Y. (2024). Diffusion model based posterior sampling for noisy linear inverse problems. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Mohan, S., Kadkhodaie, Z., Simoncelli, E. P., and Fernandez-Granda, C. (2020). Robust and interpretable blind image denoising via bias-free convolutional neural networks. In *International Conference on Learning Representations*.

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics,*, 7(1):1750861.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8162–8171.

Om, H. and Biswas, M. (2014). MMSE based MAP estimation for image denoising. *Optics & Laser Technology*, 57:252–264.

Ongie, G., Jalal, A., Metzler, C. A., Baraniuk, R. G., Dimakis, A. G., and Willett, R. (2020). Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. CC-BY 4.0 license.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. BSD 3-Clause license.

Pedrotti, F., Maas, J., and Mondelli, M. (2024). Improved convergence of score-based diffusion models via prediction-correction. *Transactions on Machine Learning Research*.

Romano, Y., Elad, M., and Milanfar, P. (2017). The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2023). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726.

Scharf, L. and Demeure, C. (1991). *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Pearson.

Shah, K., Chen, S., and Klivans, A. (2023). Learning mixtures of Gaussians using the DDPM objective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265.

Song, J., Meng, C., and Ermon, S. (2021a). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

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

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021b). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Tai, W., Zhou, F., Trajcevski, G., and Zhong, T. (2023). Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13627–13635.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Venkatakrishnan, S. V., Bouman, C. A., and Wohlberg, B. (2013). Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, pages 945–948.

Welker, S., Richter, J., and Gerkmann, T. (2022). Speech enhancement with score-based generative models in the complex STFT domain. In *Proceedings of Interspeech*, pages 2928–2932.

Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A. G., and Milanfar, P. (2022). Deblurring via stochastic refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16272–16282.

Xiao, C., Chen, Z., Jin, K., Wang, J., Nie, W., Liu, M., Anandkumar, A., Li, B., and Song, D. (2023). Densepure: Understanding diffusion models for adversarial robustness. In *The Eleventh International Conference on Learning Representations*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*. MIT license.

Xie, Y., Yuan, M., Dong, B., and Li, Q. (2023). Diffusion model for generative image denoising. *arXiv preprint arXiv:2302.02398*.

Yang, J., Liao, X., Yuan, X., Llull, P., Brady, D. J., Sapiro, G., and Carin, L. (2015). Compressive sensing by learning a Gaussian mixture model from measurements. *IEEE Transactions on Image Processing*, 24(1):106–119.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4).

Yang, Z., Feng, R., Zhang, H., Shen, Y., Zhu, K., Huang, L., Zhang, Y., Liu, Y., Zhao, D., Zhou, J., and Cheng, F. (2024). Lipschitz singularities in diffusion models. In *The Twelfth International Conference on Learning Representations*.

Zagoruyko, S. and Komodakis, N. (2017). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57.

Zhou, Z., Chen, D., Wang, C., and Chen, C. (2024). Fast ODE-based sampling for diffusion models in around 5 steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7777–7786.

Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., and Gool, L. V. (2023). Denoising diffusion models for plug-and-play image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1219–1229.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes] We provide a problem formulation and a concise introduction of the DM in Section 3, a detailed description of the mathematical setup and all assumptions in Section 4, and an algorithmic summary in Algorithm 1.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Yes] We provide a runtime and parameter analysis depending on the sample size in Appendix J.3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [Yes] The source code including all specifications and dependencies is provided in an open-source GitHub repository, see Section 5.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes] For every theoretical result in Section 4, we provide the full set of assumptions.
   (b) Complete proofs of all theoretical results.
   [Yes] For every theoretical result, we provide a proof sketch in the main paper and a rigorous proof in the appendix.
   (c) Clear explanations of any assumptions.
   [Yes] We provide clear statements of all assumptions, including references for similar choices and claims.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).

[Yes] We provide the simulation code, including the pre-trained models, to reproduce all experimental results in an open-source GitHub repository.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
[Yes] All training details are specified in Section 5, Appendix I, Appendix J, and can be found in the simulation code in the open-source GitHub repository.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
[Yes] We used a reasonably high number of test samples, see Section 5.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
[Yes] We provided the type of GPU, runtime, and the number of parameters of our models for all experiments in Appendix J.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets.
[Yes] We properly cite all owners of code, data, and models (see Section 5, Appendix I, and Appendix J).

(b) The license information of the assets, if applicable.
[Yes] We added the license information for all assets in the references.

(c) New assets either in the supplemental material or as a URL, if applicable.
[Yes] We provide the simulation code in the referenced open-source GitHub repository.

(d) Information about consent from data providers/curators.
[Not applicable] We do not use data that needs consent from providers/curators.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
[Not applicable] The paper contains no sensible content.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots.
[Not applicable] We have not used crowdsourcing or conducted research with human subjects.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
[Not applicable] We have not used crowdsourcing or conducted research with human subjects.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
[Not applicable] We have not used crowdsourcing or conducted research with human subjects.

A Training of the Diffusion Model

The training of the DM is performed by maximizing the ELBO on the log-likelihood $\log p(\boldsymbol{x}_0)$ of the form (Ho et al., 2020)

$$\mathbb{E}_q[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) - \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T)) - \sum_{t=2}^{T} \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)]. \tag{34}$$

The first term is called the *reconstruction term*, which addresses the first latent step. The second term is the *prior matching term* that measures how close the last DM step is to pure noise. Since it has no trainable parameters, it can be ignored during training. The third term is the *denoising matching term*, which is designed to match the reverse process transition in (7) to the tractable posterior (8) for each timestep of the DM. Since both distributions are Gaussian, a straightforward implementation of the t-th summand of the denoising matching term is to minimize the MSE between their conditional first moments (Ho et al., 2020), i.e.,

$$\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0) - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|_2^2\right]. \tag{35}$$

However, it was observed by Ho et al. (2020) that reparameterizing

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{x}_t, t)\right) \tag{36}$$

to predict the noise rather than the conditional mean leads to better results, yielding the following expression for the tth summand in the denoising matching term of the ELBO:

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\frac{(1-\alpha_t)^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\|\boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{x}_t, t)\|_2^2\right]. \tag{37}$$

For further details about the training procedure of the DM, we refer to Ho et al. (2020).

B Proof of Lemma 4.1

By the multivariate mean value theorem (Callahan, 2010) we have that for any convex and compact $\mathcal{K} \subseteq \mathbb{R}^N$

$$\|f_{\boldsymbol{\theta},t}(\boldsymbol{a}) - f_{\boldsymbol{\theta},t}(\boldsymbol{b})\| \leq L_t\|\boldsymbol{a} - \boldsymbol{b}\|, \quad \boldsymbol{a}, \boldsymbol{b} \in \mathcal{K}, \tag{38}$$

$$L_t = \sup_{\boldsymbol{x}\in\mathcal{K}} \|\boldsymbol{J}_{\boldsymbol{x}} f_{\boldsymbol{\theta},t}(\boldsymbol{x})\|, \tag{39}$$

where $\boldsymbol{J}_{\boldsymbol{x}} f_{\boldsymbol{\theta},t}(\boldsymbol{x})$ is the Jacobian of $f_{\boldsymbol{\theta},t}$ at the point $\boldsymbol{x}$. Noticing that $\boldsymbol{J}_{\boldsymbol{x}} f_{\boldsymbol{\theta},t}(\boldsymbol{x})$ is the derivative of the CME with respect to the parameterized distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, we can use the derivative identity for the CME from Dytso et al. (2020). By substituting $\tilde{\boldsymbol{x}}_{t-1} = \sqrt{\alpha_t}\boldsymbol{x}_{t-1}$ and $\tilde{\boldsymbol{\epsilon}}_{t-1} = \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}$ in (6), we have $\boldsymbol{x}_t = \tilde{\boldsymbol{x}}_{t-1} + \tilde{\boldsymbol{\epsilon}}_{t-1}$ with the conditional covariance $\boldsymbol{C}_{\tilde{\boldsymbol{x}}_{t-1}|\boldsymbol{x}_t} = \alpha_t\sigma_t^2\,\mathbf{I}$, being constant in $\boldsymbol{x}_t$ and $\tilde{\boldsymbol{\epsilon}}_{t-1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_{\tilde{\boldsymbol{\epsilon}}_{t-1}} = (1-\alpha_t)\,\mathbf{I})$ for all $t = 2, \ldots, T$. Thus, by plugging in the definition of σ_t^2 from (9) and taking into account the substitution, we get the identity, cf. Dytso et al. (2020),

$$\boldsymbol{J}_{\boldsymbol{x}} f_{\boldsymbol{\theta},t}(\boldsymbol{x}) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{J}_{\boldsymbol{x}}\mathbb{E}_{p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{t-1}|\boldsymbol{x}_t)}[\tilde{\boldsymbol{x}}_{t-1}|\boldsymbol{x}] = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{C}_{\tilde{\boldsymbol{x}}_{t-1}|\boldsymbol{x}_t}\boldsymbol{C}_{\tilde{\boldsymbol{\epsilon}}_{t-1}}^{-1} \tag{40}$$

$$= \frac{\sqrt{\alpha_t}\sigma_t^2}{1-\alpha_t}\,\mathbf{I} = \sqrt{\alpha_t}\frac{1-\alpha_t}{1-\alpha_t}\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\mathbf{I} \tag{41}$$

$$= \sqrt{\alpha_t}\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\mathbf{I} \tag{42}$$

for all $\boldsymbol{x} \in \mathcal{K}$. Since the induced norm of the identity matrix is always one, we get the result in (18) where the supremum can be dropped since the Jacobian is constant for all $\boldsymbol{x} \in \mathcal{K}$. Further, since the SNR is strictly decreasing with increasing t such that $1 - \bar{\alpha}_{t-1} < 1 - \bar{\alpha}_t$, it follows that $L_t < 1$ for all $t = 2, \ldots, T$. The

concatenation of Lipschitz functions is again Lipschitz with the product of the individual Lipschitz constants. Thus, we have

$$L_{\tau_1:\tau_2} = \prod_{i=\tau_1}^{\tau_2} L_i = \prod_{i=\tau_1}^{\tau_2} \sqrt{\alpha_i}\frac{1-\bar{\alpha}_{i-1}}{1-\bar{\alpha}_i} = \frac{1-\bar{\alpha}_{\tau_1-1}}{1-\bar{\alpha}_{\tau_2}} \prod_{i=\tau_1}^{\tau_2} \sqrt{\alpha_i} = \frac{(1-\bar{\alpha}_{\tau_1-1})\sqrt{\bar{\alpha}_{\tau_2}}}{(1-\bar{\alpha}_{\tau_2})\sqrt{\bar{\alpha}_{\tau_1-1}}} < 1. \tag{43}$$

The third equation follows because of the telescope product, where every numerator except the first cancels out with every denominator except the last one.

C Lipschitz Constant for Noise Estimation

Although we have derived the Lipschitz constant for $f_{\boldsymbol{\theta},t}(\boldsymbol{x}_t) := \boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ in Lemma 4.1, it implicitly depends on the chosen parameterization (e.g., the estimation of the noise part via $\boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{x}_t, t)$ in (37)). Consequently, a slightly modified Lipschitz constant can be straightforwardly computed for $\boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{x}_t, t)$. This can be seen by plugging $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ from (36) into (18), which yields

$$\begin{aligned} L_t\|\boldsymbol{a}-\boldsymbol{b}\| &\geq \|f_{\boldsymbol{\theta},t}(\boldsymbol{a}) - f_{\boldsymbol{\theta},t}(\boldsymbol{b})\| && (44)\\ &= \frac{1}{\sqrt{\alpha_t}}\|\boldsymbol{a}-\boldsymbol{b} + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}(\boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{b},t) - \boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{a},t))\| && (45)\\ &\geq \frac{1}{\sqrt{\alpha_t}}\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\|\boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{a},t) - \boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{b},t)\| - \frac{1}{\sqrt{\alpha_t}}\|\boldsymbol{a}-\boldsymbol{b}\| && (46) \end{aligned}$$

Thus, we get

$$\|\boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{a},t) - \boldsymbol{\epsilon}_{\boldsymbol{\theta},0}(\boldsymbol{b},t)\| \leq \tilde{L}_t\|\boldsymbol{a}-\boldsymbol{b}\|, \quad \tilde{L}_t = \frac{\sqrt{1-\bar{\alpha}_t}}{1-\alpha_t}(\sqrt{\alpha_t}L_t + 1). \tag{47}$$

However, for the later results, the Lipschitz constant with respect to $f_{\boldsymbol{\theta},t}(\boldsymbol{x}_t)$ is primarily used, simplifying the expressions.

D Connection of the DM Denoiser's Lipschitz Constant to the SNR

The DM estimator $f_{\boldsymbol{\theta},1:\hat{t}}$ in (13) has the Lipschitz constant $L_{1:\hat{t}} = L_1 L_{2:\hat{t}}$. Utilizing the SNR representation of the DM (11) allows to reformulate

$$L_{2:\hat{t}} = \frac{(1-\alpha_1)\sqrt{\bar{\alpha}_{\hat{t}}}}{(1-\bar{\alpha}_{\hat{t}})\sqrt{\alpha_1}} = \frac{1-\alpha_1}{\sqrt{\bar{\alpha}_{\hat{t}}\alpha_1}}\mathrm{SNR}_{\mathrm{DM}}(\hat{t}) = \frac{1-\alpha_1}{\sqrt{\alpha_1}}\sqrt{\mathrm{SNR}_{\mathrm{DM}}(\hat{t})\left(\mathrm{SNR}_{\mathrm{DM}}(\hat{t})+1\right)}. \tag{48}$$

Interestingly, the Lipschitz constant is monotonically decreasing with decreasing SNR. Since the Lipschitz constant is related to the impact of the observation onto the estimated value, the SNR-dependency of the DM denoiser's Lipschitz constant resembles a well-known property of the CME, i.e., the optimization of the bias-variance trade-off.

In particular, for a low SNR value, i.e., $\eta \to \infty$, the impact of the observation onto the estimated value is vanishing since $\lim_{\eta\to\infty} \mathrm{SNR}_{\mathrm{DM}}(\hat{t}) = 0$, cf. (12), and thus $\lim_{\eta\to\infty} L_{2:\hat{t}} = 0$, i.e., the DM estimator (13) effectively becomes a constant, regardless of the observation. The same holds for the CME, which yields the prior mean in the asymptotically low SNR regime, indicating a vanishing variance error at the expense of a higher bias, optimizing the bias-variance trade-off of the MSE (Kay, 1993, Ch. 10). In contrast, for a high SNR value, the observation has a high impact on the estimated value, indicated by a larger Lipschitz constant in (48). In Appendix J.4, these findings are experimentally validated.

E Proof of Theorem 4.2

Building upon the error decomposition (22), we bound the first term by utilizing Tweedie's formula (Efron, 2011) and the assumption (24) as

$$\left\|\mathbb{E}^{\boldsymbol{\theta}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}]\right\| = \eta^2\|\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{y}) - \nabla \log p(\boldsymbol{y})\| \leq \eta^2\Xi, \tag{49}$$

yielding the first summand in (25).

Regarding the second term in (22) attributed to the denoising procedure, we start by rewriting the parameterized CME (21) in terms of the DM estimator with a residual error by iteratively applying the law of total expectation and utilizing the Markov property of the DM, similar as in (15), where we denote $\mathbb{E}^{\boldsymbol{\theta}}_{t-1|t} := \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}$ and $f_{\boldsymbol{\theta},t}(\boldsymbol{x}_t) := \mathbb{E}^{\boldsymbol{\theta}}_{t-1|t}[\boldsymbol{x}_{t-1}|\boldsymbol{x}_t]$ for notational convenience:

$$\mathbb{E}^{\boldsymbol{\theta}}_{0|\hat{t}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] = \mathbb{E}^{\boldsymbol{\theta}}_{\hat{t}-1|\hat{t}}\left[\cdots \mathbb{E}^{\boldsymbol{\theta}}_{2|3}\left[\mathbb{E}^{\boldsymbol{\theta}}_{1|2}\left[\mathbb{E}^{\boldsymbol{\theta}}_{0|1}\left[\boldsymbol{x}_0|\boldsymbol{x}_1\right]|\boldsymbol{x}_2\right]|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] \tag{50}$$

$$= \mathbb{E}^{\boldsymbol{\theta}}_{\hat{t}-1|\hat{t}}\left[\cdots \mathbb{E}^{\boldsymbol{\theta}}_{2|3}\left[\mathbb{E}^{\boldsymbol{\theta}}_{1|2}\left[f_{\boldsymbol{\theta},1}(\boldsymbol{x}_1)|\boldsymbol{x}_2\right]|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] \tag{51}$$

$$= \mathbb{E}^{\boldsymbol{\theta}}_{\hat{t}-1|\hat{t}}\left[\cdots \mathbb{E}^{\boldsymbol{\theta}}_{2|3}\left[\mathbb{E}^{\boldsymbol{\theta}}_{1|2}\left[f_{\boldsymbol{\theta},1}(\boldsymbol{x}_1)|\boldsymbol{x}_2\right] + f_{\boldsymbol{\theta},1:2}(\boldsymbol{x}_2) - f_{\boldsymbol{\theta},1:2}(\boldsymbol{x}_2)|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] \tag{52}$$

$$= \mathbb{E}^{\boldsymbol{\theta}}_{\hat{t}-1|\hat{t}}\left[\cdots \mathbb{E}^{\boldsymbol{\theta}}_{2|3}\left[f_{\boldsymbol{\theta},1:2}(\boldsymbol{x}_2)|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] + \mathbb{E}^{\boldsymbol{\theta}}_{1|\hat{t}}\left[f_{\boldsymbol{\theta},1}(\boldsymbol{x}_1)|\boldsymbol{x}_{\hat{t}}\right] - \mathbb{E}^{\boldsymbol{\theta}}_{2|\hat{t}}\left[f_{\boldsymbol{\theta},1:2}(\boldsymbol{x}_2)|\boldsymbol{x}_{\hat{t}}\right] \tag{53}$$

$$= f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}}) + \sum_{t=1}^{\hat{t}-1} \mathbb{E}^{\boldsymbol{\theta}}_{t|\hat{t}}\left[f_{\boldsymbol{\theta},1:t}(\boldsymbol{x}_t)|\boldsymbol{x}_{\hat{t}}\right] - \mathbb{E}^{\boldsymbol{\theta}}_{t+1|\hat{t}}\left[f_{\boldsymbol{\theta},1:t+1}(\boldsymbol{x}_{t+1})|\boldsymbol{x}_{\hat{t}}\right]. \tag{54}$$

Note that the residual error is the sum of the corresponding Jensen gaps for each step. Plugging the above result into the second term in (22), we get:

$$\left\|\mathbb{E}^{\boldsymbol{\theta}}_{0|\hat{t}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\right\| = \left\|\sum_{t=1}^{\hat{t}-1} \mathbb{E}^{\boldsymbol{\theta}}_{t|\hat{t}}\left[f_{\boldsymbol{\theta},1:t}(\boldsymbol{x}_t)|\boldsymbol{x}_{\hat{t}}\right] - \mathbb{E}^{\boldsymbol{\theta}}_{t+1|\hat{t}}\left[f_{\boldsymbol{\theta},1:t+1}(\boldsymbol{x}_{t+1})|\boldsymbol{x}_{\hat{t}}\right]\right\| \tag{55}$$

$$= \left\|\sum_{t=1}^{\hat{t}-1} \mathbb{E}^{\boldsymbol{\theta}}_{t+1|\hat{t}}\left[\mathbb{E}^{\boldsymbol{\theta}}_{t|t+1}\left[f_{\boldsymbol{\theta},1:t}(\boldsymbol{x}_t)|\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] - \mathbb{E}^{\boldsymbol{\theta}}_{t+1|\hat{t}}\left[f_{\boldsymbol{\theta},1:t+1}(\boldsymbol{x}_{t+1})|\boldsymbol{x}_{\hat{t}}\right]\right\| \tag{56}$$

$$= \left\|\sum_{t=1}^{\hat{t}-1} \mathbb{E}^{\boldsymbol{\theta}}_{t+1|\hat{t}}\left[\mathbb{E}^{\boldsymbol{\theta}}_{t|t+1}\left[f_{\boldsymbol{\theta},1:t}(\boldsymbol{x}_t) - f_{\boldsymbol{\theta},1:t+1}(\boldsymbol{x}_{t+1})|\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right]\right\| \tag{57}$$

$$\leq \sum_{t=1}^{\hat{t}-1} \mathbb{E}^{\boldsymbol{\theta}}_{t+1|\hat{t}}\left[\mathbb{E}^{\boldsymbol{\theta}}_{t|t+1}\left[\|f_{\boldsymbol{\theta},1:t}(\boldsymbol{x}_t) - f_{\boldsymbol{\theta},1:t+1}(\boldsymbol{x}_{t+1})\|\,|\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] \tag{58}$$

$$\leq \sum_{t=1}^{\hat{t}-1} L_{1:t}\mathbb{E}^{\boldsymbol{\theta}}_{t+1|\hat{t}}\left[\mathbb{E}^{\boldsymbol{\theta}}_{t|t+1}\left[\left\|\boldsymbol{x}_t - \mathbb{E}^{\boldsymbol{\theta}}_{t|t+1}[\boldsymbol{x}_t|\boldsymbol{x}_{t+1}]\right\||\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] \tag{59}$$

$$= \sum_{t=1}^{\hat{t}-1} L_{1:t}\sigma_{t+1}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}\|\right] \tag{60}$$

$$\leq \sum_{t=1}^{\hat{t}-1} L_{1:t}\sigma_{t+1}N \tag{61}$$

where in (56), we used the law of total expectation, (58) follows from the triangle inequality and Jensen's inequality in combination with the convexity of the norm, and (59) follows from the Lipschitz continuity, cf. Lemma 4.1. The identity in (60) follows from the reparameterization

$$\boldsymbol{x}_t|\boldsymbol{x}_{t+1},\boldsymbol{\theta} = \mathbb{E}^{\boldsymbol{\theta}}_{t|t+1}[\boldsymbol{x}_t|\boldsymbol{x}_{t+1}] + \sigma_{t+1}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{62}$$

and the fact that the conditional variance is a time-dependent constant. In (61), we use that $\mathbb{E}[\|\boldsymbol{\epsilon}\|_p] < N$ holds for any $1 \leq p \leq \infty$. Plugging in the Lipschitz constant from Lemma 4.1 and the definition of the conditional variance (9) into (61), we get

$$\sum_{t=1}^{\hat{t}-1} L_{1:t}\sigma_{t+1}N = NL_1\sum_{t=1}^{\hat{t}-1} L_{2:t}\sigma_{t+1} = NL_1\sum_{t=1}^{\hat{t}-1}\frac{1-\alpha_1}{1-\bar{\alpha}_t}\sqrt{\frac{\bar{\alpha}_t}{\alpha_1}(1-\alpha_{t+1})\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t+1}}} \tag{63}$$

$$\leq NL_1 \sum_{t=1}^{\hat{t}-1} \frac{(1-\alpha_1)\sqrt{\bar{\alpha}_t}\sqrt{1-\alpha_{t+1}}}{\sqrt{\alpha_1}(1-\bar{\alpha}_t)} \tag{64}$$

$$\leq NL_1 \sum_{t=1}^{\hat{t}-1} \frac{(1-\alpha_1)\sqrt{\alpha_1^t}\sqrt{1-\alpha_{t+1}}}{\sqrt{\alpha_1}(1-\alpha_1^t)} \tag{65}$$

$$= NL_1 \sum_{t=1}^{\hat{t}-1} \frac{\beta_1\sqrt{(1-\beta_1)^{t-1}}\sqrt{\beta_{t+1}}}{1-(1-\beta_1)^t} \tag{66}$$

where in (64) we used $1-\bar{\alpha}_t \leq 1-\bar{\alpha}_{t+1}$, and in (65) we bounded $\bar{\alpha}_t \leq \alpha_1^t$, both following from the ordering $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_T$, cf. (23). By using a variant of Bernoulli's inequality of the form

$$(1-\beta_1)^t \leq \frac{1}{1+t\beta_1}, \tag{67}$$

we can further write

$$NL_1 \sum_{t=1}^{\hat{t}-1} \frac{\beta_1\sqrt{(1-\beta_1)^{t-1}}\sqrt{\beta_{t+1}}}{1-(1-\beta_1)^t} \leq \lim_{T\to\infty} NL_1 \sum_{t=1}^{\hat{t}-1} \frac{\beta_1\sqrt{\beta_{t+1}}(1+t\beta_1)}{(1+\frac{t-1}{2}\beta_1)t\beta_1} \tag{68}$$

$$\leq \lim_{T\to\infty} 2NL_1 \sum_{t=1}^{\hat{t}-1} \frac{\sqrt{\beta_{t+1}}}{t} \tag{69}$$

$$= 2NL_1\mathcal{O}(T^{-\gamma/2}) \sum_{t=1}^{\hat{t}-1} \frac{1}{t} \tag{70}$$

where we bounded

$$1+t\beta_1 \leq 2(1+\tfrac{t-1}{2}\beta_1) \tag{71}$$

and used that $\beta_t = \mathcal{O}\left(T^{-\gamma}\right)$, cf. (23). Further, using the bound on the harmonic number

$$\sum_{t=1}^{\hat{t}-1} \frac{1}{t} \leq 1+\log\hat{t} \tag{72}$$

results in (25). The result in (26) is straightforward using (25) and (Koller et al., 2022, Theorem 2), finishing the proof.

F Proof of Theorem 4.4

In the following, we denote $\mathbb{E}_{t-1|t} := \mathbb{E}_{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}$ for notational convenience. Let us further define $g_t(\boldsymbol{x}_t) := \mathbb{E}_{t-1|t}[\boldsymbol{x}_{t-1}|\boldsymbol{x}_t]$. Then, utilizing the definition of the DM's NN function $f_{\boldsymbol{\theta},t}$ in (28), we can conclude that

$$g_t(\boldsymbol{x}_t) = \mathbb{E}_{t-1|t}[\boldsymbol{x}_{t-1}|\boldsymbol{x}_t] = \mathbb{E}_{0|t}[\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t,\boldsymbol{x}_0)|\boldsymbol{x}_t] \tag{73}$$

$$= f_{\boldsymbol{\theta},t}(\boldsymbol{x}_t) + \mathbb{E}_{0|t}[\boldsymbol{\delta}_t(\boldsymbol{x}_t,\boldsymbol{x}_0)|\boldsymbol{x}_t] \tag{74}$$

where we used the law of total expectation. We now rewrite the ground-truth CME $\mathbb{E}_{0|\hat{t}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}]$ in terms of a concatenation of the functions g_t, $t=1,\ldots,\hat{t}$, plus an error term by iteratively applying the law of total expectation and utilizing the Markov property of the DM, similar as in (15), i.e.,

$$\mathbb{E}_{0|\hat{t}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] = \mathbb{E}_{1|\hat{t}}[\mathbb{E}_{0|1,\hat{t}}[\boldsymbol{x}_0|\boldsymbol{x}_1,\boldsymbol{x}_{\hat{t}}]|\boldsymbol{x}_{\hat{t}}] = \mathbb{E}_{1|\hat{t}}[\mathbb{E}_{0|1}[\boldsymbol{x}_0|\boldsymbol{x}_1]|\boldsymbol{x}_{\hat{t}}] \tag{75}$$

$$= \mathbb{E}_{\hat{t}-1|\hat{t}}\left[\cdots\mathbb{E}_{2|3}\left[\mathbb{E}_{1|2}\left[\mathbb{E}_{0|1}\left[\boldsymbol{x}_0|\boldsymbol{x}_1\right]|\boldsymbol{x}_2\right]|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] \tag{76}$$

$$= \mathbb{E}_{\hat{t}-1|\hat{t}}\left[\cdots\mathbb{E}_{2|3}\left[\mathbb{E}_{1|2}\left[g_1(\boldsymbol{x}_1)|\boldsymbol{x}_2\right]|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] \tag{77}$$

$$= \mathbb{E}_{\hat{t}-1|\hat{t}}\left[\cdots\mathbb{E}_{2|3}\left[\mathbb{E}_{1|2}\left[g_1(\boldsymbol{x}_1)|\boldsymbol{x}_2\right]+g_{1:2}(\boldsymbol{x}_2)-g_{1:2}(\boldsymbol{x}_2)|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] \tag{78}$$

$$= \mathbb{E}_{\hat{t}-1|\hat{t}}\left[\cdots \mathbb{E}_{2|3}\left[g_{1:2}(\boldsymbol{x}_2)|\boldsymbol{x}_3\right]\cdots|\boldsymbol{x}_{\hat{t}}\right] + \mathbb{E}_{1|\hat{t}}\left[g_1(\boldsymbol{x}_1)|\boldsymbol{x}_{\hat{t}}\right] - \mathbb{E}_{2|\hat{t}}\left[g_{1:2}(\boldsymbol{x}_2)|\boldsymbol{x}_{\hat{t}}\right] \tag{79}$$

$$= g_{1:\hat{t}}(\boldsymbol{x}_{\hat{t}}) + \underbrace{\sum_{t=1}^{\hat{t}-1} \mathbb{E}_{t|\hat{t}}\left[g_{1:t}(\boldsymbol{x}_t)|\boldsymbol{x}_{\hat{t}}\right] - \mathbb{E}_{t+1|\hat{t}}\left[g_{1:t+1}(\boldsymbol{x}_{t+1})|\boldsymbol{x}_{\hat{t}}\right]}_{\boldsymbol{\delta}_{\mathrm{JG}}}. \tag{80}$$

Note that the additive error term $\boldsymbol{\delta}_{\mathrm{JG}}$ in (80) represents the sum of the corresponding Jensen gaps for each step. Now, we find a bound on the difference between the DM estimator from (13) to the ground-truth CME as

$$\|\mathbb{E}_{0|\hat{t}}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| \leq \|g_{1:\hat{t}}(\boldsymbol{x}_{\hat{t}}) - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| + \|\boldsymbol{\delta}_{\mathrm{JG}}\|. \tag{81}$$

Considering the first term, we utilize (74) to express g_t as a function of $f_{\boldsymbol{\theta},t}$ plus an additive error term. After iteratively applying the Lipschitz property of $f_{\boldsymbol{\theta},t}$, cf. Lemma 4.1, together with the triangle inequality, we find an upper bound as

$$\|g_{1:\hat{t}}(\boldsymbol{x}_{\hat{t}}) - f_{\boldsymbol{\theta},1:\hat{t}}(\boldsymbol{x}_{\hat{t}})\| \leq \sum_{t=1}^{\hat{t}} L_{1:t-1}\|\mathbb{E}_{0|t}[\boldsymbol{\delta}_t(g_{t:\hat{t}}(\boldsymbol{x}_{\hat{t}}), \boldsymbol{x}_0)|g_{t:\hat{t}}(\boldsymbol{x}_{\hat{t}})]\| \leq \Delta \sum_{t=1}^{\hat{t}} L_{1:t-1} \tag{82}$$

with $L_{i:j} := 1$ if $j < i$ and using Assumption 4.3. Using that the Lipschitz constant (18) is a function of the DM's hyperparameters (23), we get

$$\sum_{t=1}^{\hat{t}} L_{1:t-1} = 1 + L_1 \sum_{t=1}^{\hat{t}-1} L_{2:t} = 1 + L_1 \sum_{t=1}^{\hat{t}-1} \frac{(1-\alpha_1)\sqrt{\bar{\alpha}_t}}{(1-\bar{\alpha}_t)\sqrt{\alpha_1}} \tag{83}$$

$$\leq 1 + L_1 \sum_{t=1}^{\hat{t}-1} \frac{(1-\alpha_1)\sqrt{\alpha_1^{t-1}}}{1-\alpha_1^t} = 1 + L_1 \sum_{t=1}^{\hat{t}-1} \frac{\beta_1\sqrt{(1-\beta_1)^{t-1}}}{1-(1-\beta_1)^t} \tag{84}$$

$$\leq 1 + L_1 \sum_{t=1}^{\hat{t}-1} \frac{\beta_1(1+t\beta_1)}{t\beta_1(1+\frac{t-1}{2}\beta_1)} \tag{85}$$

$$\leq 1 + 2L_1 \sum_{t=1}^{\hat{t}-1} \frac{1}{t} \tag{86}$$

$$\leq 1 + 2L_1(1+\log\hat{t}) \tag{87}$$

by using $\bar{\alpha}_t \leq \alpha_1^t$, cf. (23), Bernoulli's inequality (67), and the bound on the harmonic number (72). The second term $\|\boldsymbol{\delta}_{\mathrm{JG}}\|$ in (81) is bounded as (cf. (55))

$$\|\boldsymbol{\delta}_{\mathrm{JG}}\| \leq \sum_{t=1}^{\hat{t}-1} \mathbb{E}_{t+1|\hat{t}}\left[\mathbb{E}_{t|t+1}\left[\|g_{1:t}(\boldsymbol{x}_t) - g_{1:t+1}(\boldsymbol{x}_{t+1})\||\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] \tag{88}$$

$$\begin{aligned} &\leq \sum_{t=1}^{\hat{t}-1} L_{1:t}\mathbb{E}_{t+1|\hat{t}}\left[\mathbb{E}_{t|t+1}\left[\|\boldsymbol{x}_t - g_{t+1}(\boldsymbol{x}_{t+1})\||\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] \\ &\quad + \sum_{t=1}^{\hat{t}-1}\sum_{i=1}^{t} L_{1:i-1}\mathbb{E}_{t+1|\hat{t}}\left[\mathbb{E}_{t|t+1}\left[\|\mathbb{E}_{0|t}[\boldsymbol{\delta}_i(g_{i:t}(\boldsymbol{x}_t), \boldsymbol{x}_0)|g_{i:t}(\boldsymbol{x}_t)]\right.\right. \\ &\quad \left.\left. -\mathbb{E}_{0|t+1}[\boldsymbol{\delta}_i(g_{i:t+1}(\boldsymbol{x}_{t+1}), \boldsymbol{x}_0)|g_{i:t+1}(\boldsymbol{x}_{t+1})]\||\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] \end{aligned} \tag{89}$$

$$\leq \sum_{t=1}^{\hat{t}-1} L_{1:t}\mathbb{E}_{t+1|\hat{t}}\left[\mathbb{E}_{t|t+1}\left[\|\boldsymbol{x}_t - g_{t+1}(\boldsymbol{x}_{t+1})\||\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] + 2\Delta\sum_{t=1}^{\hat{t}-1}\sum_{i=1}^{t} L_{1:i-1} \tag{90}$$

where (74) to express g_t as a function of $f_{\boldsymbol{\theta},t}$ plus an additive error term, the iterative application of the Lipschitz property of $f_{\boldsymbol{\theta},t}$, cf. Lemma 4.1, and the triangle inequality is used together with Assumption 4.3. Using the law of total expectation allows to rewrite the first term in (90) as

$$\sum_{t=1}^{\hat{t}-1} L_{1:t}\mathbb{E}_{t+1|\hat{t}}\left[\mathbb{E}_{t|t+1}\left[\|\boldsymbol{x}_t - g_{t+1}(\boldsymbol{x}_{t+1})\||\boldsymbol{x}_{t+1}\right]|\boldsymbol{x}_{\hat{t}}\right] \tag{91}$$

$$= \sum_{t=1}^{\hat{t}-1} L_{1:t} \mathbb{E}_{t+1|\hat{t}} \left[\mathbb{E}_{0|t+1} \left[\mathbb{E}_{t|t+1} \left[\|\boldsymbol{x}_t - g_{t+1}(\boldsymbol{x}_{t+1})\| | \boldsymbol{x}_{t+1}, \boldsymbol{x}_0 \right] | \boldsymbol{x}_{t+1} \right] | \boldsymbol{x}_{\hat{t}} \right]. \tag{92}$$

Utilizing the reparameterization

$$\boldsymbol{x}_t | \boldsymbol{x}_{t+1}, \boldsymbol{x}_0 = \tilde{\boldsymbol{\mu}}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) + \sigma_{t+1} \boldsymbol{\epsilon} \tag{93}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and using (28) and (74)

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) - g_{t+1}(\boldsymbol{x}_{t+1}) = \boldsymbol{\delta}_{t+1}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) - \mathbb{E}_{0|t+1}[\boldsymbol{\delta}_{t+1}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) | \boldsymbol{x}_{t+1}], \tag{94}$$

we get

$$\begin{aligned}
&\sum_{t=1}^{\hat{t}-1} L_{1:t} \mathbb{E}_{t+1|\hat{t}} \left[\mathbb{E}_{0|t+1} \left[\mathbb{E}_{t|t+1} \left[\|\boldsymbol{x}_t - g_{t+1}(\boldsymbol{x}_{t+1})\| | \boldsymbol{x}_{t+1}, \boldsymbol{x}_0 \right] | \boldsymbol{x}_{t+1} \right] | \boldsymbol{x}_{\hat{t}} \right] \\
&\leq \sum_{t=1}^{\hat{t}-1} L_{1:t} \mathbb{E}_{t+1|\hat{t}} \left[\mathbb{E}_{0|t+1} \left[\|\boldsymbol{\delta}_{t+1}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) - \mathbb{E}_{0|t+1}[\boldsymbol{\delta}_{t+1}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) | \boldsymbol{x}_{t+1}]\| | \boldsymbol{x}_{t+1} \right] | \boldsymbol{x}_{\hat{t}} \right] \\
&\quad + L_{1:t} \sigma_{t+1} \mathbb{E}_{\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon}\| \right] \\
&\leq 2\Delta \sum_{t=1}^{\hat{t}-1} L_{1:t} + N \sum_{t=1}^{\hat{t}-1} L_{1:t} \sigma_{t+1}
\end{aligned} \tag{95, 96}$$

where we used that $\mathbb{E}_{\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon}\|_p] < N$ holds for any $1 \leq p \leq \infty$. Similarly as in (83)–(87), we can bound the first term as

$$2\Delta \sum_{t=1}^{\hat{t}-1} L_{1:t} \leq 4\Delta L_1 (1 + \log \hat{t}) \tag{97}$$

and the second term as, cf. (63)–(70),

$$\begin{aligned}
N \sum_{t=1}^{\hat{t}-1} L_{1:t} \sigma_{t+1} &\leq 2L_1 N \sum_{t=1}^{\hat{t}-1} \frac{\sigma_{t+1}}{t} = 2L_1 N \sum_{t=1}^{\hat{t}-1} \frac{\sqrt{1-\alpha_t}}{t} \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} && (98) \\
&\leq 2L_1 N \sum_{t=1}^{\hat{t}-1} \frac{\sqrt{\beta_t}}{t} && (99)
\end{aligned}$$

where we used $1 - \bar{\alpha}_{t-1} \leq 1 - \bar{\alpha}_t$, following from the ordering $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_T$, cf. (23). Using $\beta_t = \mathcal{O}(T^{-\gamma})$, cf. (23), together with the bound on the harmonic number from (72), we conclude that

$$2L_1 N \sum_{t=1}^{\hat{t}-1} \frac{\sqrt{\beta_t}}{t} \leq 2NL_1 \mathcal{O}(T^{-\gamma/2})(1 + \log \hat{t}). \tag{100}$$

It is left to bound the second term in (90). Using the insights from (83)–(87), we get

$$\begin{aligned}
2\Delta \sum_{t=1}^{\hat{t}-1} \sum_{i=1}^{t} L_{1:i-1} &\leq 2\Delta \sum_{t=1}^{\hat{t}-1} (1 + 2L_1(1 + \log t)) && (101) \\
&\leq 2\Delta(1 + 2L_1)(\hat{t} - 1) + 4L_1 \Delta \log(\hat{t}!) && (102) \\
&\leq 2\Delta(1 + 2L_1)(\hat{t} - 1) + 4L_1 \Delta(\hat{t} - 1) \log \hat{t}. && (103)
\end{aligned}$$

Summarizing (87), (97), (100), and (103) yields (29) and considering only the worst-case asymptotic complexity yields (30), finishing the proof.

G Inference Steps for Constant Hyperparameters

Example: Consider $\beta_t = \beta = T^{-\xi}$ with $0 < \xi \leq 1$ for all $t = 1, \dots, T$. Then, solving (11)

$$\mathrm{SNR}_{\mathrm{DM}}(\hat{t}) = \frac{\bar{\alpha}_{\hat{t}}}{1-\bar{\alpha}_{\hat{t}}} = \frac{(1-\beta)^{\hat{t}}}{1-(1-\beta)^{\hat{t}}} \tag{104}$$

$$\Leftrightarrow \hat{t} = \frac{\log\left(\frac{\mathrm{SNR}_{\mathrm{DM}}(\hat{t})}{1+\mathrm{SNR}_{\mathrm{DM}}(\hat{t})}\right)}{\log(1-T^{-\xi})} \leq \underbrace{\log\left(\frac{\mathrm{SNR}_{\mathrm{DM}}(\hat{t})}{1+\mathrm{SNR}_{\mathrm{DM}}(\hat{t})}\right)}_{\leq 0}(1-T^{\xi}) = \mathcal{O}(T^{\xi}), \tag{105}$$

following from the logarithm inequality. Thus, if $\xi < 1$, the growth of the inference steps is only sub-linear in T for a fixed SNR.

H Convergence Analysis of the Diffusion Model with Re-sampling

Proposition H.1. *Let $p_{\boldsymbol{\theta}}(\boldsymbol{x}_0) = p(\boldsymbol{x})$ be a non-degenerate density function. Then, the DM with re-sampling in each step does not converge pointwise to the CME for any given observation $\boldsymbol{x}_{\hat{t}} \in \mathbb{R}^N$.*

Proof: We prove this statement by providing a counterexample to the converse statement. Thus, assume the re-sampling-based DM converges to the CME. Now, let $\boldsymbol{y} = \boldsymbol{n}$, i.e., we have a purely noisy and uninformative observation at an SNR of zero and thus $\hat{t} = T$. Then, the ground-truth CME yields $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_{\hat{t}}] = \mathbb{E}[\boldsymbol{x}_0]$, i.e., the CME is equal to the prior mean. Consequently, the output of the re-sampling-based DM estimator for every noise input is a deterministic point estimate by assumption. However, the re-sampling based DM is equal to the vanilla-DM, which is a generative model and thus generates samples from $p(\boldsymbol{x})$ when sampling from pure noise at timestep T. This raises a contradiction since $p(\boldsymbol{x})$ is non-degenerate. Thus, the converse statement is true, which proves the proposition.

I Network Architecture

We mainly use the DM architecture from (Ho et al., 2020) with minor adaptations. Specifically, we adopt the U-net architecture (Ronneberger et al., 2015) with wide ResNets (Zagoruyko and Komodakis, 2017). We use time-sharing of the NN parameters with a sinusoidal position embedding of the time information (Vaswani et al., 2017). The DM is trained on estimating the noise with the denoising matching term in (37) but ignoring the pre-factor, cf. Ho et al. (2020). We have $C_{\mathrm{init}} = 16$ initial convolutional channels, which are first upsampled with factors $(1, 2, 4, 8)$ and $(1, 2, 4)$ for the 1D and 2D data, respectively, and afterward downsampled likewise to the initial resolution with residual connections between the individual up-/downsampling streams. Each up-/downsampling step consists of two ResNet blocks with batch normalization, sigmoid linear unit (SiLU) activation functions, a linear layer for the time embedding input, and a fixed kernel size of (3×1) and (3×3) for the 1D and 2D data, respectively. As compared to Ho et al. (2020), we leave out attention blocks as this has not shown performance gains for the denoising task. We use a batch size of 128 and perform a random search over the learning rate. For different timesteps, we adapt the linear schedule of β_t such that approximately the same range of SNR values is covered by the DM, which has shown good results in all considered setups. The individual hyperparameters for the different timesteps and the resulting SNR range of the DM are given in Table 1.

Table 1: Hyperparameters of the DM for different timesteps T.

T	5	10	50	100	300	1,000
β_T	0.95	0.7	0.2	0.1	0.035	0.01
β_1	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
$\mathrm{SNR}_{\mathrm{DM}}(T)$	$-22.377\,\mathrm{dB}$	$-21.571\,\mathrm{dB}$	$-23.337\,\mathrm{dB}$	$-22.479\,\mathrm{dB}$	$-23.117\,\mathrm{dB}$	$-21.978\,\mathrm{dB}$
$\mathrm{SNR}_{\mathrm{DM}}(1)$	$40\,\mathrm{dB}$	$40\,\mathrm{dB}$	$40\,\mathrm{dB}$	$40\,\mathrm{dB}$	$40\,\mathrm{dB}$	$40\,\mathrm{dB}$

J Additional Experiments and Details

J.1 Random GMM

We choose the means $\boldsymbol{\mu}_k$ where every entry is i.i.d. $\mathcal{N}(0, 1/\sqrt{N})$ and random positive definite covariances following the procedure of Pedregosa et al. (2011), i.e., we first generate a matrix $\boldsymbol{S}$ where every entry is i.i.d. $\mathcal{U}(0, 1)$. Then, we compute the eigenvalue decomposition of $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$. Finally, we construct $\boldsymbol{C}_k = \boldsymbol{V}\,\mathrm{diag}(\mathbf{1} + \boldsymbol{\xi})\boldsymbol{V}^{\mathrm{T}}$ for all $k = 1, \ldots, K$, where every entry of $\boldsymbol{\xi}$ is i.i.d. $\mathcal{U}(0, 1)$. The weights $p(k)$ are chosen i.i.d. $\mathcal{U}(0, 1)$ and afterward normalized such that $\sum_{k=1}^{K} p(k) = 1$.

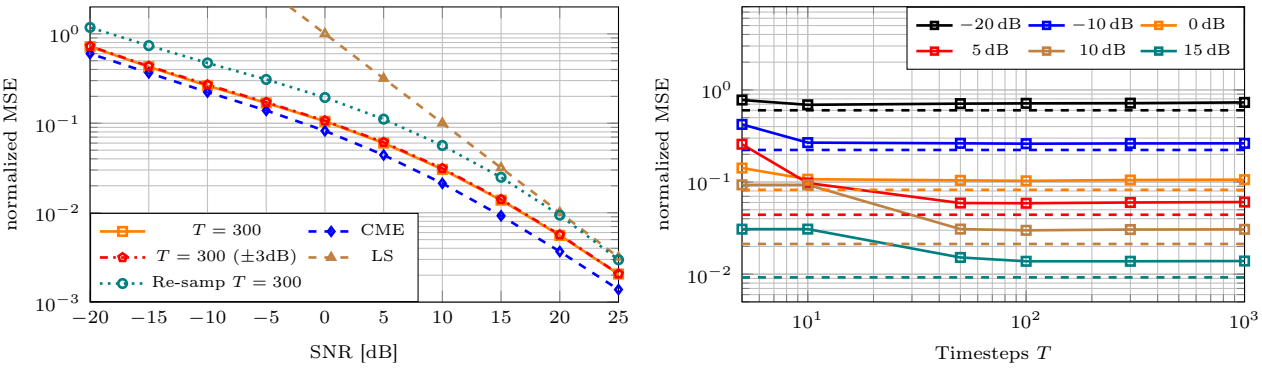

Figure 4: Evaluation of the **pre-trained** GMM based on **Fashion-MNIST** with $K = 128$ components.

J.2 Pre-trained GMM

Fashion-MNIST Data. In Figure 4 and 6 (left), the analysis is performed with a pre-trained GMM based on Fashion-MNIST instead of MNIST data. Essentially, the qualitative results are very similar to those of the pre-trained GMM based on MNIST but with a trend towards a larger number of T until saturation of the MSE and simultaneously a more significant decrease in the MSE over the reverse process. This may lead to the conclusion that for more complex data distributions, a higher number T is necessary until convergence, and the more significant the performance gap of a practically trained DM may become in comparison to the CME, which has complete knowledge of the prior distribution. Similar to earlier results, the impact of a mismatch in the SNR information (±3 dB) is negligible, highlighting the great robustness also for more complex data distributions.

Speech Data. As an additional dataset containing natural signals, we evaluate the Librispeech audio dataset (Panayotov et al., 2015), which contains 16 kHz read English speech. To limit the number of training samples for pre-training the GMM for this denoising example, we take the "test-clean" dataset and only keep samples no longer than two seconds in duration. We perform a similar pre-processing as Welker et al. (2022), i.e., we transform the raw waveform to the complex-valued one-sided short-time Fourier transform (STFT) domain by using a frame length of 512 samples, a hop size of 128 samples and a Hann window, after which we obtain $N = 256$-dimensional complex-valued data samples. As input to the DM, we stack the real and imaginary parts as separate convolution channels. The pre-trained GMM is fitted to the STFT data; also, the normalized MSE is calculated in the STFT domain, for which we add complex-valued circularly-symmetric Gaussian noise to obtain complex-valued observations at different SNR levels.

Figure 5 (left) shows the normalized MSE over the SNR of the observations for the proposed deterministic DM denoiser with $T = 300$ timesteps in comparison to the ground-truth CME, which has a large gap to the LS even in the high SNR regime. It can be observed that the DM estimator is almost on par with the CME over the whole range of SNRs (even for ±3 dB mismatch in the SNR information), validating the convergence analysis also for this dataset. In contrast to the image datasets, the DM is slightly worse in the low rather than the high SNR regime, illustrating the practical differences of the datasets from different domains. Moreover, the analysis verifies that the DM estimator on complex-valued data has the same estimation properties as analyzed for the real-valued case. In Figure 5 (right), we assess the number T of timesteps for convergence to the CME. Similar to the above results, a higher number T of timesteps is necessary for the high SNR regime until saturation; in contrast, for low SNR values, a small to moderate number of T is sufficient for a performance close to the ground-truth CME. This

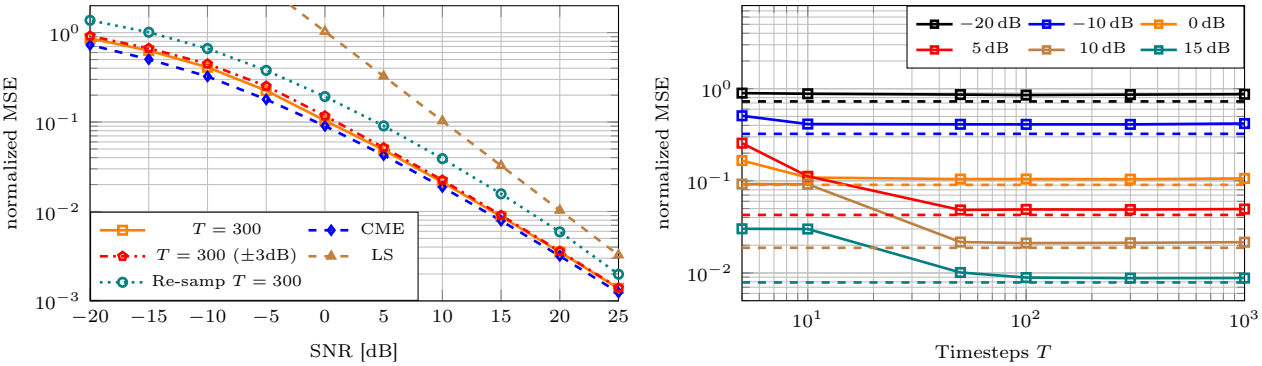

Figure 5: Evaluation of the **pre-trained** GMM based on **Librispeech** with $K = 128$ components.

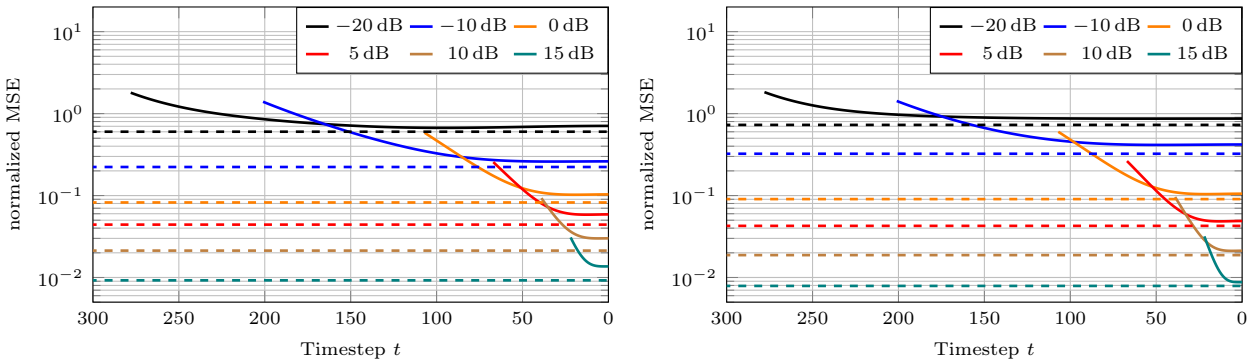

Figure 6: Comparison of the DM (solid) with the CME (dashed) after each timestep t with $T = 300$ for the **pre-trained** GMM based on **Fashion-MNIST** (left) and **Librispeech** (right) with $K = 128$.

further validates the convergence analysis on the speech dataset. Figure 6 (right) analyzes the MSE behavior over the DM's timesteps t in the reverse process. Similar to the insights from the image datasets, a steep decrease in the MSE can be observed for increasing t, ultimately converging to the CME, highlighting the structuredness of the data that can be utilized to enhance the estimation performance.

J.3 Runtime and Parameter Analysis

We conducted the runtime analysis on an Nvidia A40 GPU, where we measured the time for denoising a batch of 512 samples averaged over 10,000 test samples for the different datasets, cf. Table 2. As expected, the runtime mainly depends on the dimension of the respective samples and significantly reduces for higher SNR values since fewer reverse steps have to be performed. For example, the time for denoising at an SNR of 20 dB is only approximately 5% of the time for −10 dB SNR. The number of parameters for the image datasets is approximately one million, going up to over five million for the speech dataset since 1D instead of 2D convolutions are used, cf. Table 2.

Table 2: Runtime and parameter analysis of the proposed DM-based denoiser.

Dataset	#Parameters	Avg. time per batch (512 samples) in seconds			
		SNR = −10 dB	SNR = 0 dB	SNR = 10 dB	SNR = 20 dB
Random GMM (64×1)	$1.49 \cdot 10^6$	1.607	0.815	0.296	0.092
MNIST (28×28)	$1.06 \cdot 10^6$	4.532	2.383	0.828	0.211
Fashion-MNIST (28×28)	$1.06 \cdot 10^6$	4.531	2.382	0.828	0.211
Librispeech (256×1)	$5.62 \cdot 10^6$	4.070	2.149	0.759	0.207

J.4 Analysis of the Lipschitz Constants

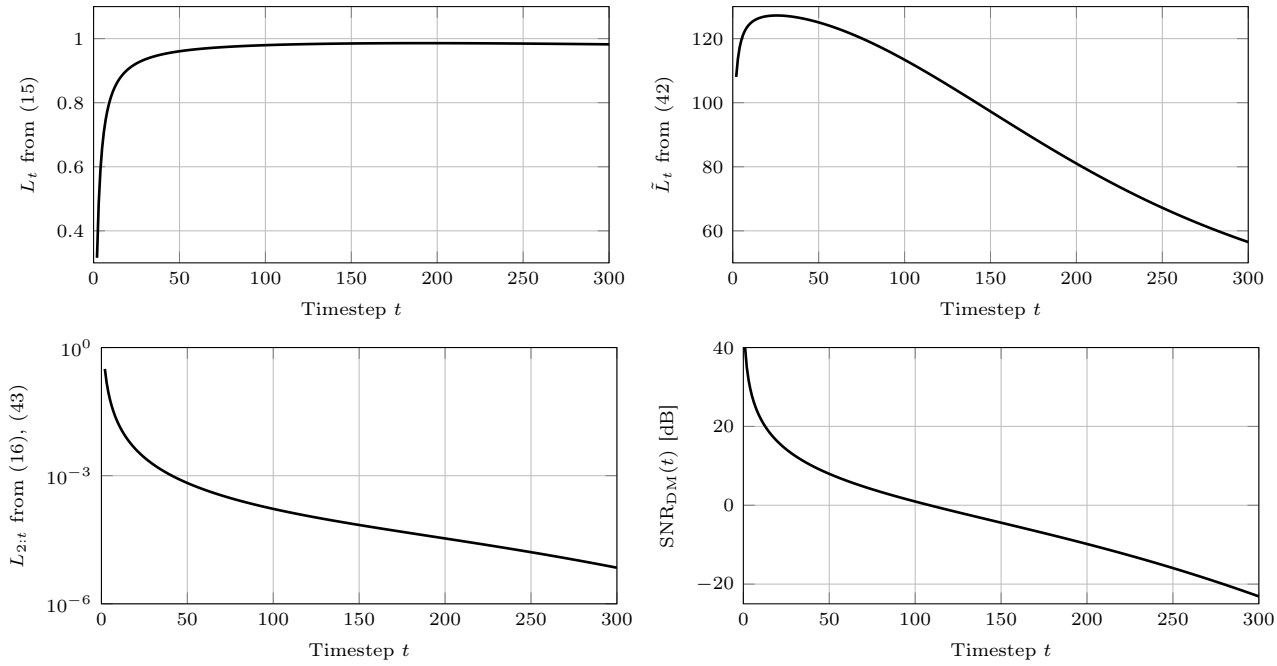

Figure 7: Lipschitz constants of $f_{\boldsymbol{\theta},t}$ (top left), $\boldsymbol{\epsilon}_{\boldsymbol{\theta},0}$ (top right), $f_{\boldsymbol{\theta},1:t}$ (bottom left), and the SNR representation of the DM (bottom right).

Figure 7 shows the analyzed Lipschitz constants from Lemma 4.1 and Appendix C and D for the linear schedule of β_t described in Appendix I. Figure 7 (top left) shows the stepwise Lipschitz constant of $f_{\boldsymbol{\theta},t} := \boldsymbol{\mu}_{\boldsymbol{\theta},t}$, being smaller than one for all $t = 2, \ldots, T$ as derived in Lemma 4.1. In contrast, the Lipschitz constant of the noise estimation NN $\boldsymbol{\epsilon}_{\boldsymbol{\theta},t}$, cf. Appendix C, in Figure 7 (top right) has a larger value for smaller t, being in accordance with the findings of Yang et al. (2024). Figure 7 (bottom) shows the relation between the Lipschitz constant of the proposed DM estimator $f_{\boldsymbol{\theta},1:\hat{t}}$ from (13) and the SNR for varying $\hat{t}$ as analyzed in Appendix D. Indeed, the Lipschitz constant is vanishing in the low SNR and monotonically increases for larger SNR values in agreement with the insights from Appendix D.

J.5 Qualitative Results

In this section, we provide qualitative analysis of the proposed DM-based denoiser for the MNIST and Fashion-MNIST datasets, illustrating the effect of the perception-distortion trade-off (Blau and Michaeli, 2018) and further validating our proposed denoising approach. We note that the noise realization is not uniform across all denoisers, leading to slightly different initial points for the denoising task.

J.5.1 Datasets Based on Pre-trained GMM

We first qualitatively analyze the denoising performance on datasets sampled from a pre-trained GMM, see Section 5 and Appendix J.2. For each dataset, we compare the performance of our proposed DM-based denoiser with the ground-truth CME, cf. (33), and the re-sampling-based approach for multiple randomly chosen samples from the test dataset across different SNR values, i.e., different noise levels in the observations. Please note that, since the "original" images are sampled from the pre-trained GMM rather than from the original MNIST and Fashion-MNIST datasets, they may contain certain artifacts arising from the mismatch between the GMM distribution and the true bounded density of images. However, the advantage of using these data lies in the ability to evaluate the ground-truth CME, thereby enabling a clearer visualization of the perception-distortion trade-off, which is a central aspect of this experiment. Later in Appendix J.5.2, we also evaluate the original datasets without the ground-truth CME baseline.

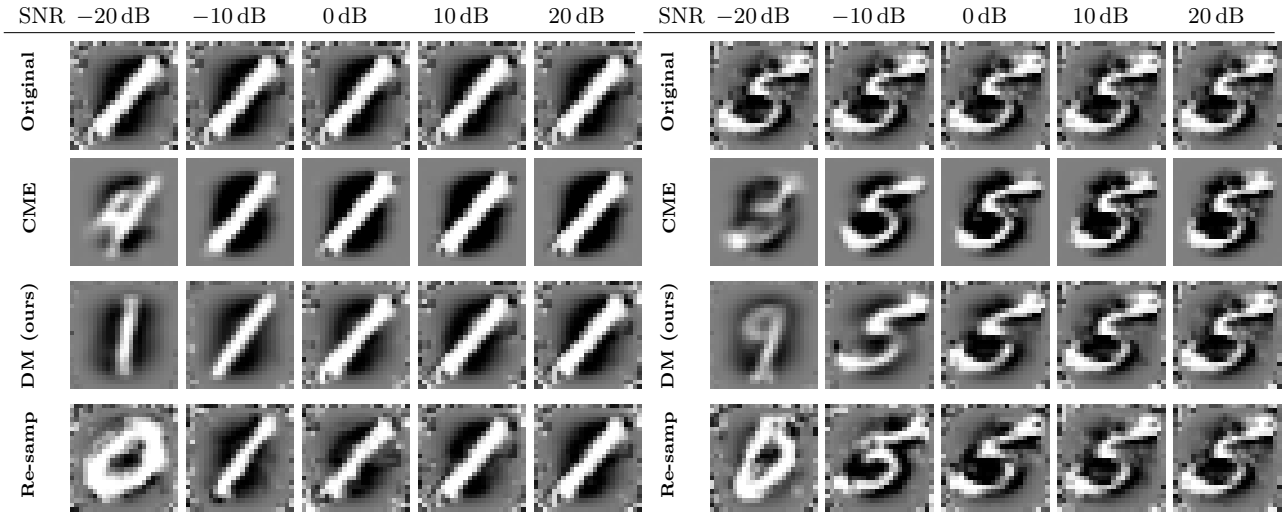

Figure 8: Qualitative denoising performance for different SNR levels for the the **pre-trained** GMM based on **MNIST** with $K = 128$.

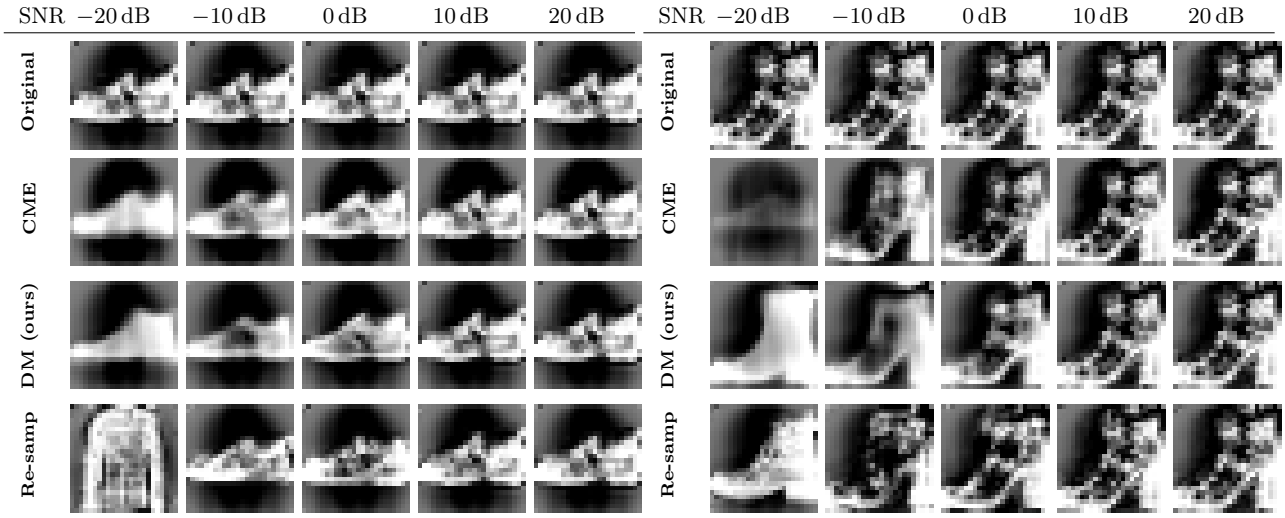

Figure 9: Qualitative denoising performance for different SNR levels for the the **pre-trained** GMM based on **Fashion-MNIST** with $K = 128$.

The qualitative analyses in Figure 8 and Figure 9 for the pre-trained GMM based on the MNIST and Fashion-MNIST datasets, respectively, show that the CME and our DM-based denoiser perform very similarly, despite differences in noise realizations for the observations. This further validates the theoretical analysis on a qualitative basis. Moreover, compared to DM-based denoising using re-sampling, the perception-distortion trade-off is clearly evident, particularly in the low SNR (high noise) regime. Specifically, the CME and our DM-based denoiser produce point estimates that appear more "blurry," resembling a weighted average over different images—consistent with the definition of the CME. In contrast, the re-sampling-based DM denoiser generates more perceptually appealing images, even in the low SNR (high noise) regime. However, this approach may yield completely different prior samples from the original data point, ultimately leading to a higher MSE (i.e., a higher distortion metric). We refer to the discussion in Section 4.5.

J.5.2 Original Datasets

In Figure 10 and Figure 11, the underlying DM is trained directly on the original MNIST and Fashion-MNIST datasets, respectively, without relying on a pre-trained GMM as the ground-truth distribution. Consequently, no

ground-truth CME baseline is available. The results further support the previous discussion on the perception-distortion trade-off and confirm the applicability of our proposed DM-based denoiser in the natural image domain. Similar to the findings in Appendix J.5.1, the difference between our deterministic DM-based denoiser and the re-sampling-based approach becomes more visually pronounced in the low SNR (high noise) regime. While our approach optimizes the distortion metric (i.e., minimizes the MSE), resulting in potentially more "blurry" but MSE-optimal reconstructions, the re-sampling-based method aims to generate high-quality perceptual features, which may not exist in the original sample. This ultimately leads to higher distortion (i.e., higher MSE).

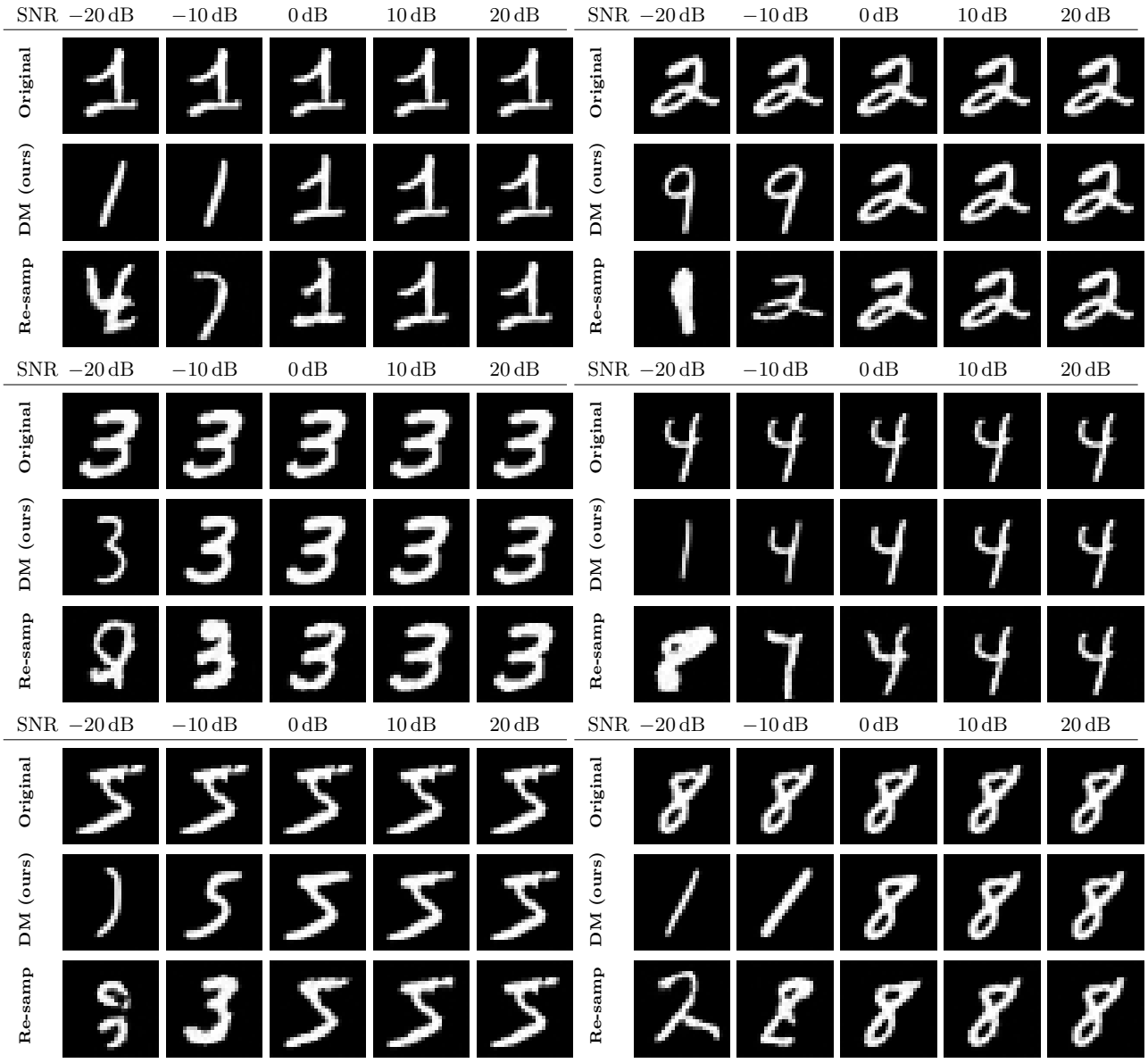

Figure 10: Qualitative denoising performance for different SNR levels based on the original **MNIST** dataset.

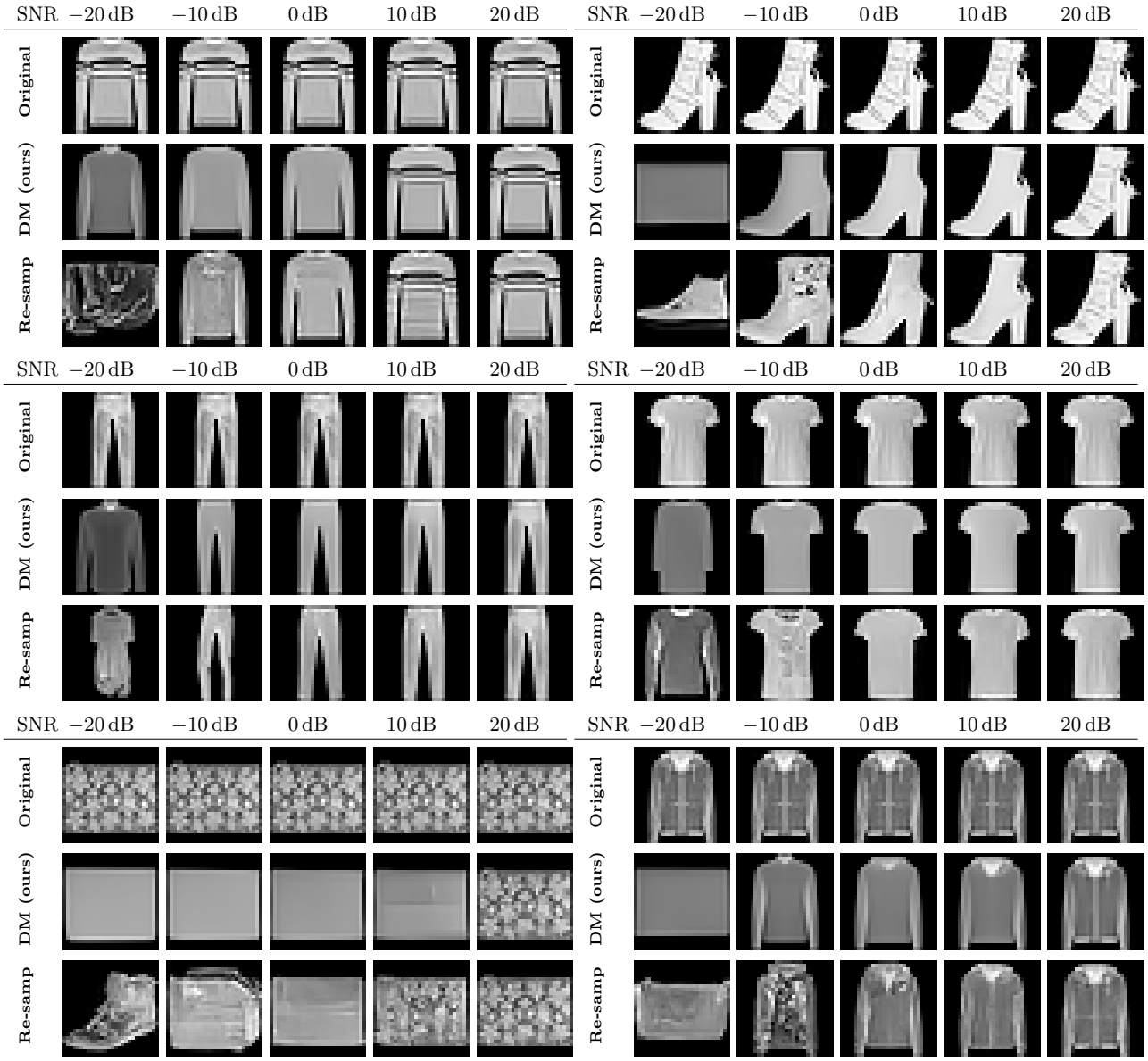

Figure 11: Qualitative denoising performance for different SNR levels based on the original **Fashion-MNIST** dataset.