Blind denoising diffusion models and adaptive sampling algorithms

Zahra Kadkhodaie*, Aram-Alexandre Pooladian*, Sinho Chewi, Eero P. Simoncelli
Flatiron institute, Simons Foundation, New York City, USA
Foundations of Data Science, Yale University, New Haven, USA
Center for computational neuroscience, New York University, New York City, USA
zkadkhodaie@flatironinstitute.org, aram-alexandre.pooladian@yale.edu

Abstract

Denoising diffusion models (DDMs) are state-of-the-art methods for learning densities from data across numerous domains, yet many aspects of the training and sampling pipeline remain poorly understood. In particular, noise conditioning requires practitioners to incorporate contrived unprincipled noise embeddings into neural network architectures and to use ad hoc noise schedules for sampling. To address these drawbacks, we provide a complete theory for *blind denoising diffusion models* (BDDMs): a variant of DDMs where the noise amplitude is not passed into the neural network during training or sampling, obviating the need for the aforementioned design choices. We justify the correctness of BDDMs as a sampling algorithm under an assumption of low intrinsic dimensionality of the underlying data distribution relative to the ambient dimension. This assumption arises through the introduction of the Bayesian problem of estimating noise levels from a single noisy sample, which might be of independent interest. We empirically compare the performance of BDDMs to standard DDMs, showcasing the benefits of an *adaptive* scheme which is rigorously justified by our analysis.

1. Introduction

Denoising diffusion models (DDMs) have emerged as the dominant methodology for generative modeling of images and solving inverse problems (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b; Chung et al., 2023). In contrast to previous machine learning methods (e.g., generative adversarial networks (Arjovsky et al., 2017), normalizing flows (Grathwohl et al., 2019), or variational autoencoders (Kingma & Welling, 2014)), diffusion models are based on corresponding noise corruption and denoising processes (Föllmer, 1985). In brief, the goal of DDMs is to sample from a data distribution p_X through discretizations of stochastic differential equations (SDEs) of the form

$$\mathrm{d}X_t = a_t \widehat{s}_\theta(X_t, \sigma_t)\,\mathrm{d}t + \sqrt{2a_t}\,\mathrm{d}B_t\,, \tag{1}$$

where $\widehat{s}_\theta : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$ is a parametric neural network trained on samples from p_X to approximate the score, $\nabla \log p_\sigma(x_\sigma)$, $\mathrm{d}B_t$ is standard Brownian motion, $(a_t)_{t\geq 0}$ are diffusion coefficients, and $(\sigma_t)_{t\geq 0}$ is a sequence of noise levels. The barriers to efficient and accurate sampling are then (a) obtaining a good approximation of the score $\widehat{s}_\theta$ and (b) choosing sequences $(\sigma_t)_{t\geq 0}$ and $(a_t)_{t\geq 0}$ that *predict* the noise level on the samples when discretizing Equation (1).

The first hurdle is resolved through an application of Tweedie's formula (Robbins, 1956; Miyasawa, 1961), which states that minimum mean square error (MMSE) denoiser (the mean of the posterior) can be re-written as the variance-scaled score function

$$r^\star_{\sigma_t}(y) \coloneqq \sigma_t^2 \nabla \log p_{\sigma_t}(y) = \mathbb{E}[X \,|\, Y = y] - y\,, \tag{2}$$

where the conditional expectation is over $X \sim p_X$ given $Y \sim \mathcal{N}(X, \sigma_t^2 I)$. With (2), the score can be learned by minimizing a regression-like objective (Hyvärinen, 2005; Raphan & Simoncelli, 2011). It is worth emphasizing that the parametric neural network takes both a noisy image y and its noise level σ_t as inputs, the latter implemented through the use of noise embeddings (Song et al., 2021b).

As for the second hurdle, the noise schedule and diffusion coefficients are often ad hoc and empirically tuned. For

*Equal contribution

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

instance, Karras et al. (2022) propose the schedule

$$\sigma_k = \left(\sigma_{\max}^{1/7} - \frac{k}{N}\left(\sigma_{\max}^{1/7} - \sigma_{\min}^{1/7}\right)\right)^7. \tag{3}$$

Perhaps surprisingly, we observe empirically that these noise schedules do not precisely predict the noise level on samples throughout the reverse process (see Figure 1). So, although the dynamics in Equation (1) are mathematically justified by the theory for forward-reverse diffusions, these justifications break down at the implementation stage. This breakdown is due to an inconsistency in the discretized reverse process: the score model $\widehat{s}_\theta$ takes in two arguments (x_σ, σ) in which the second argument σ obtained from the schedule does not accurately predict the σ of the noisy sample x_σ accurately (see Figure 1).

One way to resolve this inconsistency is to entirely remove the second argument (i.e., the noise level) and allow the model to infer it from the noisy sample. Such a construction is called a "blind" model in the denoising signal processing literature (Liu et al., 2013; Zhang et al., 2017; Mohan et al., 2020). Despite the absence of an explicit noise level, these models have been shown to achieve denoising performance rivaling that of the non-blind models.

In this work, we study denoising diffusion models that employ blind denoisers, which we call *blind denoising diffusion models* (BDDMs). BDDMs are trained *without* conditioning on the noise level of the image and also require no explicit noise schedule $(\sigma_k)_{k\geq 0}$ during inference, nor an external diffusion rate $(a_k)_{k\geq 0}$. Such a sampling scheme was initially proposed by Kadkhodaie & Simoncelli (2020): letting $\widehat{f}_\theta$ denote a trained blind denoiser (see (5)), their proposed scheme essentially amounts to

$$Y_{(k+1)h} = Y_{kh} + h\left(\widehat{f}_\theta(Y_{kh}) - Y_{kh}\right) + \sqrt{2h\beta\widehat{\sigma}_k^2}\xi_k \tag{4}$$

where $\widehat{\sigma}_k^2 := \|\widehat{f}_\theta(Y_{kh}) - Y_{kh}\|^2/d$ is an estimate of noise variance computed at each iteration, $h = \Delta t$ is the discretization constant, and $\beta \in (0, 1]$ is a constant diffusion coefficient. The rightmost plot in Figure 1 demonstrates that unlike traditional denoising schedules (e.g., log-uniform or (3)), this implicit choice exactly tracks the noise level of the image, resulting in minimal mismatch along the denoising process.

Contributions

Our main contribution is to provide the first end-to-end theoretical justification for BDDMs. In particular, we prove that sampling schemes like (4) sample from the true data distribution p_X (see Theorem 3.6). Our main assumption is that p_X has *low intrinsic dimensionality* with respect to the ambient space (see Definition 3.5).

Along the way, we develop a number of useful insights. Namely, we derive an analytical formula for the evolution of noise level along the reverse trajectory—see (12)—and provide statistical theory justifying why it can be estimated along the reverse diffusion trajectory. Moreover, while our theory supports *any* diffusion coefficient (see (6)), our discretization analysis in Section 3.5 suggests that schemes of the form (4) are possibly optimal—see Theorem 3.9 and Corollary 3.10.

Some recent works have investigated blind denoising for DDMs and related flow-based models, however their conclusions remain far from comprehensive or rigorous (Sun et al., 2025; Wang & Du, 2025).

NOTATION

We denote the centered Gaussian with variance $\sigma^2 > 0$ as $\gamma_{\sigma^2} = \mathcal{N}(0, \sigma^2 I)$. Throughout, we denote the data distribution as p_X, and the noisy data distribution(s) as $p_\sigma = p_X * \gamma_{\sigma^2}$. We let $\mathrm{d}B_t$ denote standard Brownian motion. The Kullback–Leibler divergence between probability densities p, q is written $\mathrm{KL}(p\|q) = \int \log(p(x)/q(x))\,\mathrm{d}p(x)$.

2. Preliminaries on blind denoisers

2.1. Training

Unlike standard denoisers in the diffusion model literature, blind denoisers are not given the noise level during training. For a neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$, a *blind denoiser* is trained using the following objective (Zhang et al., 2017; Gnanasambandam & Chan, 2020;

Algorithm 1. Training a blind denoiser

Input: Distributions p_X, Π_0, neural network f_θ
while `not converged` **do**
  Draw $x_1, \ldots, x_B \sim p_X$
  Draw $\sigma_1, \ldots, \sigma_B \sim \Pi_0$
  Draw $z_1, \ldots, z_B \sim \mathcal{N}(0, I)$
  Compute $\mathcal{L}(\theta) = B^{-1}\sum_{i=1}^B \|x_i - f_\theta(x_i + \sigma_i z_i)\|^2$
  Update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}(\theta)$.
end while

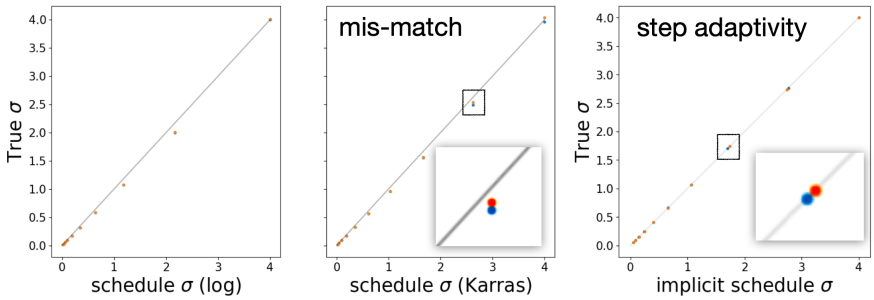

Figure 1. Empirical comparison of scheduled noise level σ_t and estimated noise level $\sigma^\star$ along trajectories for the CelebA dataset. The DDPM reverse process outpaces a log schedule **(left)**, or (3) proposed by Karras et al. (2022) **(middle)**, leading to an excess error manifested in the loss of detail in samples. In contrast, our BDDM tracks the implicit schedule σ_t via $\hat{\sigma}_k = \|x_k - f(x_k)\|/\sqrt{d}$ **(right)**.

Mohan et al., 2020):

$$\widehat{f}_\theta = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\substack{\sigma\sim\Pi_0\\ z\sim\gamma}} \|\boldsymbol{x}_i - f_\theta(\boldsymbol{x}_i + \sigma z)\|^2 \tag{5}$$

where $\{\boldsymbol{x}_i\}_{i=1}^n \sim p_X$ are the clean training samples, γ is the standard Gaussian in $\mathbb{R}^d$, and Π_0 is a prior distribution over noise levels in the range $[\sigma_T, \sigma_0]$, with $0 < \sigma_T < \sigma_0 < +\infty$ (since we are interested in the reverse process, σ_0 is the largest noise level, and σ_T is the smallest). In practice, we minimize (5) using (batch) stochastic gradient descent, see Algorithm 1 for pseudocode.

2.2. Sampling

To sample, we initialize $Y_0 \sim \mathcal{N}(0, \sigma_0^2 I)$ and consider various discretizations of

$$\mathrm{d}Y_t = (\widehat{f}_\theta(Y_t) - Y_t)\,\mathrm{d}t + \sqrt{2a_t}\,\mathrm{d}B_t\,; \tag{6}$$

This is the general continuous analogue of Equation (4) with an arbitrary sequence $(a_t)_{t\geq 0}$. For example, a simple Euler discretization gives, for a constant step size $h > 0$ and $\xi_k \sim \mathcal{N}(0, I)$

$$Y_{(k+1)h} = Y_{kh} + h\left(\widehat{f}_\theta(Y_{kh}) - Y_{kh}\right) + \sqrt{2a_k h}\,\xi_k\,.$$

See Algorithm 2 for the inference algorithm. We additionally discuss the exponential (Euler) integrator in Appendix C.1, as it plays a role in our analysis.

Algorithm 2. Sampling via BDDMs

Input: Trained network $\widehat{f}_\theta$, stepsize $h > 0$
Diffusion $(a_t)_{t\in[0,T]}$, $\sigma_{\max}, \sigma_{\min} > 0$
Initialize $X_0 \sim \mathcal{N}(\widehat{m}_X, \sigma_{\max}^2 I)$, $k = 0$
`keepgoing` ← `True`
while `keepgoing` == `True` **do**
  Compute $r_k \leftarrow \widehat{f}_\theta(X_{kh}) - X_{kh}$
  Compute $\hat{\sigma}_k^2 \leftarrow \|r_k\|^2/d$
  if $\hat{\sigma}_k \leq \sigma_{\min}$ **then**
    `keepgoing` ← `False`
  else
    Draw $\xi \sim \mathcal{N}(0, (\int_{kh}^{(k+1)h} 2a_t\,\mathrm{d}t)I)$
    Update $X_{(k+1)h} \leftarrow X_{kh} + hr_k + \xi$
  end if
  Update $k \leftarrow k + 1$
end while

2.3. Related work

BDDMs were introduced by Kadkhodaie & Simoncelli (2020; 2021), where a blind deep neural network denoiser (Zhang et al., 2017) was used to sample from p_X. The capabilities of BDDMs for sampling and solving linear inverse problems were demonstrated empirically and their generalization (with respect to training set size) was examined using U-Net architectures (Ronneberger et al., 2015) in Kadkhodaie et al. (2024).

Closest to our work is that of Sun et al. (2025), where the authors show that noise conditioning is not required for sampling. They verify this empirically on dynamics which require explicit schedules (e.g., DDPM, DDIM, EDM), and thus do not fully investigate adaptive algorithms as we do. On the theoretical front, they make initial progress by studying the simple cases where p_X is a single Dirac mass or mixture of Dirac masses. We depart from these simple scenarios by showing that blind denoisers succeed due to low intrinsic dimensionality of the underlying data distribution. We provide rigorous derivations under this assumption, and provide numerous other technical and empirical contributions of interest to diverse communities.

In other work (e.g., Du & Mordatch, 2019; Wang & Du, 2025) neural networks have been trained to learn dynamics without time inputs. We believe our work provides a basis for understanding how these training dynamics are possible, accompanied bytheoretical guarantees.

3. Theoretical contributions

This section contains our theoretical contributions, which rigorously justify the use of BDDMs as generative models. All proofs are provided in the appendix.

3.1. The optimal blind denoiser

The first step is to understand the population minimizer of the blind denoising problem of Equation (5): Given infinite data and perfect optimization, what is learned? Since the noise level is not known, it should be treated as a random variable (as opposed to a known parameter). From a Bayesian perspective, the optimal solution then comes from marginalizing over this random variable. The following theorem makes this precise and indicates that the optimal blind denoiser can be expressed as a *conditional average* of score functions.

Proposition 3.1. *The population minimizer of* (5) *is*

$$f^\star : y \mapsto y + \int \sigma^2 \nabla \log p_\sigma(y) \, \mathrm{d}\Pi(\sigma|y) \,,$$

where $p_\sigma \coloneqq p_X * \mathcal{N}(0, \sigma^2 I)$*, and*

$$\Pi(\sigma|y) \propto \frac{\Pi_0(\sigma)}{\sigma^d} \, \mathbb{E}_{X\sim p_X} \exp\Big(-\frac{1}{2\sigma^2} \, \|X - y\|^2\Big) \,. \tag{7}$$

This allows us to express the optimal blind variance-scaled score function as an integral over the noise posterior Π, i.e.,

$$r^\star(y) \coloneqq f^\star(y) - y = \int \sigma^2 \nabla \log p_\sigma(y) \, \mathrm{d}\Pi(\sigma|y) \,. \tag{8}$$

3.2. Derivation of the implicit noise schedule

We now examine the dynamics arising from the optimal blind score function. To start, we consider the general continuous-time dynamics of Equation (6) with a perfectly trained denoiser, and an arbitrary sequence $(a_t)_{t\geq 0}$, which are given by

$$\mathrm{d}X_t^\star = r^\star(X_t^\star) \, \mathrm{d}t + \sqrt{2a_t} \, \mathrm{d}B_t \tag{9}$$

where we initialize with $X_0^\star \sim \mathcal{N}(0, \sigma_0^2 I)$.

We consider the following ansatz: for all t, $X_t^\star$ is approximately distributed as p_{σ_t}, for some noise sequence $(\sigma_t)_{t\geq 0}$ (to be derived). Further, let us suppose that when $y = X_t^\star$, then the conditional distribution $\Pi(\sigma|y)$ in (7) concentrates on the true noise level σ_t. As a result, (8) collapses to a Laplace approximation. This suggests considering the following process in which we replace the conditional average over noise levels with σ_t:

$$\mathrm{d}X_t = r_{\sigma_t}^\star(X_t) \, \mathrm{d}t + \sqrt{2a_t} \, \mathrm{d}B_t \coloneqq \sigma_t^2 \nabla \log p_{\sigma_t}(X_t) \, \mathrm{d}t + \sqrt{2a_t} \, \mathrm{d}B_t \,. \tag{10}$$

However, along (10), the evolution of the density is given by SDE theory, in particular by the Fokker–Planck equation. In order to be consistent with the ansatz $\mathsf{Law}(X_t) \approx p_{\sigma_t}$, this implies the following ODE for σ_t (see Appendix A.2):

$$\frac{1}{2} \, \partial_t(\sigma_t^2) = -\sigma_t^2 + a_t \,. \tag{11}$$

Solving this ODE, we arrive at the following result.

Proposition 3.2. *The ideal SDE process* (10) *satisfies* $\mathsf{Law}(X_t) = p_{\sigma_t}$ *for all* $t \geq 0$*, provided that* $X_0 \sim p_{\sigma_0}$ *and*

$$\sigma_t^2 = \sigma_0^2 e^{-2t} + 2 \int_0^t a_s e^{-2(t-s)} \, \mathrm{d}s \,. \tag{12}$$

We record two properties of this emergent noise evolution.

Lemma 3.3. *If* $t \mapsto a_t$ *is decreasing and* $a_0 \leq \sigma_0^2$*, then* $t \mapsto \sigma_t$ *is decreasing.*

Lemma 3.4. *If* $t \mapsto a_t$ *is decreasing and* $a_t \to 0$ *as* $t \to \infty$*, then* $\sigma_t \to 0$ *as well.*

Henceforth, we assume that the conditions of Lemma 3.4 hold. In brief, this discussion suggests that if $\Pi(\cdot|X_t)$ concentrates around σ_t such that

$$\int \sigma^2 \nabla \log p_\sigma(y) \, \mathrm{d}\Pi(\sigma|X_t) \approx \sigma_t^2 \nabla \log p_{\sigma_t}(X_t) \,, \tag{13}$$

and if $a_t \to 0$, then heuristically we expect $\mathsf{Law}(X_t^\star) \approx \mathsf{Law}(X_t) \to p_X$ as $t \to \infty$.

We stress that at no point during the sampling process will the blind denoiser be given information about the schedule σ_t. Instead, it must adapt to this sequence automatically by estimating σ_t from the current input Y_t. In other words, if $\Pi(\cdot|X_t)$ concentrates, the true σ_t for which $p_{\sigma_t} = p_X * \mathcal{N}(0, \sigma_t^2 I)$ can be inferred from a single input Y_t. This is because $p_\sigma(Y_t)$ would be negligible for all σ but a small set around σ_t. As a result, under the concentration assumption, the ideal dynamics in the reverse process should follow the true noise sequence σ_t.

EMPIRICAL VERIFICATION OF THE CONCENTRATION ASSUMPTION

Since our proof relies on concentration of $\Pi(\cdot|y)$, we verify this main assumption empirically on synthetic data, before stating the rest of theoretical results. Results on real data are postponed to Section 4.

We study an analytical blind denoiser model when p_X is a 2-component mixture of Gaussians supported on a k-dimensional subspace in $\mathbb{R}^d$. Here, the optimal score function $\nabla \log p_\sigma$ is known in closed form, allowing us to isolate the error induced by the uncertainty in noise estimation. Our analytical blind denoiser will be given by a maximum likelihood estimate (MLE) of $\Pi(\cdot|y)$ for a noisy sample y:

$$\widehat{f}(y) \coloneqq y - \widehat{\sigma}^2 \nabla \log p_{\widehat{\sigma}}(y)\,, \quad \widehat{\sigma} = \arg\max \Pi(\cdot|y)$$

Figure 2 interleaves the distribution of estimated noise values and the samples generated via Algorithm 2 for a mixture of two Gaussians of dimensionality k in a d-dimensional ambient space. Estimates are broadly distributed when $\mathsf{k} \approx d$, but concentrate for $\mathsf{k} \ll d$, illustrating a "blessing of dimensionality"; the sampling results show the corresponding failure and success modes of our analytical denoiser. In these experiments, $h = 0.3$ and the dynamics are deterministic ($a_t = 0$). Similar results are obtained when $a_t > 0$. For more details, see Appendix G.1.

3.3. Error bound without discretization

In this section, we study the conditions required for our concentration assumption to hold, and then quantify the excess error arising from deviations from the perfect concentration.

Suppose that we have an empirical minimizer $\widehat{f}_\theta$ that will act as the blind denoiser, and let $\widehat{p}_t \coloneqq \mathsf{Law}(Y_t)$ along the trained dynamics (6). For now, we omit discretization error, deferring this discussion to Section 3.5. Our goal is to bound the KL divergence between $\widehat{p}_T$, the distribution corresponding to our algorithm at time T, and the target distribution with some small additional noise $p_{\sigma_T} = p_X * \mathcal{N}(0, \sigma_T^2 I)$.

By a standard application of Girsanov's theorem, we can decompose this error into the following terms:

$$\mathrm{KL}(p_{\sigma_T} \| \widehat{p}_T) \lesssim \mathrm{KL}(p_{\sigma_0} \| \widehat{p}_0) + \int_0^T a_t^{-1} \|\widehat{f}_\theta - f^\star\|^2_{L^2(p_{\sigma_t})}\,\mathrm{d}t + \int_0^T a_t^{-1} \|r^\star_{\sigma_t} - r^\star\|^2_{L^2(p_{\sigma_t})}\,\mathrm{d}t\,. \tag{14}$$

The first term on the right-hand side corresponds to the initialization error. The second term measures how close the trained denoiser is to the optimal one (the "score error"). Finally, the third term is related to the error in the approximation (13), i.e., the error in estimating the noise level. This last term is the primary novelty over prior work on diffusion models, so we focus our attention on it.

Before stating our main results, we require the following assumptions and definitions. The support of our data distribution will be denoted by $\mathcal{X} \coloneqq \mathrm{supp}(p_X)$, and we let k denote the "intrinsic dimension" of p_X, defined as follows.

Definition 3.5 (Intrinsic dimension)**.** If p_X has support $\mathcal{X}$, we define the *intrinsic dimension* of p_X at scale r_0 to be $\mathsf{k} \coloneqq 1 + \log N_{\mathcal{X}}(r_0)$, where $N_{\mathcal{X}}(r_0)$ denotes the minimal number of balls of radius r_0 needed to cover $\mathcal{X}$.

For example, if $\mathcal{X}$ is a k-dimensional subspace and $\mathcal{X}$ has diameter R, then $\mathsf{k} \asymp k \log(R/r_0)$. But the intrinsic dimension captures a broader range of situations such as manifold structure, or when $\mathcal{X}$ is actually full-dimensional but resembles a k-dimensional manifold at scale r_0. Our proofs require $r_0 = \sigma_T^2/(\sigma_0\sqrt{d})$ as the choice of scale.

Our assumptions are the following.

(A1) p_X has finite second moment, denoted m_2^2.
(A2) p_X has low intrinsic dimensionality, with $d \geq C(\mathsf{k} + \log(\sigma_0/\sigma_T))$ for a sufficiently large absolute constant $C > 0$.
(A3) For $\varepsilon_{\mathrm{BD}} > 0$, $\int_0^T a_t^{-1} \|\widehat{f}_\theta - f^\star\|^2_{L^2(p_{\sigma_t})}\,\mathrm{d}t \leq \varepsilon_{\mathrm{BD}}^2$.

Note that **(A1)**, **(A2)**, and **(A3)** are relatively mild. In particular, **(A2)** is inspired from a slew of previous empirical observations (Olshausen & Field, 1996; Roweis & Saul, 2000; Chandler & Field, 2007; Hénaff et al., 2014; Pope et al.,

2021; Brown et al., 2023) and has been leveraged in several other works on sampling via diffusion models (Li & Yan, 2024; Pooladian & Niles-Weed, 2025; Liang et al., 2025). **(A3)** is analogous to the standard assumption that the score functions are accurately learned in L^2. In its current form, it is somewhat obscure; we provide further discussion in Appendix E when we specialize to specific noise schedules.

We are now in a position to state our first main result about excess error due to noise estimation.

Theorem 3.6. *Under **(A1)–(A3)**, the KL divergence between the (reverse) processes (6) and (10) is bounded by*

$$\mathrm{KL}(p_{\sigma_T}\|\widehat{p}_T) \lesssim \frac{\mathsf{m}_2^2}{\sigma_0^2} + \varepsilon_{\mathrm{BD}}^2 + \big(\frac{\mathsf{k}^3}{d} + \frac{\mathsf{k}^5}{d^2}\big)\int_0^T \frac{\sigma_t^2}{a_t}\,\mathrm{d}t\,.$$

The first term (initialization error) is made small with a suitably large choice of σ_0^2, and the second term (the "score error") is small if the training is successful. The third term, which scales with k, is discussed next.

3.4. Interpretation as a Bayesian problem

We now provide an overview of our proof with rigorous details deferred to the appendix. As the first two terms in Theorem 3.6 are standard, we focus on the novel term $\int_0^T a_t^{-1}\|r^\star_{\sigma_t} - r^\star\|^2_{L^2(p_{\sigma_t})}\,\mathrm{d}t$ which captures the error in estimating the noise level.

Note that for $X_t \sim p_{\sigma_t}$ we can re-write the integrand with respect to deviations of noise posterior, $\Pi(\sigma|X_t)$, from the perfect posterior concentrated on a Dirac mass δ_{σ_t}

$$\|(r^\star_{\sigma_t} - r^\star)(X_t)\|^2 = \Big\|\int \sigma^2\,\nabla\log p_\sigma(X_t)\,\mathrm{d}(\delta_{\sigma_t} - \Pi(\sigma|X_t))\Big\|^2 = \Big\|\iint_\sigma^{\sigma_t} \partial_\omega(\omega^2\,\nabla\log p_\omega(X_t))\,\mathrm{d}\omega\,\mathrm{d}\Pi(\sigma|X_t)\Big\|^2\,,$$

where we used the fundamental theorem of calculus in the last line. A simple calculation shows that

$$\partial_\omega(\omega^2\,\nabla\log p_\omega(y)) = \omega^{-3}\,\mathrm{Cov}(X, \|X-Y\|^2 \mid Y = y)\,,$$

under the distribution of $X \sim p_X$, $Y \sim \mathcal{N}(X, \omega^2 I)$. We show in Corollary F.3 that the conditional covariance is bounded in terms of the intrinsic dimension k, essentially leading to a bound on the above term of order

$$\sigma_t^3\,\mathsf{k}^3 \int |\sigma_t^{-2} - \sigma^{-2}|^2\,\mathrm{d}\Pi(\sigma|X_t)\,.$$

The remainder of the analysis is a frequentist analysis of a Bayesian method: We have an observation $X_t \sim p_{\sigma_t}$, where σ_t is the "ground truth" noise level. We ask whether the posterior distribution on the noise level $\Pi(\cdot|X_t)$ concentrates on σ_t given a single sample. For interpretability, we focus on the class of power law priors $\Pi_0(\sigma) \propto \sigma^{\alpha-3}$ on $[\sigma_T, \sigma_0]$ for $\alpha \in \mathbb{R}$, $\alpha = O(1)$, and we use the following change of variables.

Lemma 3.7. *If $\sigma \sim \Pi(\cdot|y)$ in (7), then $\lambda := \sigma^{-2}$ is distributed according to*

$$\overline{\Pi}(\lambda|y) \propto \lambda^{(d-\alpha)/2}\,\mathbb{E}_{X\sim p_X}\big[\exp\big(-\tfrac{\lambda}{2}\,\|X-y\|^2\big)\big]\,. \tag{15}$$

Writing $\ell(\lambda|y) := -\log\overline{\Pi}(\lambda|y)$, then

$$\ell'(\lambda|y) = -\frac{d-\alpha}{2\lambda} + \frac{1}{2}\,\mathbb{E}_{q_\lambda}[\|X-Y\|^2 \mid Y = y]\,, \qquad \ell''(\lambda|y) = \frac{d-\alpha}{2\lambda^2} - \frac{1}{4}\,\mathrm{Var}_{q_\lambda}(\|X-Y\|^2 \mid Y = y)\,,$$

where, in an abuse of notation, we use q_λ to denote the joint distribution for which $X \sim p_X$, $Y \sim \mathcal{N}(X, \lambda^{-1}I)$.

Notice that there is no reason for $\ell''(\cdot|y)$ to always be convex, given the presence of a negative sign on the second term. However, our *low intrinsic dimensionality assumption* **(A2)** nevertheless allows us to show that the noise variance posterior concentrates on the ground truth signal $\lambda_t := \sigma_t^{-2}$.

Proposition 3.8. *Under **(A1)–(A2)**, with high probability over $X_t \sim p_{\sigma_t}$, and for $\lambda \sim \overline{\Pi}(\cdot|X_t)$,*

$$\mathbb{E}|\lambda - \lambda_t|^2 \lesssim \lambda_t^2\,(d^{-1} + \mathsf{k}^2 d^{-2})\,.$$

See (24) for a precise result. Combining these ingredients yields the claimed bound on the noise level term.

3.5. Discretization, noise, and step size schedules

A family of convenient diffusion schedules. Although our results apply to general choices of $(a_t)_{t\geq 0}$, for the sake of exposition we specialize to a particular convenient family in order to state our next results in a more interpretable form. Namely, if we take $a_t = \mathsf{a}\sigma_t^2$ for some $\mathsf{a} \in (0,1)$, then the ODE (11) becomes particularly easy to solve: $\sigma_t = \sigma_0 e^{-(1-\mathsf{a})t}$. In fact, this is the choice made in Equation (4) in which $\mathsf{a} = \beta$. We henceforth fix this particular choice.

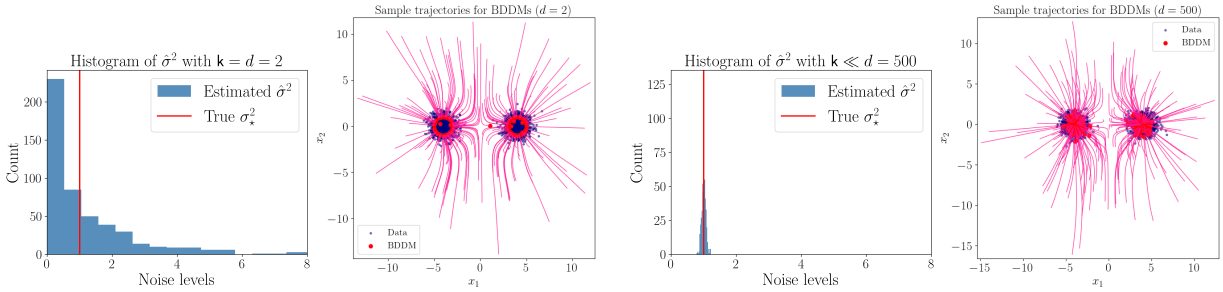

Figure 2. **Illustration of Propositions 3.8 and Corollary 3.10:** Maximum likelihood estimates of noise level and samples for an analytical blind denoiser applied to a mixture of two Gaussians with intrinsic dimensionality $\mathsf{k} = 2$. **(A)** Empirical density of $\widehat{\sigma} = \arg\max \Pi(\sigma|y)$ for $d = \mathsf{k} = 2$, where estimates are broadly distributed. **(B)** Sampling fails in the same low-dimensional setting due to errors in the MLE estimates of noise level. **(C)** The estimates concentrate when $d = 500 \gg \mathsf{k}^2$. **(D)** Sampling succeeds in the same high-dimensional setting, illustrating the blessing of dimensionality.

Discretization error. If the SDE (6) is discretized, then in addition to the error terms present in Theorem 3.6, we will also incur a discretization error. In order to establish an iteration bound which also scales with the intrinsic dimension, we consider the exponential Euler integrator (see Appendix C.1). With this choice, the discretization error takes the following form: Writing $t_- := \lfloor t/h \rfloor h$ where h is the step size, thus $N = T/h$ is the total number of iterations,

$$\mathsf{Disc}(h) = \int_0^T a_t^{-1}\, \mathbb{E}\|f^\star_{\sigma_{t_-}}(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2 \,\mathrm{d}t\,.$$

Here, $f^\star_\sigma$ denotes the Bayes denoiser when the noise level is known: $f^\star_\sigma(y) := \mathbb{E}_{q_\sigma}[X \mid Y = y]$.[1] We analyze the discretization error and establish the following bound.

Theorem 3.9. *Under (A1)–(A3) and with the exponential Euler scheme, for $h \lesssim 1$,*

$$\mathsf{Disc}(h) \lesssim \left(C_{\mathsf{a},1}\, \mathsf{k}^3 h^2 + C_{\mathsf{a},2}\, \mathsf{k} h\right) \log \frac{\sigma_0}{\sigma_T}\,,$$

where $C_{\mathsf{a},1} := \frac{(\mathsf{a}-1/2)^2}{\mathsf{a}\,(1-\mathsf{a})}$ and $C_{\mathsf{a},2} := \frac{1}{1-\mathsf{a}}$.

Remarkably, the first error term vanishes *exactly* when $\mathsf{a} = 1/2$. Therefore, our analysis reveals a canonical choice of a_t sequence, namely $a_t = \frac{1}{2}\sigma_0^2 e^{-t}$, which leads to smaller discretization error.

Moreover, if $\mathsf{a} = 1/2$, the error bound shows that we can provably choose a constant step size h, scaling with the intrinsic dimension k. This is in stark contrast to theoretical guarantees for the variance-preserving DDPM algorithm, which requires carefully tuned exponentially decaying step sizes to compensate for the singularity of the score for time $t \to T$ (Benton et al., 2024; Conforti et al., 2025).

3.6. End-to-end sampling guarantees

Combining the noise-estimation and discretization bounds with early stopping yields an end-to-end guarantee for sampling from p_X. We state the result in bounded-Lipschitz distance D_{BL}, which metrizes weak convergence; additional discussion and proof details are deferred to Appendix D.

Corollary 3.10. *Assume (A1)–(A3) and let p_{alg} be the output of BDDM with exponential Euler discretization with $a_t = \frac{1}{2}\,\sigma_t^2$. For $\varepsilon \in (0,1)$, choose $\sigma_0 \asymp \mathsf{m}_2/\varepsilon$, $\sigma_T \asymp \varepsilon/\sqrt{d}$, and $h \asymp \varepsilon^2/(\mathsf{k}\log(\mathsf{m}_2\sqrt{d}/\varepsilon^2))$. Then, $\mathsf{D}_{\mathrm{BL}}(p_X, p_{\mathsf{alg}}) \lesssim \tilde{\varepsilon}_{\mathrm{BD}} + \varepsilon$, provided that*

$$d \gtrsim \frac{\mathsf{k}^3}{\varepsilon^2}\log(\mathsf{m}_2\sqrt{d}/\varepsilon^2) \text{ and } N \asymp \frac{\mathsf{k}}{\varepsilon^2}\log^2(\mathsf{m}_2\sqrt{d}/\varepsilon^2)\,.$$

4. Empirical results on photographic images

Experiments on synthetic data shown in previous sections verified that (1) the noise variance can be accurately estimated from a single noisy observation in high dimensions; (2) the reverse process of Algorithm 2 closely adheres to the theoretical implicit schedule of Proposition 3.2. Do these results extend beyond the synthetic data to real-world models and signals? Additionally, can BDDMs offer a substantial gain in sampling performance compared to non-blind models,

[1] Indeed $\mathsf{Disc}(h)$ arises from (14) if we also incorporate discretization error.

Figure 3. **Left panel.** Comparison of denoising performance of blind and non-blind denoisers. Performance is evaluated on the test sets of CelebA and Bedroom class of LSUN datasets, and is reported in terms of PSNR. For consistency, the noise level on the horizontal axis is also expressed as $\mathrm{PSNR}(x, x_\sigma)$. **Right panel.** An example test image with corresponding PSNR values. **Top:** Noisy images. **Middle:** Denoised image by a non-blind denoiser. **Bottom:** Denoised image by a blind denoiser.

by eliminating the mismatch between true noise level and a proposed noise schedule? To investigate these questions, we consider deep neural networks trained on natural images. We trained a non-blind and blind model to approximate the optimal denoisers in (2) and (8), respectively. Importantly, the two models are exactly the same in all respects except for noise conditioning.

Experimental setup. Under the manifold hypothesis, natural images concentrate near a union of low-dimensional manifolds (Brown et al., 2023), making image datasets a suitable testbed for our theory. We evaluate on CelebA (Liu et al., 2015) and LSUN bedrooms (Yu et al., 2015), using matched blind and non-blind U-Net denoisers that differ only in noise conditioning; dataset, architecture, and training details are given in Appendix G.3.

Can the noise level be accurately estimated from a single noisy image by a trained neural network denoiser? If natural images are truly low dimensional, then we expect $\Pi(\cdot|x_\sigma) \simeq \delta_\sigma$. Additionally, the inductive biases of the neural network should allow for leveraging the concentration to get a precise estimate of true variance. If both of these conditions are satisfied, then the performance gap between blind and non-blind denoisers will be small. Figure 3 compares denoising error of the two models on two datasets in terms of their peak signal-to-noise ratio (PSNR) averaged across the data, $\mathrm{PSNR}(x, \hat{x}) \coloneqq -10 \log_{10} \|x - \hat{x}\|^2$ where $\hat{x}$ is the output of a blind or non-blind denoiser. Performance results in Figure 3 demonstrate that blind denoisers are as good as the non-blind ones, implying that the blind model estimates the noise level from data almost perfectly.

Now we sample from the space of natural images via Algorithm 2. Hyperparameters are set to $\sigma_{\max} = 4, \sigma_{\min} = 0.05, h = 0.05, a_t = 0.3\hat{\sigma}_t^2$, which leads to a total number of steps $N \approx 1000$. We compare to the non-blind denoisers via variance exploding (VE) of DDPM algorithm (Song et al., 2021a), with a $\log \sigma$ schedule with $N = 1000$.

Figure 4 shows the output of both sampling algorithms (see Figures 11 and 10 for additional examples and Figure 12 for LSUN samples). Surprisingly, samples generated by BDDMs appear to have higher visual quality than samples generated by the non-blind DDMs. This is while the two models have the same denoising performance as shown in Figure 3. Nevertheless BDDM samples appear to be drawn from a distribution more similar to the underlying true density than DDPM samples. *This suggests that the differences are due to the sampling algorithms.*

We hypothesize that using non-blind model in a reverse process with an explicit schedule incurs a mismatch error: the noise level most consistent with the image (the ML estimate) diverges from the noise level dictated by the schedule of the algorithm. Figure 5 illustrates the nature of the mismatch error in one-shot denoising in a non-blind model. Each plot shows the mean-squared error (MSE) between denoised and clean images, as a function of the second argument. MSE is lowest when $\sigma = \sigma^\star$. If σ is too small, the denoised image is still noisy; too large, and it looks blurry.

To quantify the mismatch in the DDPM backward sampling process, we train a separate small neural network to estimate noise level. Figure 9 in the appendix shows that this model can nearly perfectly estimate the noise level. We now use

Figure 4. Comparison of samples from BDDM and VE-DDM. **Top row:** Randomly selected subset of training images from the CelebA dataset. **Second row:** Samples generated by BDDM with $N \approx 1000$. **Third row:** Samples generated by a non-blind DDM (VE-DDPM), with $N = 1000$. Samples in each column are initialized with the same random seed, and use matched injected noise. Seeds are random and *not curated for quality*. See Appendix G.3 for lower and higher N, and for LSUN dataset.

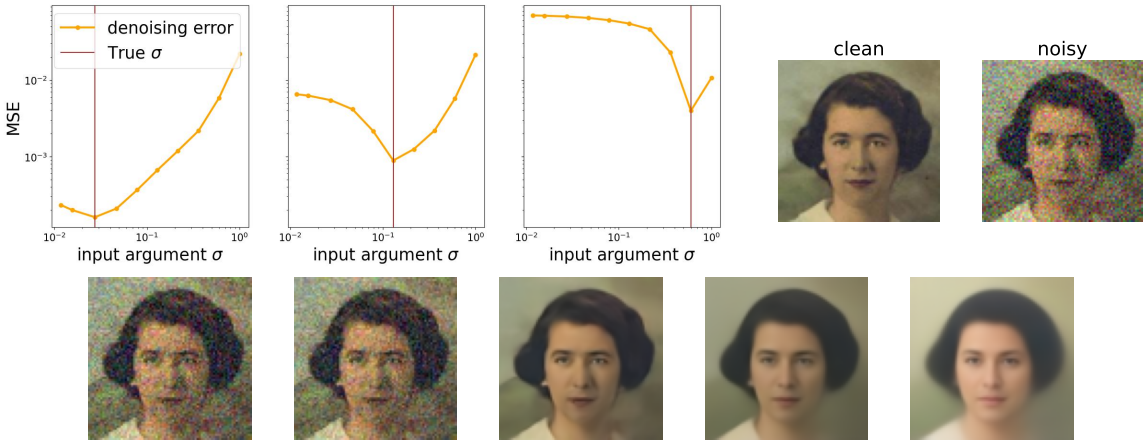

Figure 5. Sensitivity of non-blind denoising performance to mismatches in true noise level and argument to the denoiser. **Top row:** Denoising error (averaged over 512 samples) as a function of argument σ for three different true noise levels (from left to right, $\sigma^\star \in [0.025, 0.15, 0.6]$). **Bottom row:** Example clean image x_0, noisy image $x_{\sigma^\star} = x_0 + \sigma^\star z$, for $\sigma^\star = 0.1$, and five images resulting from a non-blind denoiser $\widetilde{x} = \widetilde{f}_\theta(x_{\sigma^\star}, \sigma)$ with $\sigma \in [0.01, 0.03, 0.10, 0.32, 1.0]$ (from left to right).

this model to measure the true noise level of the intermediate samples generated by DDPM, and compare them to the σ imposed by the schedule. Figure 1 shows that scheduled noise levels systematically fall behind the true σ value. This suggests that the low quality of the samples in Figures 4, 11, and 12 result from a mismatch error in which $\sigma_t > \sigma^\star$ along the trajectory. Per the leftmost plot in Figure 1, this same noise estimator model shows that BDDMs precisely track the implicit schedule of the reverse process.

5. Conclusion

This work presents a comprehensive mathematical analysis and justification of blind denoising diffusion models, closing a gap in the literature. We identify *low intrinsic dimensionality* of the underlying data distribution as a critical component underlying the success of these models both theoretically and empirically. Moreover, we have shown that BDDMs can offer improved sample diversity and quality by avoiding errors due to mismatch in noise schedules. The use of BDDMs merits further investigation at larger scales, as well as in application to other downstream tasks such as fine-tuning and inverse problems.

Impact Statement

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

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223, 2017.

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

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2023.

Chandler, D. M. and Field, D. J. Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America A*, 24(4):922–941, 2007.

Chen, F., Chewi, S., Daskalakis, C., and Rakhlin, A. High-accuracy sampling for diffusion models and log-concave distributions. *arXiv preprint 2602.01338*, 2026.

Chung, H., Kim, J., McCann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Du, Y. and Mordatch, I. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

El Alaoui, A. and Montanari, A. An information-theoretic view of stochastic localization. *IEEE Trans. Inform. Theory*, 68(11):7423–7426, 2022.

Föllmer, H. An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lect. Notes Control Inf. Sci.*, pp. 156–163. Springer, Berlin, 1985.

Gatmiry, K., Chen, S., and Salim, A. High-accuracy and dimension-free sampling with diffusions. *arXiv preprint arXiv:2601.10708*, 2026.

Gnanasambandam, A. and Chan, S. One size fits all: can we train one denoiser for all noise levels? In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3576–3586. PMLR, 7 2020.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., and Duvenaud, D. FFJORD: free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Hénaff, O. J., Ballé, J., Rabinowitz, N. C., and Simoncelli, E. P. The local low-dimensionality of natural images. *arXiv preprint 1412.6626*, 2014.

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

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

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

Kadkhodaie, Z. and Simoncelli, E. P. Solving linear inverse problems using the prior implicit in a denoiser. *arXiv preprint arXiv:2007.13640*, 2020.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Li, G. and Yan, Y. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024.

Liang, J., Huang, Z., and Chen, Y. Low-dimensional adaptation of diffusion models: convergence in total variation (extended abstract). In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*. PMLR, 7 2025.

Liu, X., Tanaka, M., and Okutomi, M. Single-image noise level estimation for blind denoising. *IEEE Transactions on Image Processing*, 22(12):5226–5237, 2013.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Miyasawa, K. An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.*, 38: 181–188, 1961.

Mohan, S., Kadkhodaie, Z., Simoncelli, E. P., and Fernandez-Granda, C. Robust and interpretable blind image denoising via bias-free convolutional neural networks. In *Int'l Conf on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 4 2020.

Olshausen, B. A. and Field, D. J. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333, 1996.

Pooladian, A.-A. and Niles-Weed, J. Plug-in estimation of Schrödinger bridges. *SIAM Journal on Mathematics of Data Science*, 7(3):1315–1336, 2025.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Raphan, M. and Simoncelli, E. P. Least squares estimation without priors or supervision. *Neural Computation*, 23(2): 374–420, 2 2011.

Robbins, H. An empirical bayes approach to statistics. In *Proc Third Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pp. 157–163. University of CA Press, 1956.

Ronneberger, O., Fischer, P., and Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Int'l Conf Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Springer, 2015.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326, 2000.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

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

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

Sun, Q., Jiang, Z., Zhao, H., and He, K. Is noise conditioning necessary for denoising generative models? In *Forty-second International Conference on Machine Learning*, 2025.

Wang, R. and Du, Y. Equilibrium matching: generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

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

A. Proofs for Sections 3.1 and 3.2

A.1. Proof of Proposition 3.1

The population counterpart to (5) is

$$f^\star = \operatorname*{arg\,min}_{f:\mathbb{R}^d\to\mathbb{R}^d} \mathbb{E}\|X - f(X+\sigma z)\|^2 \,,$$

where the expectation is under $X \sim p_X$, $\sigma \sim \Pi_0$, and $z \sim \mathcal{N}(0, I)$. Let $y := X + \sigma z$. The optimal function f of y to predict X is the conditional expectation $\mathbb{E}[X|y]$, due to orthogonality:

$$\mathbb{E}\|X - f(y)\|^2 = \mathbb{E}\|X - \mathbb{E}[X|y]\|^2 + \mathbb{E}\|\mathbb{E}[X|y] - f(y)\|^2 \,. \tag{16}$$

On the other hand, by Bayes, we can express

$$\mathbb{E}[X|y] = \int \mathbb{E}[X|\sigma, y] \,\mathrm{d}\Pi(\sigma|y) \,.$$

Applying Tweedie's identity (2),

$$\mathbb{E}[X|y] = \int \big(y + \sigma^2 \,\nabla \log p_\sigma(y)\big) \,\mathrm{d}\Pi(\sigma|y) \,.$$

We remark that by (16), for any estimator $\widehat{f}$, the excess risk is

$$\begin{aligned} \mathcal{E}(\widehat{f}) &:= \mathbb{E}\|X - \widehat{f}(y)\|^2 - \mathbb{E}\|X - f^\star(y)\|^2 \\ &= \mathbb{E}\|\widehat{f}(y) - f^\star(y)\|^2 = \int \|\widehat{f} - f^\star\|^2_{L^2(p_\sigma)} \,\mathrm{d}\Pi_0(\sigma) \,. \end{aligned} \tag{17}$$

A.2. Proof of Proposition 3.2

Let $\rho_t := \mathsf{Law}(X_t)$ and $\tilde{\rho}_t := p_{\sigma_t} := p_X * \gamma_{\sigma_t^2}$, where we abbreviate $\gamma_{\sigma^2} := \mathcal{N}(0, \sigma^2 I)$ for arbitrary $\sigma > 0$. We first compute directly $\partial_t \tilde{\rho}_t$ from this definition using the fact that the Gaussian density is the solution to the heat equation:

$$\begin{aligned} \partial_t \tilde{\rho}_t(y) &= \partial_t \int \gamma_{\sigma_t^2}(y) \, p_X(y-x) \,\mathrm{d}x \\ &= \int (\partial_t \gamma_{\sigma_t^2}(y)) \, p_X(y-x) \,\mathrm{d}x \\ &= \int (\partial_t \sigma_t^2) \, (\partial_s \gamma_s|_{s=\sigma_t^2}(y)) \, p_X(y-x) \,\mathrm{d}x \\ &= \frac{1}{2} (\partial_t \sigma_t^2) \int \Delta \gamma_{\sigma_t^2}(y) \, p_X(y-x) \,\mathrm{d}x \\ &= \frac{1}{2} (\partial_t \sigma_t^2) \, \Delta_y \int \gamma_{\sigma_t^2}(y) \, p_X(y-x) \,\mathrm{d}x \\ &= \frac{1}{2} (\partial_t \sigma_t^2) \, \Delta \tilde{\rho}_t(y) \,. \end{aligned} \tag{18}$$

On the other hand, under the ideal dynamics (10), the Fokker–Planck equation reads

$$\begin{aligned} \partial_t \rho_t &= -\nabla \cdot (\rho_t \sigma_t^2 \nabla \log p_{\sigma_t}) + a_t \, \Delta \rho_t \\ &= -\nabla \cdot (\rho_t \sigma_t^2 \nabla \log \tilde{\rho}_t) + a_t \, \Delta \rho_t \,. \end{aligned} \tag{19}$$

Since $\Delta \tilde{\rho}_t = \nabla \cdot (\tilde{\rho}_t \nabla \log \tilde{\rho}_t)$, (18) shows that $(\tilde{\rho}_t)_{t\in[0,T]}$ solves the equation (19), provided that

$$\frac{1}{2} \, \partial_t(\sigma_t^2) = -\sigma_t^2 + a_t \,. \tag{20}$$

By uniqueness of solutions to the Fokker–Planck equation, we see that with this choice of $(\sigma_t)_{t\in[0,T]}$ and for $\rho_0 = \tilde{\rho}_0$, we have $\rho_t = \tilde{\rho}_t$ for all $t \in [0, T]$. Solving the ODE (20) yields the claim.

A.3. Proof of Lemma 3.3

Recall that $\frac{1}{2}\,\partial_t(\sigma_t^2) = -\sigma_t^2 + a_t$, so we want to show $a_t \le \sigma_t^2$ for all $t \ge 0$. Note that if $a_t \le a_0 \le \sigma_0^2$ for all $t \ge 0$, then

$$\sigma_t^2 = \sigma_0^2\, e^{-2t} + 2\int_0^t a_s\, e^{-2(t-s)}\,\mathrm{d}s \ge \sigma_0^2\, e^{-2t} + a_t\,(1 - e^{-2t})\,.$$

This is lower bounded by a_t provided $a_t \le \sigma_0^2$.

A.4. Proof of Lemma 3.4

The first term in Proposition 3.2 is obviously decaying to zero, so we must show that $2\int_0^t a_s e^{-2(t-s)}\,\mathrm{d}s \to 0$. Since $a_t \to 0$, for any $\varepsilon > 0$ there exists t_0 such that $a_t \le \varepsilon$ for all $t \ge t_0$. Then, for $t \ge t_0$,

$$\begin{aligned}\int_0^t 2a_s e^{-2(t-s)}\,\mathrm{d}s &\le \int_0^{t_0}(\cdots) + \int_{t_0}^t(\cdots)\\ &\le a_0\,(e^{-2(t-t_0)} - e^{-2t}) + \varepsilon\,(1 - e^{-2(t-t_0)})\\ &\le a_0\,(e^{-2(t-t_0)} - e^{-2t}) + \varepsilon\,.\end{aligned}$$

We can choose t sufficiently large to make the first term at most ε as well.

B. Proofs for Sections 3.3 and 3.4

We use the following notation throughout the appendix. Recalling that $p_\sigma \coloneqq p * \mathcal{N}(0, \sigma^2 I)$, we write q_σ for the joint distribution under which $X \sim p_X$, $Y \sim \mathcal{N}(X, \sigma^2 I)$. We also denote

$$f_\sigma^\star(y) \coloneqq \mathbb{E}_{q_\sigma}[X \mid Y = y]\,, \tag{21}$$

$$C_\sigma(y) \coloneqq \mathbb{E}_{q_\sigma}[(X - f_\sigma^\star(y))^{\otimes 2} \mid Y = y] = \mathrm{Cov}_{q_\sigma}(X \mid Y = y)\,, \tag{22}$$

$$W_\sigma(y) \coloneqq \mathrm{Cov}_{q_\sigma}(X, \|X - y\|^2 \mid Y = y)\,. \tag{23}$$

Also, since we repeatedly encounter the quantity $\mathsf{k} + \log(1/\delta)$ for some $\delta > 0$, we abbreviate this by k_δ.

B.1. Proof of Lemma 3.7

The change-of-variables formula is straightforward as $|\frac{\mathrm{d}\sigma}{\mathrm{d}\lambda}| = \frac{1}{2}\,\lambda^{-3/2}$. For the next part, we write

$$\ell(\lambda|y) = -\log\overline{\Pi}(\lambda|y) = -\frac{d-\alpha}{2}\log\lambda - \log\mathbb{E}_{X\sim p_X}\big[\exp\big(-\frac{\lambda}{2}\,\|X - y\|^2\big)\big] + c\,,$$

where $c > 0$ is an absolute constant. We compute via chain rule

$$\ell'(\lambda|y) = -\frac{d-\alpha}{2\lambda} + \frac{1}{2}\,\mathbb{E}_{q_\lambda}[\|X - Y\|^2 \mid Y = y]\,,$$

where $q_\lambda(x|y) \propto \exp\big(-\frac{\lambda}{2}\,\|x - y\|^2\big)\,p_X(x)$. Using the following fact (which can be verified via chain rule)

$$\partial_\lambda \mathbb{E}_{q_\lambda}[\|X - y\|^2 \mid Y = y] = -\frac{1}{2}\,\mathrm{Var}(\|X - y\|^2 \mid Y = y)\,,$$

the second derivative is computed similarly, resulting in

$$\ell''(\lambda|y) = \frac{d-\alpha}{2\lambda^2} - \frac{1}{4}\,\mathrm{Var}_{q_\lambda}[\|X - Y\|^2 \mid Y = y]\,.$$

B.2. Proof of Proposition 3.8

In this section, we use the notation

$$\mathsf{e}_\delta \coloneqq \sqrt{\frac{\log(1/\delta)}{d}} + \frac{\mathsf{k}_\delta}{d}\,.$$

We will show that the noise posterior concentrates up to a relative error of e_δ.

Theorem B.1. *Let $\delta \in (0,1)$ and assume that $d \gg \mathsf{k}_\delta$. Then, the following holds with probability at least $1-\delta$ over $X_t \sim p_{\sigma_t}$. For all $a \gg \lambda_t \mathsf{e}_\delta$,*

$$\overline{\Pi}(|\lambda - \lambda_t| \geq a \mid X_t) \lesssim \frac{\mathsf{e}_\delta}{\sqrt{d}} \exp\big(-\Omega\big(\frac{da^2}{\lambda_t^2}\big)\big) + \frac{\lambda_{\max}}{\lambda_t \mathsf{e}_\delta} \exp(-\Omega(d)) \,.$$

Proof. We apply Proposition F.2, noting that

$$\mathbb{E}_{q_\lambda}[\|X - Y\|^2 \mid Y = y] = D_{\lambda^{-1/2},2}(y,y) \,.$$

Hence, with probability at least $1-\delta$,

$$\begin{aligned} 2\ell'(\lambda|X_t) &= -\frac{d-\alpha}{\lambda} + D_{\lambda^{-1/2},2}(X_t, X_t) \\ &= -\frac{d \pm O(\mathsf{k}_\delta)}{\lambda} + \frac{d \pm O(\sqrt{d\log(1/\delta)} + \mathsf{k}_\delta))}{\lambda_t} \,. \end{aligned}$$

Define the error term

$$\mathsf{E} \coloneqq \sqrt{d\log(1/\delta)} + \mathsf{k}_\delta \,.$$

Note that for $\lambda > \lambda_t$,

$$2\ell'(\lambda|X_t) \geq \frac{d}{\lambda_t}\Big[1 - \frac{\lambda_t}{\lambda} - \frac{O(\mathsf{E})}{d}\Big] \gtrsim \begin{cases} d\,(\lambda - \lambda_t)/\lambda_t^2 \,, & C_0 \lambda_t \mathsf{E}/d \leq \lambda - \lambda_t \leq \lambda_t \,, \\ d/\lambda_t \,, & \lambda - \lambda_t \geq \lambda_t \,, \end{cases}$$

where $C_0 > 0$ is a universal constant and we require that $\mathsf{E}/d \ll 1$, i.e., $d \gg \mathsf{k}_\delta$. Therefore, for $a_0 \coloneqq C_0 \lambda_t \mathsf{E}/d$ and $\lambda \geq \lambda_t + 2a_0$, we can integrate to find that

$$\begin{aligned} \overline{\Pi}(\lambda|X_t) &= \overline{\Pi}(\lambda_t + a_0|X_t) \exp\Big(-\int_{\lambda_t + a_0}^{\lambda} \ell'(\cdot|X_t)\Big) \\ &\leq \overline{\Pi}(\lambda_t + a_0|X_t) \times \begin{cases} \exp\big(-\Omega\big(\frac{d\,(\lambda-\lambda_t)^2}{\lambda_t^2}\big)\big) \,, & \lambda_t + 2a_0 \leq \lambda \leq 2\lambda_t \,, \\ \exp(-\Omega(d)) \,, & \lambda \geq 2\lambda_t \,. \end{cases} \end{aligned}$$

Hence, for $a \geq 2a_0$,

$$\begin{aligned} \int_{\lambda_t + a}^{\infty} \mathrm{d}\overline{\Pi}(\lambda|X_t) &\leq \int_{\lambda_t + a}^{2\lambda_t} \overline{\Pi}(\lambda_t + a_0|X_t) \exp\big(-\Omega\big(\frac{d\,(\lambda-\lambda_t)^2}{\lambda_t^2}\big)\big)\,\mathrm{d}\lambda \\ &\qquad + \int_{2\lambda_t}^{\lambda_{\max}} \overline{\Pi}(\lambda_t + a_0|X_t) \exp(-\Omega(d))\,\mathrm{d}\lambda \\ &\lesssim \Big[\frac{\lambda_t}{d^{1/2}} \exp\big(-\Omega\big(\frac{da^2}{\lambda_t^2}\big)\big) + \lambda_{\max} \exp(-\Omega(d))\Big] \overline{\Pi}(\lambda_t + a_0|X_t) \,. \end{aligned}$$

On the other hand, for $\lambda \leq \lambda_t + a_0$, one has $|\ell'(\lambda|X_t)| \lesssim \mathsf{E}/\lambda_t$. Hence,

$$\begin{aligned} \overline{\Pi}(\lambda_t + a_0|X_t) &= \frac{1}{a_0} \int_{\lambda_t}^{\lambda_t + a_0} \overline{\Pi}(\lambda_t + a_0|X_t)\,\mathrm{d}\lambda \\ &\lesssim \frac{d}{\lambda_t \mathsf{E}} \int_{\lambda_t}^{\lambda_t + a_0} \overline{\Pi}(\lambda|X_t) \exp\big(O\big(\frac{a_0 \mathsf{E}}{\lambda_t}\big)\big)\,\mathrm{d}\lambda \\ &\leq \frac{d}{\lambda_t \mathsf{E}} \exp\big(O\big(\frac{\mathsf{E}^2}{d}\big)\big) \,, \end{aligned}$$

and thus

$$\int_{\lambda_t + a}^{\infty} \mathrm{d}\overline{\Pi}(\lambda|X_t) \lesssim \frac{d^{1/2}}{\mathsf{E}} \exp\big(O\big(\frac{\mathsf{E}^2}{d}\big) - \Omega\big(\frac{da^2}{\lambda_t^2}\big)\big) + \frac{\lambda_{\max} d}{\lambda_t \mathsf{E}} \exp\big(O\big(\frac{\mathsf{E}^2}{d}\big) - \Omega(d)\big) \,.$$

Since we are assuming that $\mathsf{E} \ll d$, the last exponential is $\exp(-\Omega(d))$. For all $a \gg \lambda_t \mathsf{E}/d$, the bound becomes

$$\int_{\lambda_t+a}^{\infty} \mathrm{d}\overline{\Pi}(\lambda|X_t) \lesssim \frac{d^{1/2}}{\mathsf{E}} \exp\big(-\Omega\big(\frac{da^2}{\lambda_t^2}\big)\big) + \frac{\lambda_{\max} d}{\lambda_t \mathsf{E}} \exp(-\Omega(d)) \,.$$

A similar argument gives the corresponding lower bound. □

Integrating this high probability bound yields a bound in mean-squared error.

Proof of Proposition 3.8. Work over the event of probability at least $1-\delta$ under which Theorem B.1 holds. Then, provided that $d \gg \mathsf{k}_\delta$, for $\kappa := \lambda_{\max}/\lambda_{\min}$,

$$\begin{aligned}
\int |\lambda - \lambda_t|^2 \,\mathrm{d}\overline{\Pi}(\lambda|X_t) &\lesssim \lambda_t^2 \mathsf{e}_\delta^2 + \int_{|\lambda-\lambda_t| \geq C\mathsf{e}_\delta} |\lambda-\lambda_t|^2 \,\mathrm{d}\overline{\Pi}(\lambda|X_t) \\
&\lesssim \lambda_t^2 \mathsf{e}_\delta^2 + \int_0^{\lambda_{\max}} a \left(\frac{\mathsf{e}_\delta}{\sqrt{d}} \exp\big(-\Omega\big(\frac{da^2}{\lambda_t^2}\big)\big) + \frac{\lambda_{\max}}{\lambda_t \mathsf{e}_\delta} \exp(-\Omega(d)) \right) \mathrm{d}a \\
&\lesssim \lambda_t^2 \left[\mathsf{e}_\delta^2 + \frac{\mathsf{e}_\delta}{d^{3/2}} + \frac{\lambda_{\max}^3}{\lambda_t^3 \mathsf{e}_\delta} \exp(-\Omega(d)) \right] \\
&\lesssim \lambda_t^2 \left[\frac{\log(1/\delta)}{d} + \frac{\mathsf{k}_\delta^2}{d^2} + \frac{\sqrt{\log(1/\delta)}}{d^2} + \frac{\mathsf{k}_\delta}{d^{5/2}} + \kappa^3 \left(\sqrt{d} + \frac{d}{\mathsf{k}}\right) \exp(-\Omega(d)) \right] \lesssim \lambda_t^2 \mathsf{e}_\delta^2 \,,
\end{aligned}$$

provided that $d \gg \log \kappa$.

In conclusion, for all $\delta \in (0,1)$, if $d \geq C\left(\mathsf{k} + \log(1/\delta) + \log(\sigma_0/\sigma_T)\right)$ for a sufficiently large absolute constant $C > 0$, then with probability at least $1-\delta$ it holds that

$$\int |\lambda - \lambda_t|^2 \,\mathrm{d}\overline{\Pi}(\lambda|X_t) \lesssim \lambda_t^2 \left(\frac{\log(1/\delta)}{d} + \frac{\mathsf{k}^2 + \log^2(1/\delta)}{d^2} \right) . \tag{24}$$

□

B.3. Proof of Theorem 3.6

Recall the KL error decomposition (14). For the first term, we use the standard fact

$$\mathrm{KL}(p_{\sigma_0} \| \widehat{p}_0) = \mathrm{KL}(p_X * \gamma_{\sigma_0^2} \| \gamma_{\sigma_0}^2) \leq \frac{1}{2\sigma_0^2} W_2^2(p_X, \delta_0) = \frac{\mathbb{E}_{X\sim p_X}[\|X\|^2]}{2\sigma_0^2} \lesssim \frac{\mathsf{m}_2^2}{\sigma_0^2} \,.$$

The second term is due to the definition of $\varepsilon_{\mathrm{BD}}^2$, and now we focus our attention on controlling the third term.

We work over the event of probability at least $1-\delta$ under which both Theorem B.1 and Corollary F.3 hold. Then,

$$\begin{aligned}
\Big| \int_\sigma^{\sigma_t} \|W_\omega(X_t)\| \, \omega^{-3} \,\mathrm{d}\omega \Big|^2 &\lesssim \mathsf{k}_\delta^3 \Big| \int_\sigma^{\sigma_t} \sigma_t^3 \, \omega^{-3} \,\mathrm{d}\omega + \int_\sigma^{\sigma_t} \mathrm{d}\omega \Big|^2 \\
&\lesssim \sigma_t^2 \, \mathsf{k}_\delta^3 \left(|1 - \sigma_t^2/\sigma^2|^2 + |1 - \sigma/\sigma_t|^2 \right) \\
&= \frac{\mathsf{k}_\delta^3}{\lambda_t} \left(|1 - \lambda/\lambda_t|^2 + |1 - \sqrt{\lambda_t/\lambda}|^2 \right) .
\end{aligned} \tag{25}$$

Now, we integrate this w.r.t. $\overline{\Pi}(\cdot|X_t)$. By (24), if $d \gg \mathsf{k}_\delta + \log \kappa$,

$$\int |1 - \lambda/\lambda_t|^2 \,\mathrm{d}\overline{\Pi}(\lambda|X_t) \lesssim \mathsf{e}_\delta^2 \,.$$

Similarly,

$$\begin{aligned}
&\int |1-\sqrt{\lambda_t/\lambda}|^2 \,\mathrm{d}\overline{\Pi}(\lambda|X_t) \\
&\quad \lesssim \mathsf{e}_\delta^2 + \frac{1}{\lambda_t^2}\int_{C\mathsf{e}_\delta \le |\lambda-\lambda_t| \le \lambda_t/2} |\lambda-\lambda_t|^2 + \frac{\lambda_t}{\lambda_{\min}}\,\overline{\Pi}(|\lambda-\lambda_t| \ge \lambda_t/2 \mid X_t) \\
&\quad \lesssim \mathsf{e}_\delta^2 + \frac{\lambda_t}{\lambda_{\min}}\big(\frac{\mathsf{e}_\delta}{\sqrt{d}} + \frac{\lambda_{\max}}{\lambda_t \mathsf{e}_\delta}\big)\exp(-\Omega(d)) \lesssim \mathsf{e}_\delta^2\,.
\end{aligned}$$

We have shown that for all δ such that $d \gg \mathsf{k}_\delta + \log\kappa$, with probability at least $1-\delta$,

$$\|(r^\star_{\sigma_t} - r^\star)(X_t)\|^2 \lesssim \frac{\mathsf{k}_\delta^3}{\lambda_t}\,\mathsf{e}_\delta^2\,.$$

On the other hand, for smaller values of δ, from (25) we still have the crude bound

$$\|(r^\star_{\sigma_t} - r^\star)(X_t)\|^2 \lesssim \frac{\mathsf{k}_\delta^3}{\lambda_t}\,\kappa^4\,.$$

Therefore, for a universal constant $c>0$, we deduce that for *all* $\delta \in (0,1)$, if $d \gg \mathsf{k} + \log\kappa$,

$$\|(r^\star_{\sigma_t} - r^\star)(X_t)\|^2 \lesssim \frac{\mathsf{k}_\delta^3}{\lambda_t}\left[\mathsf{e}_\delta^2 + \kappa^4\,\mathbf{1}_{\delta \le \exp(-cd)}\right].$$

Integrating this high probability bound yields

$$\|r^\star_{\sigma_t} - r^\star\|^2_{L^2(p_{\sigma_t})} \lesssim \frac{\mathsf{k}^3}{\lambda_t}\left[\frac{1}{d} + \frac{\mathsf{k}^2}{d^2} + \kappa^4\exp(-\Omega(d))\right] \lesssim \frac{\mathsf{k}^3}{\lambda_t}\big(\frac{1}{d} + \frac{\mathsf{k}^2}{d^2}\big)\,.$$

Thus, the noise estimation term becomes

$$\int_0^T \frac{1}{a_t}\,\|r_{\sigma_t} - r^\star\|^2_{L^2(p_{\sigma_t})}\,\mathrm{d}t \lesssim \big(\frac{\mathsf{k}^3}{d} + \frac{\mathsf{k}^5}{d^2}\big)\int_0^T \frac{\sigma_t^2}{a_t}\,\mathrm{d}t\,.$$

C. Proofs for Section 3.5

C.1. Exponential Euler discretization

In order to establish discretization guarantees which only scale with the intrinsic dimension, we use the exponential Euler discretization: for fixed step size $h>0$, the algorithm follows

$$\mathrm{d}Y_t = (\widehat{f}(Y_{t_-}) - Y_t)\,\mathrm{d}t + \sqrt{2a_t}\,\mathrm{d}B_t\,, \tag{26}$$

where $t_- := \lfloor t/h \rfloor\, h$. In other words, we freeze the non-linear term $\widehat{f}$ and exactly integrate the rest, including the linear term. Some intuition for this can be seen as follows: if, instead of $\widehat{f}$, we had $f^\star_{\sigma_t}$, then the Jacobian of the drift is $C_{\sigma_t}/\sigma_t^2 - I$. Generally, discretization error bounds rely on Lipschitz bounds on the drift, i.e., bounds on the operator norm of this Jacobian. However, among the two terms in the Jacobian, only the first—a conditional covariance matrix—is expected to be controlled in terms of the intrinsic dimension. This suggests that to avoid dependence on the ambient dimension, we should exactly integrate the $-Y_t$ term.

C.2. Proof of Theorem 3.9

Recall that $f^\star_{\sigma_t} = \mathrm{id} + \sigma_t^2\,\nabla\log p_{\sigma_t}$. To control the discretization error, we first need an Itô calculation.

Lemma C.1. *Let $(\sigma_t)_{t\ge0}$ be as in Proposition 3.2. Then*

$$\partial_t f^\star_{\sigma_t}(y) = \dot\sigma_t\,\partial_\omega(\omega^2\,\nabla\log p_\omega(y))\big|_{\omega=\sigma_t} = \Big(\frac{a_t}{\sigma_t^4} - \frac{1}{\sigma_t^2}\Big)\,W_{\sigma_t}(y)\,.$$

Proof. Follows from the chain rule and the final expression for $\dot\sigma_t$ from (20). □

Proposition C.2. *For $(X_t)_{t\geq 0}$ following (10), it holds that*

$$\begin{aligned}\mathrm{d}f^\star_{\sigma_t}(X_t) = \Big\{ \frac{2a_t - \sigma_t^2}{\sigma_t^4} \left(W_{\sigma_t}(X_t) - C_{\sigma_t}(X_t) \left(f^\star_{\sigma_t}(X_t) - X_t \right) \right) \Big\} \, \mathrm{d}t \\ + \sqrt{2a_t} \sigma_t^{-2} \, C_{\sigma_t}(X_t) \, \mathrm{d}B_t \, .\end{aligned}$$

Proof. Writing out Itô's lemma,

$$\mathrm{d}f^\star_{\sigma_t}(X_t) = (\partial_t f^\star_{\sigma_t})(X_t) \, \mathrm{d}t + \langle \nabla f^\star_{\sigma_t}(X_t), \, \mathrm{d}X_t \rangle + \frac{1}{2} \left\langle \mathrm{d}X_t, \nabla^2 f^\star_{\sigma_t}(X_t) \, \mathrm{d}X_t \right\rangle ,$$

we see that the only terms that remain are

$$\begin{aligned}\mathrm{d}f^\star_{\sigma_t}(X_t) = \left[(\partial_t f^\star_{\sigma_t})(X_t) + \nabla f^\star_{\sigma_t}(X_t) \left(f^\star_{\sigma_t}(X_t) - X_t \right) + a_t \, \Delta f^\star_{\sigma_t}(X_t) \right] \mathrm{d}t \\ + \sqrt{2a_t} \, \nabla f^\star_{\sigma_t}(X_t) \, \mathrm{d}B_t \, .\end{aligned}$$

From Lemma C.1 we have that

$$\partial_t f^\star_{\sigma_t}(y) = \frac{a_t - \sigma_t^2}{\sigma_t^4} \, W_{\sigma_t}(y)$$

and by Tweedie's formula (recall Lemma F.1)

$$a_t \, \Delta f^\star_{\sigma_t} = a_t \sigma_t^2 \, \Delta \nabla \log p_{\sigma_t} = a_t \sigma_t^2 \, \nabla \Delta \log p_{\sigma_t} = a_t \sigma_t^{-2} \, \nabla \operatorname{tr} C_{\sigma_t} \, ,$$

where the identity matrix drops due to the additional gradient. Finally, note that

$$\nabla f^\star_{\sigma_t}(y) \left(f^\star_{\sigma_t}(y) - y \right) = \sigma_t^{-2} \, C_{\sigma_t}(y) \left(f^\star_{\sigma_t}(y) - y \right) .$$

Collecting these three terms and invoking the last result in Lemma F.1, we obtain

$$\begin{aligned}&\big((\partial_t f^\star_{\sigma_t}) + \nabla f^\star_{\sigma_t} \left(f^\star_{\sigma_t} - \mathrm{id} \right) + a_t \, \Delta f^\star_{\sigma_t} \big)(y) \\ &= \frac{a_t - \sigma_t^2}{\sigma_t^4} \, W_{\sigma_t}(y) + a_t \sigma_t^{-2} \, \nabla \operatorname{tr} C_{\sigma_t}(y) + \sigma_t^{-2} \, C_{\sigma_t}(y) \left(f^\star_{\sigma_t}(y) - y \right) \\ &= \Big(\frac{a_t - \sigma_t^2}{\sigma_t^4} \Big) W_{\sigma_t}(y) + \Big(\frac{a_t}{\sigma_t^4} \, W_{\sigma_t}(y) - \frac{2a_t}{\sigma_t^4} \, C_{\sigma_t}(y) \left(f^\star_{\sigma_t}(y) - y \right) \Big) \\ &\qquad + \sigma_t^{-2} \, C_{\sigma_t}(y) \left(f^\star_{\sigma_t}(y) - y \right) \\ &= \frac{2a_t - \sigma_t^2}{\sigma_t^4} \left(W_{\sigma_t}(y) - C_{\sigma_t}(y) \left(f^\star_{\sigma_t}(y) - y \right) \right) .\end{aligned}$$

□

Lemma C.3. *For $X_\sigma \sim p_\sigma$,*

$$\partial_\sigma \, \mathbb{E} \operatorname{tr} C_\sigma(X_\sigma) = \frac{2}{\sigma^3} \, \mathbb{E} \operatorname{tr} \big(C_\sigma(X_\sigma)^2 \big) .$$

Proof. This is a variant of the standard MMSE identity. For example, El Alaoui & Montanari (2022, §II.C) with $\boldsymbol{Q} = \boldsymbol{R} = I$ shows that

$$\partial_\lambda \, \mathbb{E} \operatorname{tr} C_{1/\sqrt{\lambda}}(X_{1/\sqrt{\lambda}}) = -\mathbb{E} \operatorname{tr} \big(C_{1/\sqrt{\lambda}}(X_{1/\sqrt{\lambda}})^2 \big) ,$$

with $\lambda := \sigma^{-2}$. The claimed result follows from the chain rule. □

We now present the main computations from Section 3.5. Applying Girsanov's theorem to the ideal process (10),

denoted P, and the exponential Euler discretization (26), denoted $\widehat{\mathsf{P}}$, the error becomes

$$\begin{aligned}
\mathrm{KL}(\mathsf{P}\|\widehat{\mathsf{P}}) &\lesssim \mathrm{KL}(p_0\|\widehat{p}_0) + \mathbb{E}\int_0^T \frac{1}{a_t}\,\|\widehat{f}_\theta(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2\,\mathrm{d}t \\
&\lesssim \mathrm{KL}(p_0\|\widehat{p}_0) + \mathbb{E}\int_0^T \frac{1}{a_t}\,\|\widehat{f}_\theta(X_{t_-}) - f^\star(X_{t_-})\|^2\,\mathrm{d}t \\
&\qquad + \mathbb{E}\int_0^T \frac{1}{a_t}\left(\|f^\star(X_{t_-}) - f^\star_{\sigma_{t_-}}(X_{t_-})\|^2 + \|f^\star_{\sigma_{t_-}}(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2\right)\mathrm{d}t \\
&\lesssim \mathrm{KL}(p_0\|\widehat{p}_0) + \tilde{\varepsilon}^2_{\mathrm{BD}} + \mathbb{E}\int_0^T \frac{1}{a_t}\,\|f^\star(X_{t_-}) - f^\star_{\sigma_{t_-}}(X_{t_-})\|^2\,\mathrm{d}t \\
&\qquad + \mathbb{E}\int_0^T \frac{1}{a_t}\,\|f^\star_{\sigma_{t_-}}(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2\,\mathrm{d}t \\
&\lesssim \frac{R^2}{\sigma_0^2} + \tilde{\varepsilon}^2_{\mathrm{BD}} + \Big(\frac{\mathsf{k}^3}{d} + \frac{\mathsf{k}^5}{d^2}\Big)\int_0^T \frac{\sigma_t^2}{a_t}\,\mathrm{d}t + \int_0^T \frac{1}{a_t}\,\mathbb{E}\|f^\star_{\sigma_{t_-}}(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2\,\mathrm{d}t
\end{aligned} \tag{27}$$

where we used the result from Theorem 3.6 in the last line.

For the last term, we have by Proposition C.2 and the Itô isometry,

$$\begin{aligned}
&\mathbb{E}\|f^\star_{\sigma_{t_-}}(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2 \\
&\quad = \mathbb{E}\Big\|\int_{t_-}^t \Big\{\frac{2a_s - \sigma_s^2}{\sigma_s^4}\left(W_{\sigma_s}(X_s) - C_{\sigma_s}(X_s)\,(f^\star_{\sigma_s}(X_s) - X_s)\right)\Big\}\,\mathrm{d}s \\
&\qquad\qquad + \int_{t_-}^t \sqrt{2a_s}\sigma_s^{-2}\,C_{\sigma_s}(X_s)\,\mathrm{d}B_s\Big\|^2 \\
&\quad \lesssim h\int_{t_-}^t \Big(\frac{2a_s - \sigma_s^2}{\sigma_s^4}\Big)^2\,\mathbb{E}\|W_{\sigma_s}(X_s) - C_{\sigma_s}(X_s)\,(f^\star_{\sigma_s}(X_s) - X_s)\|^2\,\mathrm{d}s \\
&\qquad\qquad + \int_{t_-}^t a_s\sigma_s^{-4}\,\mathbb{E}\|C_{\sigma_s}(X_s)\|_{\mathrm{F}}^2\,\mathrm{d}s\,.
\end{aligned}$$

We now specialize to the class of noise schedules with $a_t = \mathsf{a}\sigma_t^2$ for some $\mathsf{a} \in (0,1)$. In this case, we have $\sigma_t = \sigma_0\exp(-(1-\mathsf{a})\,t)$ for $t \geq 0$, and

$$\begin{aligned}
&\mathbb{E}\|f^\star_{\sigma_{t_-}}(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2 \\
&\quad \lesssim \Big(\mathsf{a} - \frac{1}{2}\Big)^2 h\int_{t_-}^t \frac{1}{\sigma_s^4}\,\mathbb{E}\|W_{\sigma_s}(X_s) - C_{\sigma_s}(X_s)\,(f^\star_{\sigma_s}(X_s) - X_s)\|^2\,\mathrm{d}s \\
&\qquad + \mathsf{a}\int_{t_-}^t \frac{1}{\sigma_s^2}\,\mathbb{E}\|C_{\sigma_s}(X_s)\|_{\mathrm{F}}^2\,\mathrm{d}s\,.
\end{aligned}$$

Note that for any non-negative $(b_t)_{t\in[0,T]}$, $T = Nh$,

$$\begin{aligned}
\int_0^T \frac{1}{a_t}\int_{t_-}^t b_s\,\mathrm{d}s\,\mathrm{d}t &= \sum_{n=0}^{(N-1)h}\int_{nh}^{(n+1)h} \frac{1}{a_t}\int_{nh}^t b_s\,\mathrm{d}s\,\mathrm{d}t \\
&= \sum_{n=0}^{(N-1)h}\int_{nh}^{(n+1)h} b_s\int_s^{(n+1)h} \frac{1}{a_t}\,\mathrm{d}t\,\mathrm{d}s \lesssim h\int_0^T \frac{b_s}{a_s}\,\mathrm{d}s\,,
\end{aligned}$$

where we use the fact that $|\partial_t \log a_t| \lesssim 1$ and hence $a_s \asymp a_t$ for $|s-t| \lesssim 1$. Thus, the total discretization error becomes

$$\begin{aligned}\mathsf{Disc}(h) &:= \int_0^T \frac{1}{a_t}\,\mathbb{E}\|f^\star_{\sigma_{t_-}}(X_{t_-}) - f^\star_{\sigma_t}(X_t)\|^2\,\mathrm{d}t \\ &\lesssim \frac{(\mathsf{a}-\frac{1}{2})^2}{\mathsf{a}}\,h^2 \int_0^T \frac{1}{\sigma_s^6}\,\mathbb{E}\|W_{\sigma_s}(X_s) - C_{\sigma_s}(X_s)\,(f^\star_{\sigma_s}(X_s) - X_s)\|^2\,\mathrm{d}s \\ &\quad + h\int_0^T \frac{1}{\sigma_s^4}\,\mathbb{E}\|C_{\sigma_s}(X_s)\|_{\mathrm{F}}^2\,\mathrm{d}s\,.\end{aligned}$$

For the last term, note that by Lemma C.3,

$$\partial_s\,\mathbb{E}\,\mathrm{tr}\,C_{\sigma_s}(X_s) = \frac{2\dot\sigma_s}{\sigma_s^3}\,\mathbb{E}\|C_{\sigma_s}(X_{\sigma_s})\|_{\mathrm{F}}^2 = -\frac{2\,(1-\mathsf{a})}{\sigma_s^2}\,\mathbb{E}\|C_{\sigma_s}(X_{\sigma_s})\|_{\mathrm{F}}^2\,.$$

Hence, the term can be written

$$\begin{aligned}-\frac{h}{1-\mathsf{a}}\int_0^T \frac{1}{\sigma_t^2}\,\partial_t\,\mathbb{E}\,\mathrm{tr}\,C_{\sigma_t}(X_t)\,\mathrm{d}t &\le \frac{h}{1-\mathsf{a}}\Big(\frac{\mathbb{E}\,\mathrm{tr}\,C_{\sigma_0}(X_0)}{\sigma_0^2} + \int_0^T \partial_t(\sigma_t^{-2})\,\mathbb{E}\,\mathrm{tr}\,C_{\sigma_t}(X_t)\,\mathrm{d}t\Big) \\ &\lesssim \frac{h}{1-\mathsf{a}}\Big(\mathsf{k} - \int_0^T \frac{1}{\sigma_t^4}\,\partial_t\sigma_t^2\,\mathbb{E}\,\mathrm{tr}\,C_{\sigma_t}(X_t)\,\mathrm{d}t\Big) \\ &\lesssim \frac{h}{1-\mathsf{a}}\Big(\mathsf{k} + (1-\mathsf{a})\int_0^T \frac{1}{\sigma_t^2}\,\mathbb{E}\,\mathrm{tr}\,C_{\sigma_t}(X_t)\,\mathrm{d}t\Big) \\ &\lesssim \big(\frac{1}{1-\mathsf{a}} + T\big)\,\mathsf{k}h\,,\end{aligned}$$

where we applied Corollary F.3.

As for the first term, by Corollary F.3, we can bound

$$\mathbb{E}\|W_{\sigma_s}(X_s) - C_{\sigma_s}(X_s)\,(f^\star_{\sigma_s}(X_s) - X_s)\|^2 \lesssim \sigma_s^6\,\mathsf{k}^3\,.$$

which gives

$$\int_0^T \frac{1}{\sigma_s^6}\,\mathbb{E}\|W_{\sigma_s}(X_s) - C_{\sigma_s}(X_s)\,(f^\star_{\sigma_s}(X_s) - X_s)\|^2\,\mathrm{d}s \lesssim \mathsf{k}^3 T\,.$$

With this, we have our final bound

$$\mathsf{Disc}(h) \lesssim \frac{(\mathsf{a}-\frac{1}{2})^2}{\mathsf{a}}\,T\mathsf{k}^3h^2 + \big(\frac{1}{1-\mathsf{a}} + T\big)\,\mathsf{k}h\,.$$

Note that $T \asymp \log(\sigma_0/\sigma_T)/(1-\mathsf{a})$, so this can be written

$$\mathsf{Disc}(h) \lesssim \frac{(\mathsf{a}-\frac{1}{2})^2}{\mathsf{a}\,(1-\mathsf{a})}\,\mathsf{k}^3h^2\log\frac{\sigma_0}{\sigma_T} + \frac{1}{1-\mathsf{a}}\,\mathsf{k}h\log\frac{\sigma_0}{\sigma_T}\,.$$

D. Additional discussion and proof for Corollary 3.10

Theorem 3.6 guarantees closeness to $p_{\sigma_T} = p_X * \mathcal{N}(0, \sigma_T^2 I)$, not directly to p_X. One approach is to simply assume that sampling from p_{σ_T} is our goal all along for some small σ_T, which is sometimes adopted in the literature (Gatmiry et al., 2026).More commonly, p_{σ_T} is used as a technical device to sample from p_X, which is known as *early stopping*. The question is therefore how small σ_T should be in order to ensure that p_{σ_T} and p_X are close, while controlling the noise-estimation and discretization errors that also depend on σ_T. With the choice $a_t = \frac{1}{2}\sigma_t^2$ from Theorem 3.9, we can balance these requirements through σ_T, σ_0, and h.

We use the bounded Lipschitz metric,

$$\mathsf{D}_{\mathrm{BL}}(p,q) := \sup\{\mathbb{E}_p[f] - \mathbb{E}_q[f] : f \in \mathrm{BLip}\}\,,$$

where BLip is the space of functions which are 1-Lipschitz and satisfy $-1 \le f \le 1$. In Corollary 3.10, the definition of the score error is slightly modified to

$$\tilde{\varepsilon}^2_{\mathrm{BD}} \coloneqq \int_0^T a_t^{-1} \|\widehat{f}_\theta - f^\star\|^2_{L^2(p_{\sigma_{t_}})} \,\mathrm{d}t\,,$$

since the error only matters at discretization time steps. Figure 2 illustrates convergence in Corollary 3.10 for a p_X constructed as a mixture of two Gaussians.

First note that the choice $\sigma_T \asymp \varepsilon/\sqrt{d}$ implies that $\mathsf{D}_{\mathrm{BL}}(p_{\sigma_T}, p_X) \le \varepsilon$, as

$$\mathsf{D}_{\mathrm{BL}}(p_{\sigma_T}, p_X) \le W_1(p_{\sigma_T}, p_X) \le W_2(p_{\sigma_T}, p_X) \le \sigma_T\sqrt{d}\,,$$

where the last inequality follows from a standard coupling argument, where (for $p \in \{1,2\}$) W_p are the p-Wasserstein distances. By the triangle inequality, we then have

$$\mathsf{D}_{\mathrm{BL}}(\widehat{p}_T, p_X) \le \mathsf{D}_{\mathrm{BL}}(p_{\sigma_T}, p_X) + \mathsf{D}_{\mathrm{BL}}(p_{\sigma_T}, \widehat{p}_T) \le \varepsilon + \mathsf{D}_{\mathrm{BL}}(p_{\sigma_T}, \widehat{p}_T)\,.$$

To bound the remaining term, note that by Pinsker's inequality (as the total variation distance is an upper bound on D_{BL}), it suffices to obtain the bound

$$\mathrm{KL}(p_{\sigma_T} \| \widehat{p}_T) \lesssim \tilde{\varepsilon}^2_{\mathrm{BD}} + \varepsilon^2\,.$$

To this end, by (27) and Theorem 3.9,

$$\mathrm{KL}(p_{\sigma_T} \| \widehat{p}_T) \lesssim \frac{R^2}{\sigma_0^2} + \tilde{\varepsilon}^2_{\mathrm{BD}} + \big(\frac{\mathsf{k}^3}{d} + \frac{\mathsf{k}^5}{d^2}\big)\,T + \mathsf{k}h\,T\,,$$

where we already used the choice $a_t = \frac{1}{2}\,\sigma_t^2$. Noting that $T \asymp \log(\sigma_0/\sigma_T)$, we now choose $\sigma_0 \asymp \mathsf{m}_2/\varepsilon$, $h \asymp \varepsilon^2/(\mathsf{k}\log(\mathsf{m}_2\sqrt{d}/\varepsilon^2))$ and obtain

$$\mathrm{KL}(p_{\sigma_T} \| \widehat{p}_T) \lesssim \varepsilon^2 + \tilde{\varepsilon}^2_{\mathrm{BD}} + \big(\frac{\mathsf{k}^3}{d} + \frac{\mathsf{k}^5}{d^2}\big)\log\frac{\mathsf{m}_2\sqrt{d}}{\varepsilon^2}\,.$$

The result concludes by invoking the remaining conditions.

E. On the choice of noise prior

Finally, we revisit the assumption **(A3)** on the "score estimation error". In standard DDPM theory, the bound on the score estimation error in L^2 which is needed for the discretization analysis can also be written as the *excess risk* of the population score matching loss. This leads to a remarkable harmony between statistical theory which can identify when the L^2 error is small, and the discretization analysis. In the context of the convenient noise sequence defined above, we prove the analogous result for BDDMs.

Theorem E.1. *Suppose that* $[\sigma_T, \sigma_0] \subseteq \operatorname{supp}\Pi_0$. *It holds that*

$$\varepsilon^2_{\mathrm{BD}} \le \big(\min_{\sigma_T \le \sigma \le \sigma_0} \mathsf{a}\,(1-\mathsf{a})\,\sigma^3\,\Pi_0(\sigma)\big)^{-1}\,\mathcal{E}(\widehat{f}_\theta)\,,$$

where $\mathcal{E}(\cdot)$ *denotes the population excess risk for* (5).

This result shows that if the neural network learns a good blind denoiser, in the sense of attaining a small excess risk, then the $\varepsilon_{\mathrm{BD}}$ quantity in our error analysis is controlled. Moreover, our analysis singles out a particularly good choice of noise prior. Namely, when $\Pi_0(\sigma) \propto \sigma^{-3}$ over $[\sigma_T, \sigma_0]$, then $\varepsilon^2_{\mathrm{BD}}$ and $\mathcal{E}(\widehat{f}_\theta)$ are in fact equal up to a constant, so that the training objective and the error propagation along the dynamics are well-aligned.[2] Although we state this for a particular family of noise schedules for concreteness, the same approach indeed identifies a well-aligned prior Π_0 for each schedule $(a_t)_{t\ge 0}$.

[2]Note that $\Pi_0(\sigma) \propto \sigma^{-3}$ corresponds to a uniform prior over λ.

E.1. Proof of Theorem E.1

By the definition of $\varepsilon_{\mathrm{BD}}^2$ and from (17),

$$\varepsilon_{\mathrm{BD}}^2 = \int_0^T \frac{1}{a_t} \|\widehat{f}_\theta - f^\star\|_{L^2(p_{\sigma_t})}^2 \,\mathrm{d}t\,, \qquad \mathcal{E}(\widehat{f}_\theta) = \int \|\widehat{f}_\theta - f^\star\|_{L^2(p_\sigma)}^2 \,\mathrm{d}\Pi_0(\sigma)\,.$$

Let us switch notation to $a(t) \coloneqq a_t$, $\sigma(t) \coloneqq \sigma_t$, and we perform the change of variables $\omega = \sigma(t)$ so that $\mathrm{d}\omega = \dot\sigma(t)\,\mathrm{d}t$. Then, for $\varepsilon(\sigma) \coloneqq \|\widehat{f}_\theta - f^\star\|_{L^2(p_\sigma)}$, the integral on the left can be written

$$\varepsilon_{\mathrm{BD}}^2 = \int_0^T \frac{\varepsilon(\sigma(t))^2}{a(t)} \frac{1}{\dot\sigma(t)} \,\dot\sigma(t)\,\mathrm{d}t = \int_{\sigma(0)}^{\sigma(T)} \frac{\varepsilon(\omega)^2}{a(\sigma^{-1}(\omega))\,\dot\sigma(\sigma^{-1}(\omega))}\,\mathrm{d}\omega\,.$$

Next, we specialize to the noise schedule with $a(t) = \mathsf{a}\sigma(t)^2$, $\dot\sigma(t) = -(1-\mathsf{a})\,\sigma(t)$. This yields

$$\begin{aligned}\varepsilon_{\mathrm{BD}}^2 &= -\int_{\sigma(0)}^{\sigma(T)} \frac{\varepsilon(\omega)^2}{\mathsf{a}\,\omega^2\,(1-\mathsf{a})\,\omega}\,\mathrm{d}\omega = \int_{\sigma(T)}^{\sigma(0)} \frac{1}{\mathsf{a}\,(1-\mathsf{a})\,\omega^3\,\Pi_0(\omega)}\,\varepsilon(\omega)^2\,\mathrm{d}\Pi_0(\omega)\\ &\le \Big(\max_{\omega\in[\sigma(T),\sigma(0)]} \frac{1}{\mathsf{a}\,(1-\mathsf{a})\,\omega^3\,\Pi_0(\omega)}\Big)\,\mathcal{E}(\widehat{f}_\theta)\,.\end{aligned}$$

When $\Pi_0(\omega) \propto \omega^{-3}$ on $[\sigma_T, \sigma_0]$, then $\varepsilon_{\mathrm{BD}}^2 \propto \mathcal{E}(\widehat{f}_\theta)$.

F. Technical results

F.1. Calculations based on Tweedie's formula

Lemma F.1. *The following hold:*

$$\omega^2\,\nabla \log p_\omega(y) = f_\omega^\star(y) - y\,, \tag{28}$$

$$\omega^2\,\nabla^2 \log p_\omega(y) = \omega^{-2}\,C_\omega(y) - I\,, \tag{29}$$

$$\omega^2\,\nabla \operatorname{tr} C_\omega(y) = W_\omega(y) - 2C_\omega(y)\,(f_\omega^\star(y) - y)\,. \tag{30}$$

Proof. We only prove the third equality, as it is the least standard. Also, we will temporarily drop the conditioning variable and write $\mathbb{E}_\omega$ as shorthand for $\mathbb{E}_{q_\omega(\cdot|y)}$.

We first notice that

$$\mathbb{E}_\omega\|X - y\|^2 = \operatorname{tr} C_\omega(y) + \|f_\omega^\star(y) - y\|^2\,.$$

Thus the expression for $W_\omega(y)$ simplifies to

$$W_\omega(y) = \mathbb{E}_\omega[(X - f_\omega^\star(y))\,\|X - f_\omega^\star(y)\|^2] + 2C_\omega(y)\,(f_\omega^\star(y) - y)\,. \tag{31}$$

We now want to show

$$\nabla \operatorname{tr} C_\omega(y) = \omega^{-2}\,\mathbb{E}_\omega[(X - f_\omega^\star(y))\,\|X - f_\omega^\star(y)\|^2]\,.$$

To this end, we compute (here note the gradients are with respect to y)

$$\begin{aligned}\nabla_y \operatorname{tr} C_\omega(y) &= \nabla_y \int \|x - f_\omega^\star(y)\|^2\,\mathrm{d}q_\omega(x|y)\\ &= \int 2\,\nabla_y f_\omega^\star(y)\,(x - f_\omega^\star(y))\,\mathrm{d}q_\omega(x|y)\\ &\qquad + \int \|x - f_\omega^\star(y)\|^2\,\big(\nabla_y \log q_\omega(x|y)\big)\,\mathrm{d}q_\omega(x|y)\\ &= \int \|x - f_\omega^\star(y)\|^2\,\big(\nabla_y \log q_\omega(x|y)\big)\,\mathrm{d}q_\omega(x|y)\,.\end{aligned}$$

And finally we complete the claim by computing the following:

$$\begin{aligned}\nabla_y \log q_\omega(x|y) &= \nabla_y \log p_X(x) - \frac{1}{2\omega^2}\nabla_y \|x-y\|^2 - \nabla \log Z_\omega(y)\\ &= 0 + \frac{1}{\omega^2}(x-y) - \nabla_y \log \int p_X(x)\exp\big(-\frac{1}{2\omega^2}\|x-y\|^2\big)\,\mathrm{d}x\\ &= \frac{1}{\omega^2}(x-y) - \int \frac{\omega^{-2}(x-y)\,p_X(x)\exp\big(-\frac{1}{2\omega^2}\|x-y\|^2\big)}{Z_\omega(y)}\,\mathrm{d}x\\ &= \frac{1}{\omega^2}(x-y-f^\star_\omega(y)+y) = \frac{1}{\omega^2}(x-f^\star_\omega(y)),\end{aligned}$$

where $Z_\omega(y)$ is the normalizing constant of $q_\omega(\cdot|y)$. □

F.2. Tools for proving the main results

Letting $N_X(r) \coloneqq N(\mathcal{X}; r, \|\cdot\|)$ be the covering number of $\mathcal{X}$ under the Euclidean norm, recall that

$$\mathsf{k} = 1 + \log N_X\big(\tfrac{\sigma_T^2}{\sigma_0\sqrt{d}}\big).$$

We use $q_\sigma(\cdot|y)$ to denote the posterior of $X \sim p_X$ given $Y \sim \mathcal{N}(X, \sigma^2 I)$, and

$$D_{\sigma,\ell}(\bar{x}, y) \coloneqq \int \|x_0 - \bar{x}\|^\ell \,\mathrm{d}q_\sigma(x_0|y).$$

Also, recall the definitions of $f^\star_\sigma$, C_σ, and W_σ from (21), (22), and (23) respectively.

The following proposition is adapted from Chen et al. (2026, Proposition E.7). To make the presentation more self-contained and because we have modified the statements, we provide a proof here.

Proposition F.2. *For $\tilde{\sigma} \geq 0$, we write $\mathbb{P}_{\tilde{\sigma}}(\cdot)$ to be the probability measure under which $\bar{x} \sim p_X$, $V \sim \mathcal{N}(0, \tilde{\sigma}^2 I)$. For any $\sigma, \tilde{\sigma} \in [\sigma_T, \sigma_0]$, $\ell \geq 1$, and $\delta \in (0,1)$, with probability at least $1-\delta$ under $\mathbb{P}_{\tilde{\sigma}}$,*

$$\begin{aligned}D_{\sigma,\ell}(\bar{x}, \bar{x}+V) &\lesssim_\ell \left\{(\sigma^2+\tilde{\sigma}^2)\,(\mathsf{k}+\log(1/\delta))\right\}^{\ell/2},\\ \sqrt{\langle V, C_\sigma(\bar{x}+V)\,V\rangle} &\lesssim (\sigma^2+\tilde{\sigma}^2)\,(\mathsf{k}+\log(1/\delta)),\end{aligned}$$

and

$$|D_{\sigma,2}(\bar{x}+V, \bar{x}+V) - \tilde{\sigma}^2 d| \lesssim \tilde{\sigma}^2\sqrt{d\log(1/\delta)} + (\sigma^2+\tilde{\sigma}^2)\,(\mathsf{k}+\log(1/\delta)). \tag{32}$$

Proof. Let $r \coloneqq \frac{\sigma_T^2}{\sigma_0\sqrt{d}}$, and fix an r-covering of $\mathcal{X}$, denoted $\mathcal{X} \subseteq \bigcup_{i=1}^N B_i$, where $N \coloneqq N_X(r)$ and $B_i \coloneqq B(z_i, r)$ for $i = 1, \ldots, N$. Additionally, let $\mathcal{I} \coloneqq \{i \in [N] : p_X(B_i) \geq \alpha\}$ for some $\alpha > 0$ to be chosen later, and let $\mathcal{X}_+ \coloneqq \bigcup_{i\in\mathcal{I}} B_i$.

Claim 1. Define $M_1 \coloneqq \tilde{\sigma}\,(\sqrt{d} + 2\sqrt{\log(2/\delta)})$, $M_2 \coloneqq \tilde{\sigma}\sqrt{2\log(4N/\delta)}$, and

$$\mathcal{V} \coloneqq \{V : \|V\| \leq M_1,\ |\langle V, x-x'\rangle| \leq M_2\,\|x-x'\| + 2r\,(M_1+M_2),\ \forall x, x' \in \overline{\mathcal{X}}\}.$$

Then for any $\tilde{\sigma} \geq 0$, it holds that $\mathbb{P}_{\tilde{\sigma}}(V \notin \mathcal{V}) \leq \delta$.

Proof of Claim 1. By Gaussian concentration, it is clear that for $t \geq 0$,

$$\mathbb{P}_{\tilde{\sigma}}(\|V\| \geq \tilde{\sigma}\,(\sqrt{d} + 2\sqrt{t})) \leq e^{-t}.$$

Further, for any $i, j \in [N]$, it holds that $\mathbb{P}_{\tilde{\sigma}}(|\langle V, z_i - z_j\rangle| \leq \tilde{\sigma}\,\|z_i - z_j\|\,\sqrt{2t}) \leq 2e^{-t}$ for $t \geq 0$. Hence, by a union bound, we can show that $\mathbb{P}_{\tilde{\sigma}}(V \notin \mathcal{V}_0) \leq \delta$, where

$$\mathcal{V}_0 = \{V : \|V\| \leq M_1,\ |\langle V, z_i - z_j\rangle| \leq M_2\,\|z_i - z_j\|,\ \forall i, j \in [N]\}.$$

Note that for any $V \in \mathcal{V}_0$, it is clear that for any $i, j \in [N]$, $x \in B_i$, $x' \in B_j$, we can bound

$$\begin{aligned} |\langle V, x - x' \rangle| &\le |\langle V, z_i - z_j \rangle| + 2r \|V\| \\ &\le M_2 \|z_i - z_j\| + 2rM_1 \\ &\le M_2 \|x - x'\| + 2r (M_1 + M_2) . \end{aligned}$$

This implies that $\mathcal{V}_0 \subseteq \mathcal{V}$, and the proof of Claim 1 is hence completed. □

Note that for any $z \in \mathbb{R}^d$ and $M > 0$, $\ell \ge 1$,

$$\begin{aligned} D_{\sigma,\ell}(\bar{x}, y) &\coloneqq \int \|x_0 - \bar{x}\|^\ell \, \mathrm{d}q_\sigma(x_0|y) = \frac{\int \|x_0 - \bar{x}\|^\ell \exp\big(-\frac{\|y - x_0\|^2}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)}{\int \exp\big(-\frac{\|y - x_0\|^2}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)} \\ &\lesssim_\ell M^\ell + \frac{\int (\|x_0 - \bar{x}\| - M)_+^\ell \exp\big(-\frac{\|y - x_0\|^2}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)}{\int \exp\big(-\frac{\|y - x_0\|^2}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)} . \end{aligned}$$

Here and throughout the proof, $\lesssim_\ell$ indicates that the bound holds up to a constant depending only on ℓ. Now specialize the above to $y = \bar{x} + V$ for $\bar{x} \in \mathcal{X}_+$ and $V \in \mathcal{V}$, and fix $M \ge 8 (M_2 + \sqrt{r (M_1 + M_2)})$. Together, we have that

$$\begin{aligned} &D_{\sigma,\ell}(\bar{x}, \bar{x} + V) \\ &\lesssim_\ell M^\ell + \frac{\int (\|x_0 - \bar{x}\| - M)_+^\ell \exp\big(-\frac{\|\bar{x} + V - x_0\|^2}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)}{\int \exp\big(-\frac{\|\bar{x} + V - x_0\|^2}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)} \\ &= M^\ell + \frac{\int (\|x_0 - \bar{x}\| - M)_+^\ell \exp\big(-\frac{\|\bar{x} - x_0\|^2 + 2\langle V, \bar{x} - x_0 \rangle}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)}{\int \exp\big(-\frac{\|\bar{x} - x_0\|^2 + 2\langle V, \bar{x} - x_0 \rangle}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0)} . \end{aligned}$$

Since $V \in \mathcal{V}$, for any $\bar{x}, x_0 \in \overline{\mathcal{X}}$ with $\|\bar{x} - x_0\| \ge M$, we can bound

$$|\langle V, \bar{x} - x_0 \rangle| \le M_2 \|\bar{x} - x_0\| + 2r (M_1 + M_2) \le \frac{1}{4} \|\bar{x} - x_0\|^2 ,$$

and hence in this case $\|\bar{x} - x_0\|^2 + 2\langle V, \bar{x} - x_0 \rangle \ge \frac{1}{2} \|\bar{x} - x_0\|^2$. This immediately implies that

$$\begin{aligned} &\int (\|x_0 - \bar{x}\| - M)_+^\ell \exp\big(-\frac{\|\bar{x} - x_0\|^2 + 2\langle V, \bar{x} - x_0 \rangle}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0) \\ &\quad \le \int (\|x_0 - \bar{x}\| - M)_+^\ell \exp\big(-\frac{\|\bar{x} - x_0\|^2}{4\sigma^2}\big) \, \mathrm{d}p_X(x_0) \\ &\quad \le M^\ell \exp\big(-\frac{M^2}{4\sigma^2}\big) , \end{aligned}$$

where we use the fact that $u \mapsto u^{\ell/2} e^{-u}$ is decreasing for $u \ge \ell/2$, and assuming further that $M^2 \ge 2\ell\sigma^2$.

On the other hand, suppose that $\bar{x} \in B_i$ such that $p_X(B_i) \ge \alpha$. Note that for any $x \in B_i$, we can bound

$$\begin{aligned} \|\bar{x} - x_0\|^2 + 2\langle V, \bar{x} - x_0 \rangle &\le 4r^2 + 4r \|V\| \le 4r^2 + 4\tilde{\sigma} r (\sqrt{d} + 2\sqrt{\log(2/\delta)}) \\ &\le 8\sigma^2 (1 + \sqrt{\log(2/\delta)}) , \end{aligned}$$

by the choice of r. This implies that

$$\int \exp\big(-\frac{\|\bar{x} - x_0\|^2 + 2\langle V, \bar{x} - x_0 \rangle}{2\sigma^2}\big) \, \mathrm{d}p_X(x_0) \ge e^{-4-4\sqrt{\log(2/\delta)}} \, p_X(B_i) \ge c_0 \alpha \delta ,$$

where $c_0 > 0$ is an absolute constant.

Combining the above bounds, we can conclude that as long as $\bar{x} \in \mathcal{X}_+$ and $V \in \mathcal{V}$, $M = 8\,(M_2 + \sqrt{r\,(M_1 + M_2)}) + \sqrt{2\ell}\sigma + 2\sigma\sqrt{\log(1/(c_0\alpha\delta))}$, it holds that

$$D_{\sigma,\ell}(\bar{x}, \bar{x} + V) \lesssim_\ell M^\ell + \frac{M^\ell}{c_0\alpha\delta} \exp\big(-\frac{M^2}{4\sigma^2}\big) \lesssim M^\ell\,.$$

Note that $\mathbb{P}_{\bar{x}\sim p_X}(\bar{x} \notin \mathcal{X}_+) \le \sum_{i\notin\mathcal{I}} p_X(B_i) \le N\alpha = \delta$, where we choose $\alpha = \delta/N$. Hence, with probability at least $1-\delta$ under $\mathbb{P}_{\tilde{\sigma}}$,

$$D_{\sigma,\ell}(\bar{x}, \bar{x} + V) \lesssim_\ell M^\ell\,.$$

Similarly, we note that for $V \in \mathcal{V}$,

$$\begin{aligned}\langle V, C_\sigma(\bar{x} + V)\, V\rangle &\le \int \langle V, x_0 - \bar{x}\rangle^2 \,\mathrm{d}q_\sigma(x_0|\bar{x} + V)\\ &\le \int (M_2\,\|x_0 - \bar{x}\| + 2r\,(M_1 + M_2))^2 \,\mathrm{d}q_\sigma(x_0|\bar{x} + V)\\ &\le 2M_2^2\, D_{\sigma,2}(\bar{x}, \bar{x} + V) + 8r^2\,(M_1 + M_2)^2\,.\end{aligned}$$

Our argument above then shows that with probability at least $1-\delta$ under $\mathbb{P}_{\tilde{\sigma}}$,

$$\langle V, C_\sigma(\bar{x} + V)\, V\rangle \lesssim M^4\,.$$

For the final claim, note that

$$\begin{aligned}D_{\sigma,2}(y, y) &= \int \|y - x_0\|^2 \,\mathrm{d}q_\sigma(x_0|y)\\ &= \|y - \bar{x}\|^2 + 2\left\langle y - \bar{x}, \int (\bar{x} - x_0)\,\mathrm{d}q_\sigma(x_0|y)\right\rangle + D_{\sigma,2}(\bar{x}, y)\,.\end{aligned}$$

Setting $y = \bar{x} + V$,

$$D_{\sigma,2}(\bar{x} + V, \bar{x} + V) = \|V\|^2 - 2\int \langle V, \bar{x} - x_0\rangle \,\mathrm{d}q_\sigma(x_0|\bar{x} + V) + D_{\sigma,2}(\bar{x}, \bar{x} + V)\,.$$

This readily yields, for $V \in \mathcal{V}$,

$$\begin{aligned}\big|D_{\sigma,2}(\bar{x} + V, \bar{x} + V) - \|V\|^2\big| &\lesssim M_2 \int \|\bar{x} - x_0\| \,\mathrm{d}q_\sigma(x_0|\bar{x} + V) + r\,(M_1 + M_2) + M^2\\ &= M_2\, D_{\sigma,1}(\bar{x}, \bar{x} + V) + r\,(M_1 + M_2) + M^2 \lesssim M^2\,.\end{aligned}$$

Together with Gaussian concentration, the claim is complete. □

As a corollary, we obtain the following inequalities.

Corollary F.3. *If $X_t \sim p_{\sigma_t}$ and $\sigma \in [\sigma_T, \sigma_0]$, then with probability at least $1-\delta$,*

$$\begin{aligned}\operatorname{tr} C_\sigma(X_t) &\lesssim (\sigma_t^2 + \sigma^2)\,(\mathsf{k} + \log(1/\delta))\,,\\ \|W_\sigma(X_t)\| &\lesssim \big[(\sigma_t^2 + \sigma^2)\,(\mathsf{k} + \log(1/\delta))\big]^{3/2}\,.\end{aligned}$$

Proof. We apply Proposition F.2 with $\tilde{\sigma} = \sigma_t$, $\bar{x} = X_0$, and $V = X_t - X_0$.

For the first statement, we note that

$$\begin{aligned}\operatorname{tr} C_\sigma(X_t) &= \int \|x_0 - f_\sigma^\star(X_t)\|^2 \,\mathrm{d}q_\sigma(x_0|X_t)\\ &\le \int \|x_0 - X_0\|^2 \,\mathrm{d}q_\sigma(x_0|X_t) = D_{\sigma,2}(X_0, X_0 + V)\end{aligned}$$

so the result follows from Proposition F.2.

For the second statement, recall from (31) that

$$W_\sigma(X_t) = \int (x_0 - f_\sigma^\star(X_t)) \, \|x_0 - f_\sigma^\star(X_t)\|^2 \, \mathrm{d}q_\sigma(x_0|X_t) + 2C_\sigma(X_t) \, (f_\sigma^\star(X_t) - X_t) \, .$$

Hence,

$$\begin{aligned}
\|W_\sigma(X_t)\| &\leq \int \|x_0 - f_\sigma^\star(X_t)\|^3 \, \mathrm{d}q_\sigma(x_0|X_t) + 2 \, \|C_\sigma(X_t) \, (f_\sigma^\star(X_t) - X_t)\| \\
&\leq 8D_{\sigma,3}(X_0, X_0 + V) + 2 \operatorname{tr} C_\sigma(X_t) \, \|f_\sigma^\star(X_t) - X_0\| + 2 \, \|C_\sigma(X_t) \, V\| \\
&\leq 8D_{\sigma,3}(X_0, X_0 + V) + 2 \operatorname{tr} C_\sigma(X_t) \, D_{\sigma,1}(X_0, X_0 + V) \\
&\qquad + 2\sqrt{\operatorname{tr} C_\sigma(X_t) \, \langle V, C_\sigma(X_t) \, V \rangle} \\
&\lesssim \left[(\sigma_t^2 + \sigma^2) \, (\mathsf{k} + \log(1/\delta)) \right]^{3/2} .
\end{aligned}$$

□

G. Experimental details

G.1. Analytical blind denoiser

In this section, we detail the results shown in Section 3.2. Additionally, we present more examples of sampling from arbitrary mixture of high dimensional Gaussian embedded in low dimensions.

G.1.1. Optimal denoiser for Gaussian Mixture

Take the true distribution to be a mixture of Gaussians

$$p(x) = \frac{1}{K} \sum_{i=1}^{K} \mathcal{N}(x; m_i, \Sigma_0)$$

If we add centered Gaussian noise with variance σ^2 to the data ($x_\sigma = x + \sigma z$), the distribution of noisy images will be

$$p(x_\sigma) = \frac{1}{K} \sum_{i=1}^{K} \mathcal{N}(x; m_i, \Sigma_0 + \sigma^2 I)$$

The score of this density is

$$\nabla \log p(x_\sigma) = -(\Sigma_0 + \sigma^2 I)^{-1} \sum_{i=1}^{K} \omega_i(x_\sigma)(x_\sigma - m_i) \tag{33}$$

where

$$\omega_i(x_\sigma) = \frac{\mathcal{N}(x_\sigma; m_i, \Sigma_0 + \sigma^2 I)}{\sum_{j=1}^{K} \mathcal{N}(x_\sigma; m_j, \Sigma_0 + \sigma^2 I)}$$

So, the optimal denoiser for a noisy observation can be written in closed-form as

$$f^\star(x_\sigma, \sigma, \{m_i\}_{i=1}^K, \Sigma_0) = x_\sigma - \sigma^2 \, (\Sigma_0 + \sigma^2 I)^{-1} \sum_{i=1}^{K} \omega_i(x_\sigma) \, (x_\sigma - m_i) \, . \tag{34}$$

To simulate the blind denoiser, we estimate the noise variance from a single noisy data point by maximizing the likelihood of noisy observation across $\log p_\sigma$. Thus, for a given noisy observation y, we compute $\hat{\sigma}$ such that

$$\hat{\sigma} = \arg\max_{\sigma} \log p_\sigma(y) \, .$$

The *blind* denoiser under maximum likelihood estimation (MLE) of the noise variance is therefore

$$f^\dagger(x_\sigma, \hat{\sigma}, \{m_i\}_{i=1}^K, \Sigma_0) = x_\sigma - \hat{\sigma}^2 \, (\Sigma_0 + \hat{\sigma}^2 I)^{-1} \sum_{i=1}^{K} \omega_i(x_\sigma) \, (x_\sigma - m_i) \, . \tag{35}$$

Algorithm 3 Sampling using an optimal blind denoiser with MLE noise variance

Input: Optimal denoiser $f^\dagger$, stepsize $h > 0$, diffusion coefficients $(a_t)_{t\in[0,T]}$, model parameters $\{m_i\}_{i=1}^K, \Sigma_0$, and $\sigma_0, \sigma_T > 0$
Initialize $X_0 \sim \mathcal{N}(0, \sigma_0^2 I)$, $k = 0$
while `keepgoing` == `True` **do**
  Compute $\hat{\sigma}_k = \text{argmax}_\sigma \log p_\sigma(X_{kh})$
  Compute $s_k = f^\dagger(X_{kh}, \hat{\sigma}_k, \{m_i\}_{i=1}^K, \Sigma_0) - X_{kh}$
  if $\hat{\sigma}_k \leq \sigma_T$ **then**
    `keepgoing` $\leftarrow$ `False`
  else
    Update $X_{(k+1)h} \leftarrow X_{kh} + hs_k + \sqrt{2}\int_{kh}^{(k+1)h} a_t \, \mathrm{d}B_t$
  end if
  Update $k \leftarrow k + 1$
end while

MORE EXAMPLES

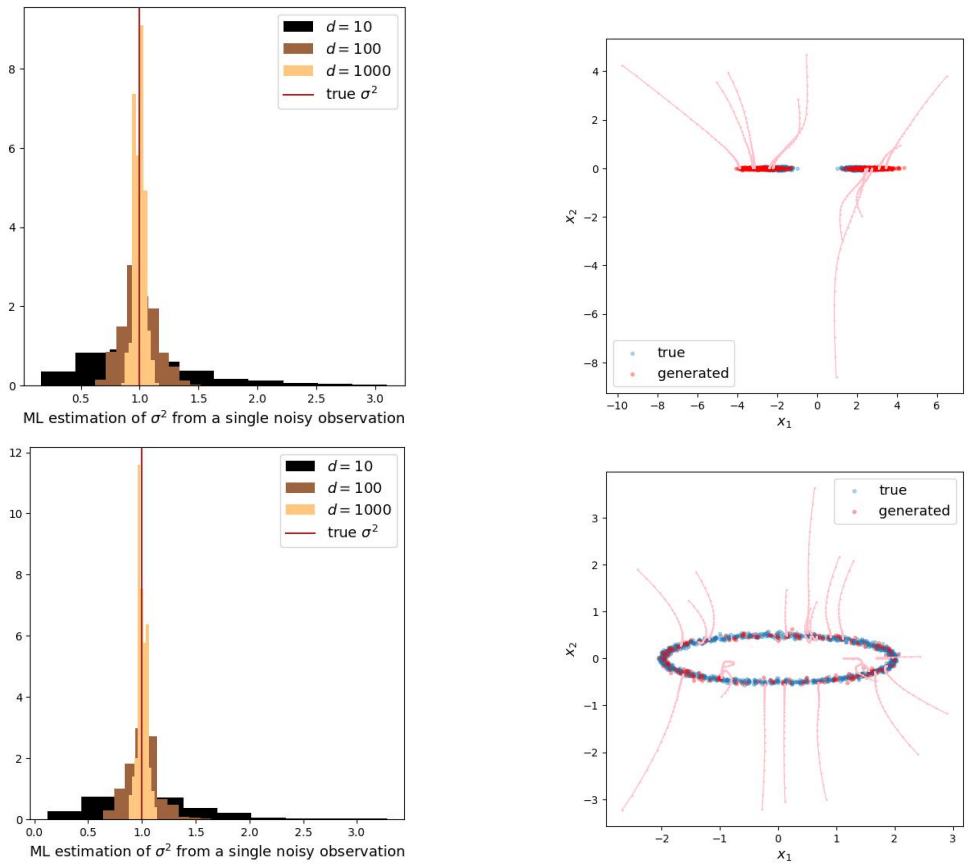

Figure 6. **Top:** Mixture of two Gaussians with covariances of rank 1. Sampling is shown for $d = 1000$. **Bottom:** Ellipse constructed as a mixture of Gaussians. Sampling is shown for $d = 100$.

G.2. Blind denoisers trained on Gaussian and Gaussian mixture data

We present two sets of experiments on toy data that elucidate the impact of intrinsic dimensionality and blind denoising. In both cases, we use a 6-layer MLP with width 512 and ReLU activations. We use the default AdamW optimizer in PyTorch, with learning rate equal to 10^{-3}, and train for 50k steps with batch size equal to 512. We take $\sigma_T = 0.01$ and $\sigma_0 = 10$. Additionally, Π_0 was chosen to be a log-uniform prior on $[\sigma_T, \sigma_0]$.

GAUSSIAN DATA

We randomly generate a $\mathsf{k} \times \mathsf{k}$ covariance Σ matrix, with $\mathsf{k} = 2$. We generate $n = 5000$ points and train two blind denoisers: $\widehat{f}_\theta^{(2)}$ and $\widehat{f}_\theta^{(100)}$. In the first scenario, the input dimension is the same as the intrinsic dimension of the data. In the latter, we pad the data with zeros until the data dimension is $d = 100$, thus changing the input dimension of the network.

The corresponding Gaussian-data samples and noise-evolution plot appear in Appendix G.2, Figure 7. Under an appropriate choice $a_t = \frac{1}{2}\sigma_t^2$ and $h = 0.5$, we see that $\hat{\sigma}_k \simeq \sigma_t$ as predicted by Proposition 3.2 when having low intrinsic dimensionality.

Figure 7 displays the output from Algorithm 2 using an exponential Euler integrator with $h = 0.5$ and $a_t = 0.5\sigma_t^2$. We see that the generated samples for $\mathsf{k} = d = 2$ are heavily concentrated and do not accurately capture the training data. Moreover, the noise schedule is both too fast at the beginning and too slow at inference. In contrast, the blind denoiser trained with larger input dimension manages to both (i) generate viable samples, and (ii) obey the noise schedule from Proposition 3.2 exactly.

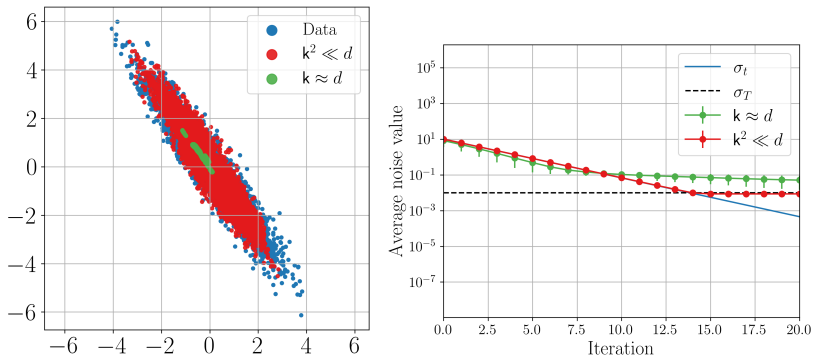

Figure 7. Sampling performance of BDDMs trained on Gaussian data with intrinsic dimension $\mathsf{k} = 2$ and two different input dimensions $d \in \{2, 100\}$. Generated samples **(left)** and evolution of the estimated noise level, corresponding to Proposition 3.2 **(right)**.

GAUSSIAN-MIXTURE DATA

We repeat the same experiment on 2-component Gaussian mixture data using randomly generated means and covariances with equal weight with $\mathsf{k} = 2$. We use $n = 50000$ data points and additionally incorporated weight-decay in the AdamW optimizer of value 5×10^{-2}. All other parameters remain the same, except now $d \in \{2, 200\}$. The behavior is essentially identical to Figure 7, where low intrinsic and high ambient dimensionalities are crucial for blind denoisers to learn and sample from the data distribution.

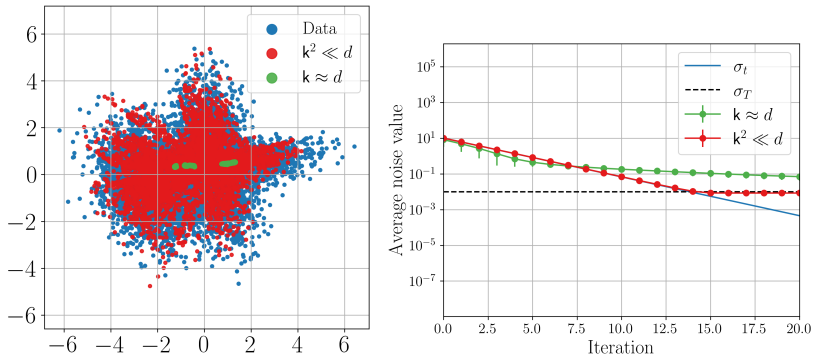

Figure 8. Sampling performance of BDDMs trained on a mixture of Gaussian dataset with intrinsic dimension $\mathsf{k} = 2$ and two different input dimensions $d \in \{2, 200\}$. Generated samples **(left)** and evolution of the estimated noise level, corresponding to Proposition 3.2 **(right)**.

G.3. Photographic images

DATASETS

Under the manifold hypothesis, natural images concentrate near a union of low-dimensional manifolds (Brown et al., 2023). This implies that natural images have low intrinsic dimensionality, making them a suitable testbed for our theoretical results. We use two popular image datasets, CelebA (Liu et al., 2015) and LSUN (bedroom class) (Yu et al., 2015). These datasets are complex enough to capture real-world structure while remaining simple enough to train smaller-size networks without text conditioning. All images are downsampled to 80×80 resolution.

ARCHITECTURE

The U-Net is the most widely used architecture for denoising and score estimation. We adopt the vanilla U-Net (Ronneberger et al., 2015), composed of convolutional layers with ReLU nonlinearities and layer normalization, organized into downsampling and upsampling blocks. In more detail, there are 3 downsampling and 3 upsampling blocks with increasing number of channels [64, 128, 256]. The middle block has 512 channels. There are also 3 skip connections that carry details from each encoder to the corresponding decoder at the same scale. The number of layers in the encoding blocks is 2 and decoding blocks is 6.

For non-blind models, we use the standard noise-level embedding: the noise level is mapped through a Fourier feature transform, processed by an MLP, and injected into all normalization layers via scale and shift parameters.

Our goal is to match architectural and training hyperparameters as closely as possible (except for noise conditioning versus blindness) to enable fair comparisons, so all models are trained from scratch. The architecture is intentionally small (**around 13 million parameters**) and simple (e.g., no attention or text conditioning) compared to commonly used models with hundreds of millions of parameters; the goal is not SOTA performance but a controlled comparison of blind versus non-blind models.

TRAINING

Blind models are trained by minimizing the empirical loss in (5), following Algorithm 1, while non-blind models are trained using the noise-conditioned counterpart. The noise level $\sigma \in (0, 3]$ is sampled according to $1/\sqrt{\sigma}$. We use a batch size of 512 for all models. The learning rate is 0.001 with a rate of decay of 100. We fix the total number of epochs (i.e., full pass over the dataset) to 1000. Training is done on four H100 NVIDIA GPUs per model, taking around 24 hours for CelebA and 34 hours for Bedroom datasets.

We observe diminishing mean-square error improvements for widening range of σ when the model is trained to predict the clean image; this generalization beyond the training noise range does not hold for models trained to predict the residual directly. This suggests that neural networks more readily represent constants (e.g., zero or the data mean) than the identity mapping.

NOISE LEVEL ESTIMATOR MODEL

We train a small network to estimate the noise level from a given noisy image. Given the simplicity of this task, our network consists of 4 convolutional layers with filter size of 3×3, and the number of channels in the intermediate layers is 64. To match the denoisers, we choose $\sigma \in (0, 3]$. The trained model can estimate the noise level almost perfectly, as seen in Figure 9.

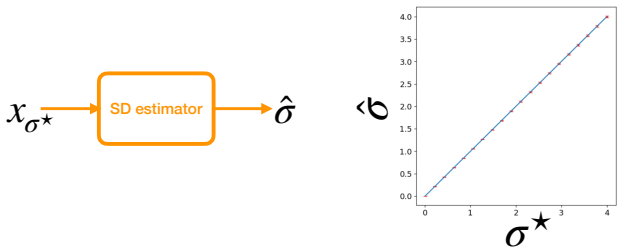

Figure 9. A smaller neural network trained to estimate σ from noisy image is very accurate. Error bars show two standard deviations. It is also worth noting that the model performs well beyond training range $\sigma \in (0, 3]$.

MORE SAMPLING EXAMPLES

Figure 10. Continued from Figure 4 with **lower** number of steps: $N \approx 100$ for BDDM shown in second row and $N = 100$ for VE-DDPM shown in third row. The samples from BDDM are still of higher quality than samples from DDPM. Samples in each column are initialized with the same seed.

Figure 11. Continued from Figure 4 with **higher** number of steps: $N \approx 17,000$ for BDDM shown in second row and $N = 17,000$ for VE-DDPM shown in third row. The samples from BDDM are still of significantly higher quality than samples from DDPM. Samples in each column are initialized with the same seed.

Figure 12. **Top row:** example training images from the LSUN dataset. **Second row:** samples generated by BDDM with average number of steps $N \approx 1300$. **Third row:** Samples generated by DDPM (VE) algorithm, with total number of steps $N = 1300$. Samples in each column are initialized with the same random seed with matched injected noise.