Interpreting and Improving Diffusion Models from an Optimization Perspective

Frank Permenter [*1] **Chenyang Yuan** [*1]

Abstract

Denoising is intuitively related to projection. Indeed, under the manifold hypothesis, adding random noise is approximately equivalent to orthogonal perturbation. Hence, learning to denoise is approximately learning to project. In this paper, we use this observation to interpret denoising diffusion models as approximate gradient descent applied to the Euclidean distance function. We then provide straight-forward convergence analysis of the DDIM sampler under simple assumptions on the projection error of the denoiser. Finally, we propose a new gradient-estimation sampler, generalizing DDIM using insights from our theoretical results. In as few as 5-10 function evaluations, our sampler achieves state-of-the-art FID scores on pretrained CIFAR-10 and CelebA models and can generate high quality samples on latent diffusion models.

1. Introduction

Diffusion models achieve state-of-the-art quality on many image generation tasks (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022). They are also successful in text-to-3D generation (Poole et al., 2022) and novel view synthesis (Liu et al., 2023). Outside the image domain, they have been used for robot path-planning (Chi et al., 2023), prompt-guided human animation (Tevet et al., 2022), and text-to-audio generation (Kong et al., 2020).

Diffusion models are presented as the reversal of a stochastic process that corrupts clean data with increasing levels of random noise (Sohl-Dickstein et al., 2015; Ho et al., 2020). This reverse process can also be interpreted as likelihood maximization of a noise-perturbed data distribution using learned gradients (called *score functions*) (Song & Ermon, 2019; Song et al., 2020b). While these interpretations are inherently probabilistic, samplers widely used in practice (e.g. (Song et al., 2020a)) are often deterministic, suggesting diffusion can be understood using a purely deterministic analysis. In this paper, we provide such analysis by interpreting denoising as approximate *projection*, and diffusion as *distance minimization* with gradient descent, using the denoiser output as an estimate of the gradient. This in turn leads to novel convergence results, algorithmic extensions, and paths towards new generalizations.

Denoising approximates projection The core object in diffusion is a *learned denoiser* $\epsilon_\theta(x,\sigma)$, which, when given a noisy point $x \in \mathbb{R}^n$ with noise level $\sigma > 0$, predicts the *noise direction* in x, i.e., it estimates ϵ satisfying $x = x_0 + \sigma\epsilon$ for a clean datapoint x_0.

Prior work (Rick Chang et al., 2017) interprets denoising as approximate *projection* onto the data manifold $\mathcal{K} \subseteq \mathbb{R}^n$. Our first contribution makes this interpretation rigorous by introducing a relative-error model, which states that $x - \sigma\epsilon_\theta(x,\sigma)$ well-approximates the projection of x onto $\mathcal{K}$ when $\sqrt{n}\sigma$ well-estimates the distance of x to $\mathcal{K}$. Specifically, we will assume that

$$\|x - \sigma\epsilon_\theta(x,\sigma) - \mathrm{proj}_{\mathcal{K}}(x)\| \leq \eta \mathrm{dist}_{\mathcal{K}}(x) \qquad (1)$$

when (x,σ) satisfies $\frac{1}{\nu}\mathrm{dist}_{\mathcal{K}}(x) \leq \sqrt{n}\sigma \leq \nu \mathrm{dist}_{\mathcal{K}}(x)$ for constants $1 > \eta \geq 0$ and $\nu \geq 1$.

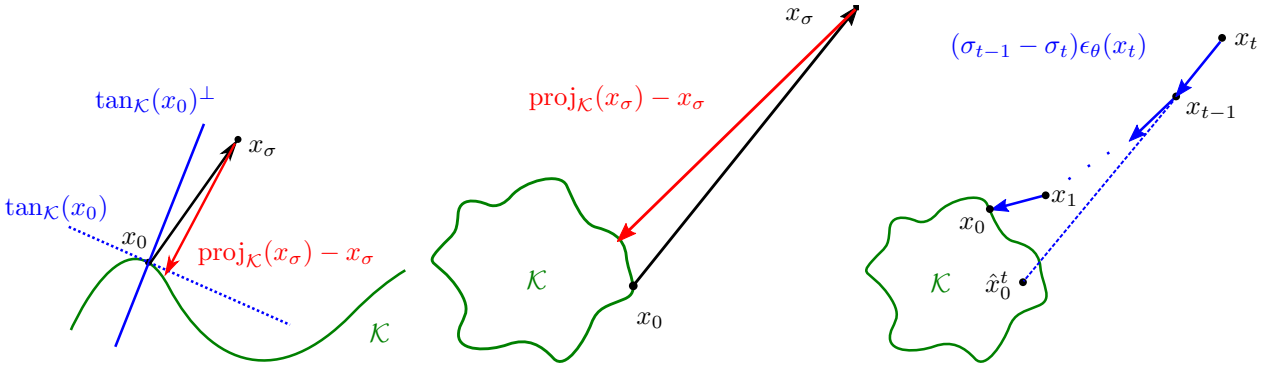
This error model is motivated by the following theoretical observations that hold when $\sigma \approx \mathrm{dist}_{\mathcal{K}}(x)/\sqrt{n}$:

1. When σ is small and the manifold hypothesis holds, denoising approximates projection given that most of the added noise is orthogonal to the data manifold; see Figure 1a and Proposition 3.1.
2. When σ is large, then any denoiser predicting any weighted mean of the data $\mathcal{K}$ has small relative error; see Figure 1b and Proposition 3.2.
3. Denoising with the *ideal denoiser* is a σ-smoothing of $\mathrm{proj}_{\mathcal{K}}(x)$ with relative error that can be controlled under mild assumptions; see Section 3.3.

We also empirically validate this error model on sequences

[*]Equal contribution [1]Toyota Research Institute, Cambridge, Massachusetts, USA. Correspondence to: Chenyang Yuan <chenyang.yuan@tri.global>, Frank Permenter <frank.permenter@tri.global>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



(a) Low noise (b) High noise (c) Denoising process (DDIM)

Figure 1: Denoising approximates projection: When σ is small (1a), most of the added noise lies in $\tan_{\mathcal{K}}(x_0)^{\perp}$ with high probability under the manifold hypothesis. When σ is large (1b), both denoising and projection point in the same direction towards $\mathcal{K}$. We interpret the denoising process (1c) as minimizing $\text{dist}^2_{\mathcal{K}}(x)$ by iteratively taking gradient steps, estimating the direction of $\nabla \frac{1}{2}\text{dist}^2_{\mathcal{K}}(x) = x_t - \text{proj}_{\mathcal{K}}(x_t)$ with $\epsilon_\theta(x_t)$.

(x_i, σ_i) generated by pretrained diffusion samplers, showing that it holds in practice on image datasets.

Diffusion as distance minimization Our second contribution analyzes diffusion sampling under the error model (1). In particular, we show it is equivalent to approximate gradient descent on the *squared* Euclidean distance function $f(x) := \frac{1}{2}\text{dist}^2_{\mathcal{K}}(x)$, which satisfies $\nabla f(x) = x - \text{proj}_{\mathcal{K}}(x)$. Indeed, in this notation, (1) is equivalent to

$$\|\nabla f(x) - \sigma\epsilon_\theta(x, \sigma)\| \leq \eta\|\nabla f(x)\|,$$

a standard relative-error assumption used in gradient-descent analysis. We also show how the error parameters (η, ν) controls the schedule of noise levels σ_t used in diffusion sampling. Theorem 4.2 shows that with bounded error parameters, a geometric σ_t schedule guarantees decrease of $\text{dist}_{\mathcal{K}}(x_t)$ in the sampling process. Finally, we leverage properties of the distance function to design a sampler that aggregates previous denoiser outputs to reduce gradient-estimation error (Section 5).

We conclude with computational evaluation of our sampler (Section 6) that demonstrates state-of-the-art FID scores on pretrained CIFAR-10 and CelebA datasets and comparable results to the best samplers for high-resolution latent models such as Stable Diffusion (Rombach et al., 2022) (Figure 3). Section 7 provides novel interpretations of existing techniques under the framework of distance functions and outlines directions for future research.

2. Background

Denoising diffusion models (along with all other generative models) treat datasets as samples from a probability distribution D supported on a subset $\mathcal{K}$ of $\mathbb{R}^n$. They are used to *generate* new points in $\mathcal{K}$ outside the training set. We overview their basic features. We then state properties of the Euclidean distance function $\text{dist}_{\mathcal{K}}(x)$ that are key to our contributions.

2.1. Denoising Diffusion Models

Denoisers Denoising diffusion models are trained to estimate a *noise vector* $\epsilon \in \mathbb{R}^n$ from a given noise level $\sigma > 0$ and noisy input $x_\sigma \in \mathbb{R}^n$ such that $x_\sigma = x_0 + \sigma\epsilon$ approximately holds for some x_0 in the data manifold $\mathcal{K}$. The learned function, denoted $\epsilon_\theta : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^n$, is called a *denoiser*. The trainable parameters, denoted jointly by $\theta \in \mathbb{R}^m$, are found by (approximately) minimizing the following loss function using stochastic gradient descent:

$$L(\theta) := \mathbb{E}\ \|\epsilon_\theta(x_0 + \sigma_t\epsilon, \sigma_t) - \epsilon\|^2 \tag{2}$$

where the expectation is taken over $x_0 \sim D$, $t \sim [N]$, and $\epsilon \sim \mathcal{N}(0, I)$. Given noisy x_σ and noise level σ, the denoiser $\epsilon_\theta(x_\sigma, \sigma)$ induces an estimate of $\hat{x}_0 \approx x_0$ via

$$\hat{x}_0(x_\sigma, \sigma) := x_\sigma - \sigma\epsilon_\theta(x_\sigma, \sigma). \tag{3}$$

Ideal Denoiser The *ideal denoiser* $\epsilon^*(x_\sigma, \sigma)$ for a particular noise level σ and data distribution D is the minimizer of the loss function

$$\mathcal{L}(\epsilon^*) = \mathop{\mathbb{E}}_{x_0 \sim D}\ \mathop{\mathbb{E}}_{x_\sigma \sim \mathcal{N}(x_0, \sigma I)} \|(x_\sigma - x_0)/\sigma - \epsilon^*(x_\sigma, \sigma)\|^2.$$

Informally, $\hat{x}_0$ predicted by ϵ^* is an estimate of the expected value of x_0 given x_σ. If D is supported on a set $\mathcal{K}$, then $\hat{x}_0$ lies in the *convex hull* of $\mathcal{K}$.

Sampling Aiming to improve accuracy, sampling algorithms construct a sequence $\hat{x}_0^t := \hat{x}_0(x_t, \sigma_t)$ of estimates

from a sequence of points x_t initialized at a given x_N. Diffusion samplers iteratively construct x_{t-1} from x_t and $\epsilon_\theta(x_t, \sigma_t)$, with a monotonically decreasing σ *schedule* $\{\sigma_t\}_{t=N}^{0}$. For simplicity of notation we use $\epsilon_\theta(\cdot, \sigma_t)$ and $\epsilon_\theta(\cdot, t)$ interchangeably based on context.

For instance, the randomized DDPM (Ho et al., 2020) sampler uses the update

$$x_{t-1} = x_t + (\sigma_{t'} - \sigma_t)\epsilon_\theta(x_t, \sigma_t) + \eta w_t, \tag{4}$$

where $w_t \sim \mathcal{N}(0, I)$, $\sigma_{t'} = \sigma_{t-1}^2/\sigma_t$ and $\eta = \sqrt{\sigma_{t-1}^2 - \sigma_{t'}^2}$ (we have $\sigma_{t'} < \sigma_{t-1} < \sigma_t$, as σ_{t-1} is the geometric mean of $\sigma_{t'}$ and σ_t). The deterministic DDIM (Song et al., 2020a) sampler, on the other hand, uses the update

$$x_{t-1} = x_t + (\sigma_{t-1} - \sigma_t)\epsilon_\theta(x_t, \sigma_t). \tag{5}$$

See Figure 1c for an illustration of this denoising process. Note that these samplers were originally presented in variables z_t satisfying $z_t = \sqrt{\alpha_t} x_t$, where α_t satisfies $\sigma_t^2 = \frac{1-\alpha_t}{\alpha_t}$. We prove equivalence of the original definitions to (4) and (5) in Appendix A and note that the change-of-variables from z_t to x_t previously appears in (Song et al., 2020b; Karras et al., 2022; Song et al., 2020a).

2.2. Distance and Projection

The *distance function* to a set $\mathcal{K} \subseteq \mathbb{R}^n$ is defined as

$$\text{dist}_\mathcal{K}(x) := \inf\{\|x - x_0\| : x_0 \in \mathcal{K}\}. \tag{6}$$

The *projection* of $x \in \mathbb{R}^n$, denoted $\text{proj}_\mathcal{K}(x)$, is the set of points that attain this distance:

$$\text{proj}_\mathcal{K}(x) := \{x_0 \in \mathcal{K} : \text{dist}_\mathcal{K}(x) = \|x - x_0\|\}. \tag{7}$$

When $\text{proj}_\mathcal{K}(x)$ is unique, i.e., when $\text{proj}_\mathcal{K}(x) = \{x_0\}$, we abuse notation and let $\text{proj}_\mathcal{K}(x)$ denote x_0. Then $x - \text{proj}_\mathcal{K}(x)$ is the direction of steepest descent of $\text{dist}_\mathcal{K}(x)$:

Proposition 2.1 (page 283, Theorem 3.3 of (Delfour & Zolésio, 2011)). *Suppose $\mathcal{K} \subseteq \mathbb{R}^n$ is closed and $x \notin \mathcal{K}$. Then $\text{proj}_\mathcal{K}(x)$ is unique for almost all $x \in \mathbb{R}^n$ (under the Lebesgue measure). If $\text{proj}_\mathcal{K}(x)$ is unique, then $\nabla \text{dist}_\mathcal{K}(x)$ exists, $\|\nabla \text{dist}_\mathcal{K}(x)\| = 1$ and*

$$\nabla \tfrac{1}{2}\text{dist}_\mathcal{K}(x)^2 = \text{dist}_\mathcal{K}(x),$$
$$\nabla \text{dist}_\mathcal{K}(x) = x - \text{proj}_\mathcal{K}(x).$$

In addition, we define a *smoothed squared-distance function* for a smoothing parameter $\sigma > 0$ by using the $\text{softmin}_{\sigma^2}$ operator instead of min.

$$\text{dist}^2_\mathcal{K}(x, \sigma) := \underset{x_0 \in \mathcal{K}}{\text{softmin}_{\sigma^2}} \|x_0 - x\|^2$$
$$= -\sigma^2 \log\left(\textstyle\sum_{x_0 \in \mathcal{K}} \exp\left(-\frac{\|x_0 - x\|^2}{2\sigma^2}\right)\right).$$

In contrast to $\text{dist}^2_\mathcal{K}(x)$, $\text{dist}^2_\mathcal{K}(x, \sigma)$ is always differentiable and lower bounds $\text{dist}^2_\mathcal{K}(x)$.

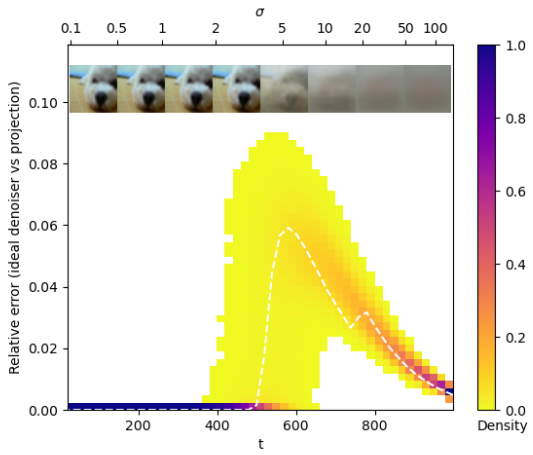


Figure 2: Ideal denoiser well-approximates projection onto the CIFAR-10 dataset. Dashed line plots error for the example shown, and density plot shows the error distribution over 10k different DDIM sampling trajectories.

3. Denoising as Approximate Projection

In this section, we provide theoretical and empirical justifications for our relative error model, formally stated below:

Definition 3.1. *We say $\epsilon_\theta(x, \sigma)$ is an (η, ν)-*approximate *projection if there exists constants $1 > \eta \geq 0$ and $\nu \geq 1$ so that for all x with unique $\text{proj}_\mathcal{K}(x)$ and for all σ satisfying $\frac{1}{\nu}\text{dist}_\mathcal{K}(x) \leq \sqrt{n}\sigma \leq \nu \text{dist}_\mathcal{K}(x)$, we have*

$$\|x - \sigma\epsilon_\theta(x, \sigma) - \text{proj}_\mathcal{K}(x)\| \leq \eta \text{dist}_\mathcal{K}(x).$$

To justify this model, we will prove relative error bounds under different assumptions on ϵ_θ, (x, σ) and $\mathcal{K}$. Analysis of DDIM based on this model is given in Section 4. Formal statements and proofs are deferred to Appendix B. Appendix E contains further experiments verifying our error model on image datasets.

3.1. Relative Error Under the Manifold Hypothesis

The *manifold hypothesis* (Bengio et al., 2013; Fefferman et al., 2016; Pope et al., 2021) asserts that "real-world" datasets are (approximately) contained in low-dimensional manifolds of $\mathbb{R}^n$. Specifically, we suppose that $\mathcal{K}$ is a manifold of dimension d with $d \ll n$. We next show that denoising is approximately equivalent to projection, when noise is small compared to the *reach* of $\mathcal{K}$, defined as the largest $\tau > 0$ such that $\text{proj}_\mathcal{K}(x)$ is unique when $\text{dist}_\mathcal{K}(x) < \tau$.

The following classical result tells us that for small perturbations w, the difference between $\text{proj}_\mathcal{K}(x_0 + w)$ and x_0 is contained in the *tangent space* $\tan_\mathcal{K}(x_0)$, a subspace of dimension d associated with each $x_0 \in \mathcal{K}$.

Lemma 3.1 (Theorem 4.8(12) in (Federer, 1959))**.** *Consider $x_0 \in \mathcal{K}$ and $w \in \tan_{\mathcal{K}}(x_0)^{\perp}$. If $\|w\| < \text{reach}(\mathcal{K})$, then $\text{proj}_{\mathcal{K}}(x_0 + w) = x_0$.*

When $n \gg d$, the orthogonal complement $\tan_{\mathcal{K}}(x)^{\perp}$ is large and will contain most of the weight of a random ϵ, intuitively suggesting $\text{proj}_{\mathcal{K}}(x_0+\sigma\epsilon) \approx x_0$ if σ is small. In other words, we intuitively expect that the *oracle denoiser*, which returns x_0 given $x_0 + \sigma\epsilon$, approximates projection.

We formalize this intuition using Gaussian concentration inequalities (Vershynin, 2018) and Lipschitz continuity of $\text{proj}_{\mathcal{K}}(x)$ when $\text{dist}_{\mathcal{K}}(x) < \text{reach}(\mathcal{K})$.

Proposition 3.1 (Oracle denoising (informal))**.** *Given $x_0 \in \mathcal{K}$, $\sigma > 0$ and $\epsilon \sim \mathcal{N}(0, I)$, let $x_\sigma = x_0 + \sigma\epsilon \in \mathbb{R}^n$. If $n \gg d$ and $\sigma\sqrt{n} \lesssim \text{reach}(\mathcal{K})$, then with high probability,*

$$\|\text{proj}_{\mathcal{K}}(x_\sigma) - x_0\| \lesssim \sqrt{\tfrac{d}{n}}\text{dist}_{\mathcal{K}}(x_\sigma)$$

and for $\nu \lesssim 1 + \sqrt{\frac{d}{n}}$, $\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x_\sigma) \le \sqrt{n}\sigma \le \nu\text{dist}_{\mathcal{K}}(x_\sigma)$.

Observe this motivates both constants η and ν used by our relative error model. We note prior analysis of diffusion under the manifold hypothesis is given by (De Bortoli, 2022).

3.2. Relative Error in Large Noise Regime

We next analyze denoisers in the large-noise regime, when $\text{dist}_{\mathcal{K}}(x)$ is much larger than $\text{diam}(\mathcal{K}) := \sup\{\|x - y\| : x, y \in \mathcal{K}\}$. In this regime (see Figure 1b for an illustration), any denoiser that predicts a point in the convex hull of $\mathcal{K}$ approximates projection with error small compared to $\text{dist}_{\mathcal{K}}(x)$.

Proposition 3.2. *Suppose $x - \sigma\epsilon_\theta(x, \sigma) \in \text{convhull}(\mathcal{K})$. If $\sqrt{n}\sigma \le \nu\text{dist}_{\mathcal{K}}(x)$, then $\|x - \sigma\epsilon_\theta(x, \sigma) - \text{proj}_{\mathcal{K}}(x)\| \le \nu\frac{\text{diam}(\mathcal{K})}{\sqrt{n}\sigma}\text{dist}_{\mathcal{K}}(x)$.*

Thus any denoiser that predicts any weighted mean of the data, for instance the ideal denoiser, well approximates projection when $\sqrt{n}\sigma \gg \text{diam}(\mathcal{K})$. For most diffusion models used in practice, $\sqrt{n}\sigma_N$ is usually 50-100 times the diameter of the training set, with $\sqrt{n}\sigma_t$ in this regime for a significant proportion of timesteps.

3.3. Relative Error of Ideal Denoisers

We now consider the setting where $\mathcal{K}$ is a finite set and $\epsilon_\theta(x, \sigma)$ is the ideal denoiser $\epsilon^*(x, \sigma)$. We first show that predicting $\hat{x}_0$ with the ideal denoiser is equivalent to projection using the σ-smoothed distance function:

Proposition 3.3. *For all $\sigma > 0$ and $x \in \mathbb{R}^n$, we have*

$$\nabla_x \tfrac{1}{2}\,\text{dist}^2_{\mathcal{K}}(x, \sigma) = \sigma\epsilon^*(x, \sigma).$$

This shows our relative error model is in fact a bound on

$$\|\nabla_x \,\text{dist}^2_{\mathcal{K}}(x, \sigma) - \nabla_x \,\text{dist}^2_{\mathcal{K}}(x)\|,$$

the error between the gradient of the smoothed distance function and that of the true distance function. Since the amount of smoothing is directly determined by σ, it is therefore natural to bound this error using $\text{dist}_{\mathcal{K}}(x)$ when $\sqrt{n}\sigma \approx \text{dist}_{\mathcal{K}}(x)$. In other words, Proposition 3.3 directly motivates our error-model.

Towards a rigorous bound, let $N_\alpha(x)$ denote the subset of $x_0 \in \mathcal{K}$ satisfying $\|x - x_0\| \le \alpha\text{dist}_{\mathcal{K}}(x)$ for $\alpha \ge 1$. Consider the following.

Proposition 3.4. *If $\alpha \ge 1 + \frac{2\nu^2}{n}\left(\frac{1}{e} + \log\left(\frac{|\mathcal{K}|}{\eta}\right)\right)$ and $\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x) \le \sqrt{n}\sigma$, then*

$$\|x - \sigma\epsilon^*(x, \sigma) - \text{proj}_{\mathcal{K}}(x)\| \le \eta\text{dist}_{\mathcal{K}}(x) + C_{x,\alpha},$$

where $C_{x,\alpha} := \sup_{x_0 \in N_\alpha} \|x_0 - \text{proj}_{\mathcal{K}}(x)\|$.

We can guarantee $C_{x,\alpha}$ is zero when $\text{dist}_{\mathcal{K}}(x)$ is small compared to the minimum pairwise distance of points in $\mathcal{K}$. Combined with Proposition 3.2, this shows the ideal denoiser has low relative error in both the large and small noise regimes. We cannot guarantee that $C_{x,\alpha}$ is small relative to $\text{dist}_{\mathcal{K}}(x)$ in all regimes, however, due to pathological cases, e.g., x is exactly between two points in N_α. Nevertheless, Figure 2 empirically shows that the relative error of the ideal denoiser $\|x_t - \sigma\epsilon^*(x, \sigma_t) - \text{proj}_{\mathcal{K}}(x_t)\| / \text{dist}_{\mathcal{K}}(x_t)$ is small at *all pairs* (x_t, σ_t) generated by the DDIM sampler, suggesting these pathologies do not appear in practice.

4. Gradient Descent Analysis of Sampling

Having justified the relative error model in Section 3, we now use it to study the DDIM sampler. As a warmup, we first consider the limiting case of zero error, where we see DDIM is precisely gradient descent on the squared-distance function with step-size determined by σ_t. We then generalize this result to arbitrary (η, ν), showing DDIM is equivalent to gradient-descent with relative error. Proofs are postponed to Appendix C.

4.1. Warmup: Exact Projection and Gradient Descent

We state our zero-error assumption in terms of the error-model as follows.

Assumption 1. *ϵ_θ is a $(0, 1)$-approximate projection.*

We can now characterize DDIM as follows.

Theorem 4.1. *Let $x_N, \ldots, x_0$ denote a sequence (5) generated by DDIM on a schedule $\{\sigma_t\}_{t=N}^0$ and $f(x) := \frac{1}{2}\text{dist}_{\mathcal{K}}(x)^2$. Suppose that Assumption 1 holds, $\nabla f(x_t)$ exists for all t and $\text{dist}_{\mathcal{K}}(x_N) = \sqrt{n}\sigma_N$. Then x_t is*

a sequence generated by gradient descent with step-size $\beta_t := 1 - \sigma_{t-1}/\sigma_t$*:*

$$x_{t-1} = x_t - \beta_t \nabla f(x_t),$$

and $\mathrm{dist}_{\mathcal{K}}(x_t) = \sqrt{n}\sigma_t$ *for all* t*.*

We remark that the existence of $\nabla f(x_t)$ is a weak assumption as it is generically satisfied by almost all $x \in \mathbb{R}^n$.

4.2. Approximate Projection and Gradient Descent with Error

We next establish upper and lower-bounds of distance under approximate gradient descent iterations. Given $\{\sigma_t\}_{t=N}^0$, let $\beta_t := 1 - \sigma_{t-1}/\sigma_t$ and

$$L_t^{\sigma,\eta} := \prod_{i=t}^{N}(1 - \beta_i(\eta+1)), U_t^{\sigma,\eta} := \prod_{i=t}^{N}(1 + \beta_i(\eta-1)).$$

Lemma 4.1. *For* $\mathcal{K} \subseteq \mathbb{R}^n$*, let* $f(x) := \frac{1}{2}\mathrm{dist}_{\mathcal{K}}(x)^2$*. If* $x_{t-1} = x_t - \beta_t(\nabla f(x_t) + e_t)$ *for* e_t *satisfying* $\|e_t\| \leq \eta \mathrm{dist}_{\mathcal{K}}(x_t)$ *and* $0 \leq \beta_t \leq 1$*, then*

$$L_t^{\sigma,\eta}\mathrm{dist}_{\mathcal{K}}(x_N) \leq \mathrm{dist}_{\mathcal{K}}(x_{t-1}) \leq U_t^{\sigma,\eta}\mathrm{dist}_{\mathcal{K}}(x_N).$$

Observe that the distance upper bound decreases only if $\beta_i < 1$ when $\eta > 0$. This conforms with our intuition that step sizes are limited by the error in our gradient estimates.

The challenge in applying Lemma 4.1 to DDIM lies in the specifics of our relative error model, which states that $\epsilon_\theta(x_t, \sigma_t)$ provides an η-accurate estimate of $\nabla f(x_t)$ *only if* σ_t provides a ν-accurate estimate of $\mathrm{dist}_{\mathcal{K}}(x_t)$. Hence we must first control the difference between σ_t and distance $\mathrm{dist}_{\mathcal{K}}(x_t)$ by imposing the following conditions on σ_t.

Definition 4.1. *We say that parameters* $\{\sigma_t\}_{t=N}^0$ *are* (η, ν)*-admissible if, for all* $t \in \{1, \ldots, N\}$*,*

$$\tfrac{1}{\nu}U_t^{\sigma,\eta} \leq \textstyle\prod_{i=t}^{N}(1-\beta_i) \leq \nu L_t^{\sigma,\eta}. \tag{8}$$

Intuitively, an admissible schedule decreases σ_t slow enough (corresponding to taking smaller gradient steps) to ensure $\frac{1}{\nu}\mathrm{dist}_{\mathcal{K}}(x_t) \leq \sqrt{n}\sigma_t \leq \nu \mathrm{dist}_{\mathcal{K}}(x_t)$ holds at each iteration. Our analysis assumes admissibility of the noise schedule and our relative-error model (Definition 3.1):

Assumption 2. *For* $\eta > 0$ *and* $\nu \geq 1$*,* $\{\sigma_t\}_{t=N}^0$ *is* (η, ν)*-admissible and* ϵ_θ *is an* (η, ν)*-approximate projection.*

Our main result follows. In simple terms, it states that DDIM is approximate gradient descent, admissible schedules σ_t are good estimates of distance, and the error bounds of Lemma 4.1 hold.

Theorem 4.2 (DDIM with relative error)**.** *Let* x_t *denote the sequence generated by DDIM. Suppose Assumption 2 holds, the gradient of* $f(x) := \frac{1}{2}\mathrm{dist}_{\mathcal{K}}(x)^2$ *exists for all* x_t *and* $\mathrm{dist}_{\mathcal{K}}(x_N) = \sqrt{n}\sigma_N$*. Then:*

- x_t *is generated by approximate gradient descent iterations of the form in Lemma 4.1 with* $\beta_t = 1 - \sigma_{t-1}/\sigma_t$*.*
- $\frac{1}{\nu}\mathrm{dist}_{\mathcal{K}}(x_t) \leq \sqrt{n}\sigma_t \leq \nu \mathrm{dist}_{\mathcal{K}}(x_t)$ *for all* t*.*
- $\mathrm{dist}_{\mathcal{K}}(x_N)L_t^{\sigma,\eta} \leq \mathrm{dist}_{\mathcal{K}}(x_{t-1}) \leq \mathrm{dist}_{\mathcal{K}}(x_N)U_t^{\sigma,\eta}$

4.3. Admissible Log-Linear Schedules for DDIM

We next characterize admissible σ_t of the form $\sigma_{t-1} = (1-\beta)\sigma_t$ where β denotes a *constant* step-size. This illustrates that admissible σ_t-sequences not only *exist*, they can also be explicitly constructed from (η, ν).

Theorem 4.3. *Fix* $\beta \in \mathbb{R}$ *satisfying* $0 \leq \beta < 1$ *and suppose that* $\sigma_{t-1} = (1-\beta)\sigma_t$*. Then* σ_t *is* (η, ν)*-admissible if and only if* $\beta \leq \beta_{*,N}$ *where* $\beta_{*,N} := \frac{c}{\eta + c}$ *for* $c := 1 - \nu^{-1/N}$*.*

Suppose we fix (η, ν) and choose, for a given N, the step-size $\beta_{*,N}$. It is natural to ask how the error bounds of Theorem 4.2 change as N increases. We establish the limiting behavior of the *final* output (σ_0, x_0) of DDIM.

Theorem 4.4. *Let* $x_N, \ldots, x_1, x_0$ *denote the sequence generated by DDIM with* σ_t *satisfying* $\sigma_{t-1} = (1-\beta_{*,N})\sigma_t$ *for* $\nu \geq 1$ *and* $\eta > 0$*. Then*

- $\lim_{N\to\infty} \frac{\sigma_0}{\sigma_N} = \lim_{N\to\infty}(1-\beta_{*,N})^N = \nu^{-1/\eta}$.
- $\lim_{N\to\infty} \frac{\mathrm{dist}_{\mathcal{K}}(x_0)}{\mathrm{dist}_{\mathcal{K}}(x_N)} \leq \lim_{N\to\infty}(1+(\eta-1)\beta_{*,N})^N = \nu^{\frac{\eta-1}{\eta}}$.

This theorem illustrates that final error, while bounded, need not converge to zero under our error model. This motivates heuristically updating the step-size from $\beta_{*,N}$ to a full step ($\beta = 1$) during the final DDIM iteration. We adopt this approach in our experiments (Section 6).

Next we demonstrate an explicit construction of an admissible schedule using numerical estimates of the error parameters on an image dataset.

Example 4.1 (Construction of admissible schedule)**.** *Let the CIFAR-10 training set be* $\mathcal{K}$ *and the ideal denoiser be* ϵ_θ*. From Figure 2, which plots the relative projection error relative to the training set, we see that* $\eta \leq 0.1$*. Our experiments comparing* $\mathrm{dist}_{\mathcal{K}}(x_t)$ *with* $\sqrt{n}\sigma_t$ *suggest that* $\nu = 2$ *is a conservative estimate, as the error in* $\mathrm{dist}_{\mathcal{K}}(x_t)$ *is bounded by this amount throughout the sampling trajectories. Theorem 4.3 shows that if* $\sigma_{t-1}/\sigma_t \geq \frac{\eta}{\eta+1-\nu^{-1/N}}$*, then* $\sigma_0, \ldots, \sigma_N$ *is an admissible schedule. With* $\eta = 0.1$*,* $\nu = 2$ *and* $N = 50$*, we obtain* $\sigma_{t-1}/\sigma_t \geq 0.88$*. This is very close to the value of* $\sigma_{t-1}/\sigma_t = 0.85$ *in the schedule used in our sampler in Section 6.1.*

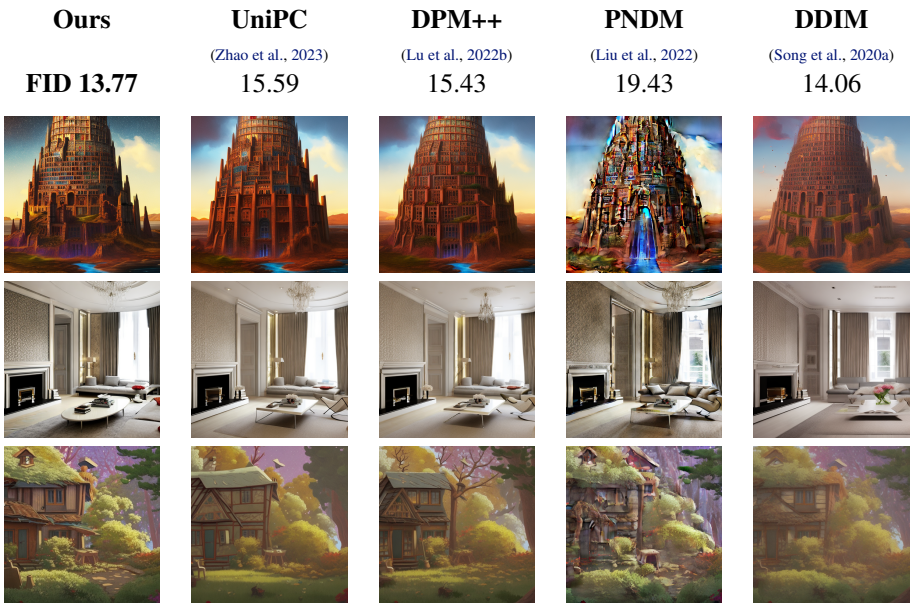


Figure 3: Outputs of our gradient-estimation sampler on text-to-image Stable Diffusion compared to other commonly used samplers, when limited to $N = 10$ function evaluations. We also report FID scores for text-to-image generation on MS-COCO 30K.

Algorithm 1 DDIM sampler (Song et al., 2020a)

Require: $(\sigma_N, \dots, \sigma_0)$, $x_N \sim \mathcal{N}(0, I)$, ϵ_θ
Ensure: Compute x_0 with N evaluations of ϵ_θ
  for $t = N, \dots, 1$ **do**
    $x_{t-1} \leftarrow x_t + (\sigma_{t-1} - \sigma_t)\epsilon_\theta(x_t, \sigma_t)$
  return x_0

Algorithm 2 Our gradient-estimation sampler

Require: $(\sigma_N, \dots, \sigma_0)$, $x_N \sim \mathcal{N}(0, I)$, ϵ_θ
Ensure: Compute x_0 with N evaluations of ϵ_θ
  $x_{N-1} \leftarrow x_N + (\sigma_{N-1} - \sigma_N)\epsilon_\theta(x_N, \sigma_N)$
  for $t = N - 1, \dots, 1$ **do**
    $\bar{\epsilon}_t \leftarrow 2\epsilon_\theta(x_t, \sigma_t) - \epsilon_\theta(x_{t+1}, \sigma_{t+1})$
    $x_{t-1} \leftarrow x_t + (\sigma_{t-1} - \sigma_t)\bar{\epsilon}_t$
  return x_0

5. Improving Deterministic Sampling Algorithms via Gradient Estimation

Section 3 establishes that $\epsilon_\theta(x, \sigma) \approx \sqrt{n}\nabla \mathrm{dist}_{\mathcal{K}}(x)$ when $\mathrm{dist}_{\mathcal{K}}(x) \approx \sqrt{n}\sigma$. We next exploit an invariant property of $\nabla \mathrm{dist}_{\mathcal{K}}(x)$ to reduce the prediction error of ϵ_θ via *gradient estimation*.

The gradient $\nabla \mathrm{dist}_{\mathcal{K}}(x)$ is *invariant* along line segments between a point x and its projection $\mathrm{proj}_{\mathcal{K}}(x)$, i.e., letting $\hat{x} = \mathrm{proj}_{\mathcal{K}}(x)$, for all $\theta \in (0, 1]$ we have

$$\nabla \mathrm{dist}_{\mathcal{K}}(\theta x + (1 - \theta)\hat{x}) = \nabla \mathrm{dist}_{\mathcal{K}}(x). \tag{9}$$

Hence, $\epsilon_\theta(x, \sigma)$ should be (approximately) constant on this line-segment under our assumption that $\epsilon_\theta(x, \sigma) \approx \sqrt{n}\nabla \mathrm{dist}_{\mathcal{K}}(x)$ when $\mathrm{dist}_{\mathcal{K}}(x) \approx \sqrt{n}\sigma$. Precisely, for x_1 and x_2 on this line-segment, we should have

$$\epsilon_\theta(x_1, \sigma_{t_1}) \approx \epsilon_\theta(x_2, \sigma_{t_2}) \tag{10}$$

if t_i satisfies $\mathrm{dist}_{\mathcal{K}}(x_i) \approx \sqrt{n}\sigma_{t_i}$. This property suggests combining previous denoiser outputs $\{\epsilon_\theta(x_i, \sigma_i)\}_{i=t+1}^{N}$ to estimate $\epsilon_t := \sqrt{n}\nabla \mathrm{dist}_{\mathcal{K}}(x_t)$. We next propose a practical *second-order method* [1] for this estimation that combines the current denoiser output with the previous. Recently introduced *consistency models* (Song et al., 2023) penalize violation of (10) during *training*. Interpreting denoiser output as $\nabla \mathrm{dist}_{\mathcal{K}}(x)$ and invoking (9) offers an alternative justification for these models.

Let $e_t(\epsilon_t) = \epsilon_t - \epsilon_\theta(x_t, \sigma_t)$ be the error of $\epsilon_\theta(x_t, \sigma_t)$ when predicting ϵ_t. To estimate ϵ_t from $\epsilon_\theta(x_t, \sigma_t)$, we minimize the norm of this error concatenated over two time-steps. Precisely, letting $y_t(\epsilon) = (e_t(\epsilon), e_{t+1}(\epsilon))$, we compute

$$\bar{\epsilon}_t := \arg\min_{\epsilon} \|y_t(\epsilon)\|_W^2, \tag{11}$$

[1] This method is second-order in the sense that the update step uses previous values of ϵ_θ, and should not be confused with second-order derivatives.

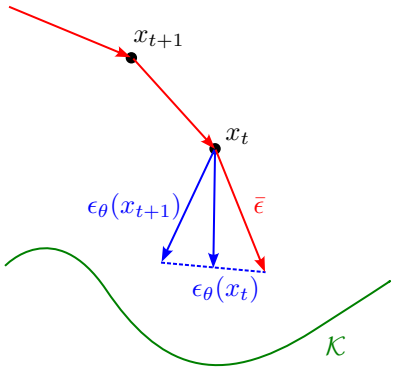


Figure 4: Illustration of our choice of $\bar{\epsilon}_t$

where W is a specified positive-definite weighting matrix. In Appendix D we show that this error model, for a particular family of weight matrices, results in the update rule

$$\bar{\epsilon}_t = \gamma\epsilon_\theta(x_t, \sigma_t) + (1-\gamma)\epsilon_\theta(x_{t+1}, \sigma_{t+1}), \qquad (12)$$

where we can search over W by searching over γ.

6. Experiments

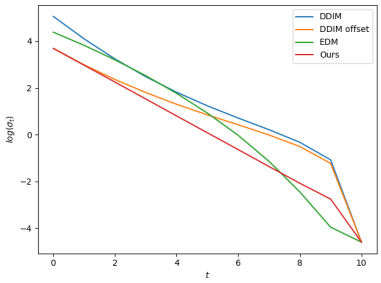


Figure 5: Plot of different choices of $\log(\sigma_t)$ for $N = 10$.

Schedule	CIFAR-10	CelebA
DDIM	16.86	18.08
DDIM Offset	14.18	15.38
EDM	20.85	16.72
Ours	**13.25**	**13.55**

Table 1: FID scores of the DDIM sampler (Algorithm 1) with different σ_t schedules on the CIFAR-10 model for $N = 10$ steps.

We evaluate modifications of DDIM (Algorithm 1) that leverage insights from Section 5 and Section 4.3. Following Section 5 we modify DDIM to use a second-order update that corrects for error in the denoiser output (Algorithm 2). Specifically, we use the Equation (12) update with an empirically tuned γ. We found that setting $\gamma = 2$ works well for $N < 20$; for larger N slightly increasing γ also improves sample quality (see Appendix E for more details). A comparison of this update with DDIM is visualized in Figure 4. Following Section 4.3, we select a noise schedule $(\sigma_N, \ldots, \sigma_0)$ that decreases at a log-linear (geometric) rate. The specific rate is determined by an initial and target noise level. Our σ_t schedule is illustrated in Figure 5, along with other commonly used schedules. We note that log-linear schedules have been previously proposed for SDE-samplers (Song et al., 2020b); to our knowledge we are the first to propose and analyze their use for DDIM[2]. All the experiments were run on a single Nvidia RTX 4090 GPU. Code for the experiments is available at `https://github.com/ToyotaResearchInstitute/gradient-estimation-sampler`

6.1. Evaluation of Noise Schedule

In Figure 5 we plot our schedule (with our choices of σ_t detailed in Appendix F) with three other commonly used schedules on a log scale. The first is the evenly spaced subsampling of the training noise levels used by DDIM. The second "DDIM Offset" uses the same even spacing but starts at a smaller σ_N, the same as that in our schedule. This type of schedule is typically used for guided image generation such as SDEdit (Meng et al., 2021). The third "EDM" is the schedule used in Karras et al. (2022, Eq. 5), with $\sigma_{\max} = 80, \sigma_{\min} = 0.002$ and $\rho = 7$.

We then test these schedules on the DDIM sampler Algorithm 1 by sampling images with $N = 10$ steps from the CIFAR-10 and CelebA models. We see that in Table 1 that our schedule improves the FID of the DDIM sampler on both datasets even without the second-order updates. This is in part due to choosing a smaller σ_N so the small number of steps can be better spent on lower noise levels (the difference between "DDIM" and "DDIM Offset"), and also because our schedule decreases σ_t at a faster rate than DDIM (the difference between "DDIM Offset" and "Ours").

6.2. Evaluation of Full Sampler

We quantitatively evaluate our gradient-estimation sampler (Algorithm 2) by computing the Fréchet Inception Distance (FID) (Heusel et al., 2017) between all the training images and 50k generated images. We use denoisers from (Ho et al., 2020; Song et al., 2020a) that were pretrained on the CIFAR-10 (32x32) and CelebA (64x64) datasets (Krizhevsky et al., 2009; Liu et al., 2015). We compare our results with other samplers using the same denoisers. The FID scores are tabulated in Table 2, showing that our sampler achieves better performance on both CIFAR-10 (for $N = 5, 10, 20, 50$) and

[2]DDIM is usually presented using not σ_t but parameters α_t satisfying $\sigma_t^2 = (1-\alpha_t)/\alpha_t$. Linear updates of σ_t are less natural when expressed in terms of α_t.

Table 2: FID scores of our gradient-estimation sampler compared to that of other samplers for pretrained CIFAR-10 and CelebA models with a discrete linear schedule. The first half of the table shows our computational results whereas the second half of the table show results taken from the respective papers. *Results for $N = 25$

	CIFAR-10 FID				CelebA FID			
Sampler	$N = 5$	$N = 10$	$N = 20$	$N = 50$	$N = 5$	$N = 10$	$N = 20$	$N = 50$
Ours	**12.57**	**3.79**	**3.32**	**3.41**	**10.76**	**4.41**	3.19	3.04
DDIM (Song et al., 2020a)	47.20	16.86	8.28	4.81	32.21	18.08	11.81	7.39
PNDM (Liu et al., 2022)	13.9	7.03	5.00	3.95	11.3	7.71	5.51	3.34
DPM (Lu et al., 2022a)		6.37	3.72	3.48		5.83	**2.82**	**2.71**
DEIS (Zhang & Chen, 2022)	18.43	7.12	4.53	3.78	25.07	6.95	3.41	2.95
UniPC (Zhao et al., 2023)	23.22	3.87						
A-DDIM (Bao et al., 2022)		14.00	5.81*	4.04		15.62	9.22*	6.13

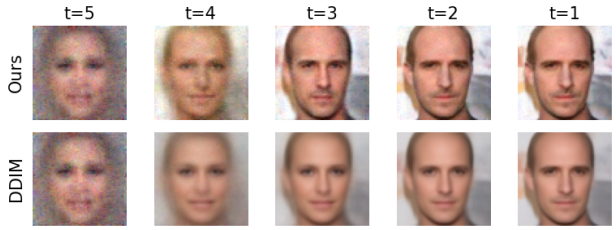


Figure 6: A comparison of our gradient-estimation sampler with DDIM on the CelebA dataset with $N = 5$ steps.

CelebA (for $N = 5, 10$).

We also incorporated our sampler into Stable Diffusion (a latent diffusion model). We change the noise schedule σ_t as described in Appendix F. In Figure 3, we show some example results for text to image generation in $N = 10$ function evaluations, as well as FID results on 30k images generated from text captions drawn the MS COCO (Lin et al., 2014) validation set. From these experiments we can see that our sampler performs comparably to other commonly used samplers, but with the advantage of being much simpler to describe and implement.

7. Related Work and Discussion

Learning diffusion models Diffusion was originally introduced as a variational inference method that learns to reverse a noising process (Sohl-Dickstein et al., 2015). This approach was empirically improved by (Ho et al., 2020; Nichol & Dhariwal, 2021) by introducing the training loss (2), which is different from the original variational lower bound. This improvement is justified from the perspective of denoising score matching (Song & Ermon, 2019; Song et al., 2020b), where the ϵ_θ is interpreted as $\nabla \log(p(x_t, \sigma_t))$, the gradient of the log density of the data distribution perturbed by noise. Score matching is also shown to be equivalent to denoising autoencoders with Gaussian noise (Vincent, 2011).

Sampling from diffusion models Samplers for diffusion models started with probabilistic methods (e.g. (Ho et al., 2020)) that formed the reverse process by conditioning on the denoiser output at each step. In parallel, score based models (Song & Ermon, 2019; Song et al., 2020b) interpret the forward noising process as a stochastic differential equation (SDE), so SDE solvers based on Langevian dynamics (Welling & Teh, 2011) are employed to reverse this process. As models get larger, computational constraints motivated the development of more efficient samplers. (Song et al., 2020a) then discovered that for smaller number of sampling steps, deterministic samplers perform better than stochastic ones. These deterministic samplers are constructed by reversing a non-Markovian process that leads to the same training objective, which is equivalent to turning the SDE into an ordinary differential equation (ODE) that matches its marginals at each sampling step.

This led to a large body of work focused on developing ODE and SDE solvers for fast sampling of diffusion models, a few of which we have evaluated in Table 2. Most notably, (Karras et al., 2022) put existing samplers into a common framework and isolated components that can be independently improved. Our gradient-estimation sampler Algorithm 2 bears most similarity to linear multistep methods, which can also be interpreted as accelerated gradient descent (Scieur et al., 2017). What differs is the error model: ODE solvers aim to minimize discretization error whereas we aim to minimize gradient estimation error, resulting in different "optimal" samplers.

Linear-inverse problems and conditioning Several authors (Kadkhodaie & Simoncelli, 2020; Chung et al., 2022; Kawar et al., 2022) have devised samplers for finding images that satisfy linear equations $Ax = b$. Such linear inverse problems generalize inpainting, colorization, and compressed sensing. In our framework, we can interpret this samplers as algorithms for equality constraint minimization of the distance function, a classical problem in optimization. Similarly, the widely used technique of *conditioning* (Dhari-

wal & Nichol, 2021) can be interpreted as multi-objective optimization, where minimization of distance is replaced with minimization of $\text{dist}_{\mathcal{K}}(x)^2 + g(x)$ for an auxiliary objective function $g(x)$.

Score distillation sampling We illustrate the potential of our framework for discovering new applications of diffusion models by deriving Score Distillation Sampling (SDS), a method for parameterized optimization introduced in (Poole et al., 2022) in the context of text to 3D object generation. At a high-level, this technique finds (x, θ) satisfying non-linear equations $x = g(\theta)$ subject to the constraint $x \in \mathcal{K}$, where $\mathcal{K}$ denotes the image manifold. It does this by iteratively updating x with a direction proportional to $(\epsilon_\theta(x + \sigma\epsilon, \sigma) - \epsilon)\nabla g(\theta)$, where σ is a randomly chosen noise level and $\epsilon \sim \mathcal{N}(0, I)$. Under our framework, this iteration can be interpreted as gradient descent on the squared-distance function with gradient $\frac{1}{2}\nabla_\theta \text{dist}_{\mathcal{K}}(g(\theta))^2 = (x - \text{proj}_{\mathcal{K}}(x))\nabla g(\theta)$, with the assumption that $\text{proj}_{\mathcal{K}}(x) \approx \text{proj}_{\mathcal{K}}(x + \sigma\epsilon)$, along with our Section 3 denoising approximation $\text{proj}_{\mathcal{K}}(x + \sigma\epsilon) \approx x + \sigma\epsilon - \sigma\epsilon_\theta(x + \sigma\epsilon, \sigma)$.

Flow matching Flow matching (Lipman et al., 2022) offers a different interpretation and generalization of diffusion models and deterministic sampling. Under this interpretation, the learned ϵ_θ represents a time-varying vector field, defining probability paths that transport the initial Gaussian distribution to the data distribution. For ϵ_θ learned with the denoising objective, we can interpret this vector field as the gradient of the smoothed squared-distance function $\text{dist}^2_{\mathcal{K}}(x, \sigma)$ (where σ changes as a function of t), thus moving along a probability path in this vector field minimizes the distance to the manifold.

Learning the distance function Reinterpreting denoising as projection, or equivalently gradient descent on the distance function, has a few immediate implications. First, it suggests generalizations that draw upon the literature for computing distance functions and projection operators. Such techniques include Fast Marching Methods (Sethian, 1996), kd-trees, and neural-network approaches, e.g., (Park et al., 2019; Rick Chang et al., 2017). Using concentration inequalities, we can also interpret training a denoiser as learning a solution to the *Eikonal PDE*, given by $\|\nabla d(x)\| = 1$. Other techniques for solving this PDE with deep neural nets include (Smith et al., 2020; Lichtenstein et al., 2019; bin Waheed et al., 2021).

8. Conclusion and Future Work

We have presented a simple framework for analyzing and generalizing diffusion models that has led to a new sampling approach and new interpretations of pre-existing techniques. Moreover, the key objects in our analysis —the distance function and the projection operator—are canonical objects in constrained optimization. We believe our work can lead to new generative models that incorporate sophisticated objectives and constraints for a variety of applications. We also believe this work can be leveraged to incorporate existing denoisers into optimization algorithms in a plug-in-play fashion, much like the work in (Chan et al., 2016; Le Pendu & Guillemot, 2023; Rick Chang et al., 2017).

Combining the multi-level noise paradigm of diffusion with distance function learning (Park et al., 2019) is an interesting direction, as are diffusion-models that carry out projection using analytic formulae or simple optimization routines.

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 which we feel must be specifically highlighted here.

Acknowledgements

The authors would like to thank Preetum Nakkiran, Vaishaal Shankar, Lingxiao Li and Pablo Parrilo for insightful discussions and comments on the manuscript. We would also like to thank the anonymous referees for their feedback.

References

Bao, F., Li, C., Zhu, J., and Zhang, B. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

bin Waheed, U., Haghighat, E., Alkhalifah, T., Song, C., and Hao, Q. Pinneik: Eikonal solution using physics-informed neural networks. *Computers & Geosciences*, 155:104833, 2021.

Chan, S. H., Wang, X., and Elgendy, O. A. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022.

De Bortoli, V. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Delfour, M. C. and Zolésio, J.-P. *Shapes and geometries: metrics, analysis, differential calculus, and optimization*. SIAM, 2011.

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

Federer, H. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

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

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Le Pendu, M. and Guillemot, C. Preconditioned plug-and-play admm with locally adjustable denoiser for image restoration. *SIAM Journal on Imaging Sciences*, 16(1): 393–422, 2023.

Lichtenstein, M., Pai, G., and Kimmel, R. Deep eikonal solvers. In *Scale Space and Variational Methods in Computer Vision: 7th International Conference, SSVM 2019, Hofgeismar, Germany, June 30–July 4, 2019, Proceedings 7*, pp. 38–50. Springer, 2019.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023.

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.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022a.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rick Chang, J., Li, C.-L., Poczos, B., Vijaya Kumar, B., and Sankaranarayanan, A. C. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5888–5897, 2017.

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

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Scieur, D., Roulet, V., Bach, F., and d'Aspremont, A. Integration methods and optimization algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Sethian, J. A. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

Smith, J. D., Azizzadenesheli, K., and Ross, Z. E. Eikonet: Solving the eikonal equation with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10685–10696, 2020.

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. *arXiv preprint arXiv:2010.02502*, 2020a.

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., and Bermano, A. H. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Zhang, Q. and Chen, Y. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

Zhao, W., Bai, L., Rao, Y., Zhou, J., and Lu, J. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *arXiv preprint arXiv:2302.04867*, 2023.

A. Equivalent Definitions of DDIM and DDPM

The DDPM and DDIM samplers are usually described in a different coordinate system z_t defined by parameters $\bar{\alpha}_t$ and the following relations , where the noise model is defined by a schedule $\bar{\alpha}_t$:

$$y \approx \sqrt{\bar{\alpha}_t} z + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{13}$$

with the estimate $\hat{z}_0^t := \hat{z}_0(z_t, t)$ given by

$$\hat{z}_0(y,t) := \frac{1}{\sqrt{\bar{\alpha}_t}}(y - \sqrt{1-\bar{\alpha}_t}\epsilon'_\theta(y,t)). \tag{14}$$

We have the following conversion identities between the x and z coordinates:

$$x_0 = z_0, \quad x_t = z_t/\sqrt{\bar{\alpha}_t}, \quad \sigma_t = \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}, \quad \epsilon_\theta(y, \sigma_t) = \epsilon'_\theta(y/\sqrt{\bar{\alpha}_t}, t). \tag{15}$$

While this change-of-coordinates is used in Song et al. (2020a, Section 4.3) and in (Karras et al., 2022)–and hence not new– we rigorously prove equivalence of the DDIM and DDPM samplers given in Section 2 with their original definitions.

DDPM Given initial z_N, the DDPM sampler constructs the sequence

$$z_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\hat{z}_0^t + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} z_t + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)}w_t, \tag{16}$$

where $\alpha_t := \bar{\alpha}_t/\bar{\alpha}_{t-1}$ and $w_t \sim \mathcal{N}(0, I)$. This is interpreted as sampling z_{t-1} from a Gaussian distribution conditioned on z_t and $\hat{z}_0^t$ (Ho et al., 2020).

Proposition A.1 (DDPM change of coordinates)**.** *The sampling update* (4) *is equivalent to the update* (16) *under the change of coordinates* (15)*.*

Proof. First we write (4) in terms of z_t, $\epsilon'_\theta(z_t, t)$ and w_t using (14):

$$\begin{aligned} z_{t-1} &= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_t)}\left(z_t - \sqrt{1-\bar{\alpha}_t}\epsilon'_\theta(z_t,t)\right) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} z_t + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)}w_t \\ &= \frac{z_t}{\sqrt{\alpha_t}} + \frac{\alpha_t - 1}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}\epsilon'_\theta(z_t,t) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)}w_t. \end{aligned}$$

Next we divide both sides by $\sqrt{\bar{\alpha}_{t-1}}$ and change z_t and z_{t-1} to x_t and x_{t-1}:

$$x_{t-1} = x_t + \frac{\alpha_t - 1}{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_t)}}\epsilon_\theta(x_t, \sigma_t) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}\frac{1-\alpha_t}{1-\bar{\alpha}_t}}w_t.$$

Now if we define

$$\begin{aligned} \eta &:= \sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}\frac{1-\alpha_t}{1-\bar{\alpha}_t}} = \sigma_{t-1}\sqrt{\frac{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}, \\ \sigma_{t'} &:= \sqrt{\sigma_{t-1}^2 - \eta^2} = \sigma_{t-1}\sqrt{\frac{\bar{\alpha}_t(1/\bar{\alpha}_{t-1}-1)}{1-\bar{\alpha}_t}} = \frac{\sigma_{t-1}^2}{\sigma_t}, \end{aligned}$$

it remains to check that

$$\sigma_{t'} - \sigma_t = \frac{\sigma_{t-1}^2 - \sigma_t^2}{\sigma_t} = \frac{1/\bar{\alpha}_{t-1} - 1/\bar{\alpha}_t}{\sqrt{1-\bar{\alpha}_t}/\sqrt{\bar{\alpha}_t}} = \frac{\alpha_t - 1}{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_t)}}.$$

□

DDIM Given initial z_N, the DDIM sampler constructs the sequence

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{z}_0^t + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon'_\theta(z_t, t), \tag{17}$$

i.e., it estimates $\hat{z}_0^t$ from z_t and then constructs z_{t-1} by simply updating $\bar{\alpha}_t$ to $\bar{\alpha}_{t-1}$. This sequence can be equivalently expressed in terms of $\hat{z}_0^t$ as

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{z}_0^t + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}(z_t - \sqrt{\bar{\alpha}_t}\hat{z}_0^t). \tag{18}$$

Proposition A.2 (DDIM change of coordinates)**.** *The sampling update* (5) *is equivalent to the update* (18) *under the change of coordinates* (15)*.*

Proof. First we write (17) in terms of z_t and $\epsilon'_\theta(z_t, t)$ using (14):

$$z_{t-1} = \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}z_t + \left(\sqrt{1-\bar{\alpha}_{t-1}} - \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}\sqrt{1-\bar{\alpha}_t}\right)\epsilon'_\theta(z_t, t).$$

Next we divide both sides by $\sqrt{\bar{\alpha}_{t-1}}$ and change z_t and z_{t-1} to x_t and x_{t-1}:

$$\begin{aligned} x_{t-1} &= x_t + \left(\sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}} - \sqrt{\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\right)\epsilon_\theta(x_t, \sigma_t) \\ &= x_t + (\sigma_{t-1} - \sigma_t)\epsilon_\theta(x_t, \sigma_t). \end{aligned}$$

□

B. Formal Comparison of Denoising and Projection

B.1. Proof of Proposition 3.1

First, we state the formal version of Proposition 3.1

Proposition B.1 (Oracle denoising)**.** *Fix $\sigma > 0$, $t > 0$ and let $\kappa(t) := \sqrt{(\sqrt{d}+t)^2 + (\sqrt{n-d}+t)^2}$. Given $x_0 \in \mathcal{K}$ and $\epsilon \sim \mathcal{N}(0, I)$, let $x_\sigma = x_0 + \sigma\epsilon$. Suppose that $\mathrm{reach}(\mathcal{K}) > \sigma\kappa(t)$ and $\sqrt{n-d} - \sqrt{d} - 2t > 0$. Then, for an absolute constant $\alpha > 0$, we have, with probability at least $(1 - \exp(-\alpha t^2))^2$, that*

$$\sigma(\sqrt{n-d} - \sqrt{d} - 2t) \le \mathrm{dist}(x_\sigma) \le \sigma(\sqrt{n-d} + \sqrt{d} + 2t)$$

and

$$\|\mathrm{proj}_{\mathcal{K}}(x_\sigma) - x_0\| \le \frac{C(t)(\sqrt{d}+t)}{\sqrt{n-d}-\sqrt{d}-2t}\mathrm{dist}_{\mathcal{K}}(x_\sigma)$$

where $C(t) := \frac{\mathrm{reach}(\mathcal{K})}{\mathrm{reach}(\mathcal{K}) - \sigma\kappa(t)}$.

Our proof uses local Lipschitz continuity of the projection operator, stated formally as follows.

Proposition B.2 (Theorem 6.2(vi), Chapter 6 of (Delfour & Zolésio, 2011))**.** *Suppose $0 < \mathrm{reach}(\mathcal{K}) < \infty$. Consider $h > 0$ and $x, y \in \mathbb{R}^n$ satisfying $0 < h < \mathrm{reach}(\mathcal{K})$ and $\mathrm{dist}_{\mathcal{K}}(x) \le h$ and $\mathrm{dist}_{\mathcal{K}}(y) \le h$. Then the projection map satisfies $\|\mathrm{proj}_{\mathcal{K}}(y) - \mathrm{proj}_{\mathcal{K}}(x)\| \le \frac{\mathrm{reach}(\mathcal{K})}{\mathrm{reach}(\mathcal{K}) - h}\|y - x\|$.*

We also use the following concentration inequalities.

Proposition B.3. *Let $w \sim \mathcal{N}(0, \sigma^2 I_n)$. Let S be a fixed subspace of dimension d and denote by w_S and $w_{S^\perp}$ the projections onto S and $S^\perp$ respectively. Then for an absolute constant α, the following statements hold*

- *With probability at least $1 - \exp(-\alpha t^2)$,*

$$\sigma(\sqrt{n} - t) \le \|w\| \le \sigma(\sqrt{n} + t)$$

- *With probability at least* $(1-\exp(-\alpha t^2))^2$,

$$\sigma(\sqrt{d}-t) \le \|w_S\| \le \sigma(\sqrt{d}+t), \qquad \sigma(\sqrt{n-d}-t) \le \|w_{S^\perp}\| \le \sigma(\sqrt{n-d}+t),$$

Proof. The first statement is proved in (Vershynin, 2018, page 44, Equation 3.3). For the second, let $B \in \mathbb{R}^{n\times d}$ denote an orthonormal basis for S and define $y = B^T w$ Then $\|y\| = \|w_S\|$. Further, $y \sim \mathcal{N}(0, \sigma^2 I_{d\times d})$ given that $\mathrm{cov}(y) = \sigma^2 B^T B = \sigma^2 I_{d\times d}$. Hence, the bounds on $\|w_S\|$ hold with probability at least $p := 1-\exp(-\alpha t^2)$ given the first statement. By similar argument, the bounds on $\|w_{S^\perp}\|$ also hold with probability p. Since w_S and $w_{S^\perp}$ are independent, we deduce that both sets of bounds simultaneously hold with probability at least p^2. □

To prove Proposition B.1, we decompose random noise $\sigma\epsilon$ as

$$\sigma\epsilon = w_N + w_T \tag{19}$$

for $w_T \in \tan_{\mathcal{K}}(x_0)$ and $w_N \in \tan_{\mathcal{K}}(x_0)^\perp$ and use Lemma 3.1. The proof follows.

Proof of Proposition B.1. Let $p := 1-\exp(-\alpha t^2)$. Proposition B.3 asserts that, with probability at least p^2,

$$\sigma(\sqrt{d}-t) \le \|w_T\| \le \sigma(\sqrt{d}+t), \qquad \sigma(\sqrt{n-d}-t) \le \|w_N\| \le \sigma(\sqrt{n-d}+t), \tag{20}$$

These inequalities imply the claimed bounds on $\mathrm{dist}_{\mathcal{K}}(x_\sigma)$, given that

$$\|w_N\| - \|w_T\| \le \mathrm{dist}_{\mathcal{K}}(x_\sigma) \le \|w_N\| + \|w_T\|$$

by Lemma C.2 and the fact $\mathrm{dist}(x_0 + w_N) = \|w_N\|$ under the reach assumption and Lemma 3.1.

Using $\mathrm{proj}(x_0 + w_N) = x_0$, we observe that

$$\begin{aligned}
\|\mathrm{proj}(x_\sigma) - x_0\| &= \|\mathrm{proj}(x_0 + w_N + w_T) - x_0\| \\
&= \|\mathrm{proj}(x_0 + w_N) - x_0 + \mathrm{proj}(x_0 + w_N + w_T) - \mathrm{proj}(x_0 + w_N)\| \\
&= \|\mathrm{proj}(x_0 + w_N) - \mathrm{proj}(x_0 + w_N + w_T)\| \\
&\le C\|w_T\| \\
&\le C\sigma(\sqrt{d}+t)
\end{aligned}$$

where the second-to-last inequality comes from Proposition B.2 using the fact that $\mathrm{reach}(\mathcal{K}) > w$ and the inequalities $\mathrm{dist}_{\mathcal{K}}(x_0 + w_N) = \|w_N\| \le \|w\|$ and $\mathrm{dist}_{\mathcal{K}}(x_0 + w_N + w_T) \le \|w\|$. The proof is completed by dividing by our lower bound of $\mathrm{dist}_{\mathcal{K}}(x_\sigma)$.

□

B.2. Proof of Proposition 3.3

First we derive an explicit expression for the ideal denoiser for a uniform distribution over a finite set.

Lemma B.1. *When D is a discrete uniform distribution over a set $\mathcal{K}$, the ideal denoiser ϵ^* is given by*

$$\epsilon^*(x_\sigma, \sigma) = \frac{\sum_{x_0\in\mathcal{K}} (x_\sigma - x_0)\exp(-\|x_\sigma - x_0\|^2/2\sigma^2)}{\sigma\sum_{x_0\in\mathcal{K}}\exp(-\|x_\sigma - x_0\|^2/2\sigma^2)}.$$

Proof. Writing the loss explicitly as

$$\mathcal{L}_\sigma(\epsilon^*) = \int \sum_{x_0\in\mathcal{K}} \frac{1}{|\mathcal{K}|\,\sigma\sqrt{2\pi}} \exp\left(-\frac{\|x_\sigma - x_0\|^2}{2\sigma^2}\right) \|(x_\sigma - x_0)/\sigma - \epsilon^*(x_\sigma,\sigma)\|^2 \, d(x_\sigma),$$

It suffices to take the point-wise minima of the expression inside the integral, which is convex in terms of ϵ^*. □

From this expression of the ideal denoiser ϵ^*, we see that $\hat{x}_0$ can be written as a convex combination of points in $\mathcal{K}$:

$$\hat{x}_0(x,\sigma) = \sigma\epsilon^*(x,\sigma) - x = \sum_{x_0\in\mathcal{K}} w(x,x_0)x_0,$$

where $\sum_{x_0\in\mathcal{K}} w(x,x_0) = 1$.

By taking the gradient of the log-sum-exp function in the definition of $\operatorname{dist}^2_{\mathcal{K}}(x,\sigma)$ then applying Lemma B.1, it is clear that

$$\nabla_x \tfrac{1}{2}\operatorname{dist}^2_{\mathcal{K}}(x,\sigma) = \sigma\epsilon^*(x,\sigma).$$

B.3. Proof of Proposition 3.4

We wish to bound the error between the gradient of the true distance function and the result of the ideal denoiser. Precisely, we want to upper-bound the following in terms of $\operatorname{dist}_{\mathcal{K}}(x) = \|x - x_0^*\|$, where $x_0^* = \operatorname{proj}_{\mathcal{K}}(x)$. We first define

$$w(x,x') := \frac{\exp(-\|x-x'\|^2/2\sigma^2)}{\sum_{x_0\in\mathcal{K}}\exp(-\|x-x_0\|^2/2\sigma^2)}.$$

Note that by definition, we have $\sum_{x_0\in\mathcal{K}} w(x,x_0) = 1$. Letting $\bar{N}_\alpha$ denote the complement of N_α in $\mathcal{K}$, we have

$$\begin{aligned}
\left\|\nabla\tfrac{1}{2}\operatorname{dist}_{\mathcal{K}}(x)^2 - \sigma\epsilon^*(x,\sigma)\right\| &= \left\|\nabla\tfrac{1}{2}\operatorname{dist}_{\mathcal{K}}(x)^2 - \nabla\tfrac{1}{2}\operatorname{dist}^2_{\mathcal{K}}(x,\sigma)\right\| \\
&= \left\|x_0^* - \textstyle\sum_{x_0\in\mathcal{K}} w(x,x_0)x_0\right\| \\
&= \left\|\textstyle\sum_{x_0\in\mathcal{K}} w(x,x_0)(x_0^* - x_0)\right\| \\
&\le \|\sum_{x_0\in\bar{N}_\alpha} w(x,x_0)(x_0^*-x_0)\| + \|\sum_{x_0\in N_\alpha} w(x,x_0)(x_0^*-x_0)\| \\
&\le \|\sum_{x_0\in\bar{N}_\alpha} w(x,x_0)(x_0^*-x_0)\| + C_{x,\alpha}
\end{aligned}$$

The claim then follows from the following theorem.

Theorem B.1. *Suppose $\mathcal{K}$ is a finite-set and let $x_0^* = \operatorname{proj}_{\mathcal{K}}(x)$. Suppose we have*

$$\alpha \ge 1 + \frac{2\sigma^2}{\operatorname{dist}_{\mathcal{K}}(x)^2}\left(\frac{1}{e} + \log\left(\frac{m}{\eta}\right)\right), \tag{21}$$

then $\left\|\sum_{x_0\in\bar{N}_\alpha} w(x,x_0)(x_0^*-x_0)\right\| \le \eta\operatorname{dist}_{\mathcal{K}}(x)$.

Proof. Applying the triangle inequality, it suffices to upper-bound each of $w(x,x_0)\|x_0^*-x_0\|$. For convenience of notation let $\delta(x_0) := \|x-x_0\| / \|x-x_0^*\|$. Note that by construction $\delta(x_0)\ge 1$ for all $x_0\in\mathcal{K}$, and $\delta(x_0)\ge\alpha$ for all $x_0\in\bar{N}_\alpha$. Then

$$\begin{aligned}
\|x_0^*-x_0\| &\le \|x_0^*-x\| + \|x-x_0\| = (1+\delta(x_0))\|x-x_0^*\|, \\
w(x,x_0) &\le \exp\left(-\frac{\|x-x_0\|^2 - \|x-x_0^*\|^2}{2\sigma^2}\right) \le \exp\left(-\frac{(\delta(x_0)^2-1)\|x-x_0^*\|^2}{2\sigma^2}\right).
\end{aligned}$$

From (21) and the fact that $1/e \ge \log(a)/a$ for $a = \delta(x_0)+1 \ge 1$, we have

$$\begin{aligned}
\delta(x_0) - 1 \ge \alpha - 1 &\ge \frac{2\sigma^2}{a\|x-x_0^*\|^2}\left(\log(a) + \log\left(\frac{m}{\epsilon}\right)\right) = \frac{2\sigma^2}{(\delta(x_0)+1)\|x-x_0^*\|^2}\log\left(\frac{m(\delta(x_0)+1)}{\epsilon}\right) \\
\delta(x_0)^2 - 1 &\ge \frac{2\sigma^2}{\|x-x_0^*\|^2}\log\left(\frac{m(\delta(x_0)+1)}{\epsilon}\right)
\end{aligned}$$

Putting these together, we have:

$$\begin{aligned}\left\|\sum_{x_0\in\bar{N}_\alpha} w(x,x_0)(x_0^*-x_0)\right\| &\le \sum_{x_0\in\bar{N}_\alpha}(1+\delta(x_0))\left\|x-x_0^*\right\|\exp\left(-\frac{(\delta(x_0)^2-1)\left\|x-x_0^*\right\|^2}{2\sigma^2}\right)\\ &\le \sum_{x_0\in\bar{N}_\alpha}\frac{\epsilon\left\|x-x_0^*\right\|}{m}\\ &\le \epsilon\mathrm{dist}_{\mathcal{K}}(x)\end{aligned}$$

□

C. DDIM with Projection Error Analysis

C.1. Proof of Theorem 4.1

We use the following lemma for gradient descent applied to the squared-distance function $f(x)$.

Lemma C.1. *Fix $x\in\mathbb{R}^n$ and suppose that $\nabla f(x)$ exists. For step-size $0<\beta\le 1$ consider the gradient descent iteration applied to $f(x)$:*

$$x_+ := x-\beta\nabla f(x)$$

Then, $\mathrm{dist}_{\mathcal{K}}(x_+)=(1-\beta)\,\mathrm{dist}_{\mathcal{K}}(x)<\mathrm{dist}_{\mathcal{K}}(x)$.

Make the inductive hypothesis that $\mathrm{dist}(x_t)=\sqrt{n}\sigma_t$. From the definition of DDIM (5), we have

$$x_{t-1}=x_t+(\frac{\sigma_{t-1}}{\sigma_t}-1)\sigma_t\epsilon_\theta(x_t,\sigma_t).$$

Under Assumption 1 and the inductive hypothesis, we conclude

$$\begin{aligned}x_{t-1}&=x_t+(\frac{\sigma_{t-1}}{\sigma_t}-1)\nabla f(x_t)\\ &=x_t-\beta_t\nabla f(x_t)\end{aligned}$$

Using Lemma C.1 we have that

$$\mathrm{dist}(x_{t-1})=(1-\beta_t)\,\mathrm{dist}(x_t)=\frac{\sigma_{t-1}}{\sigma_t}\,\mathrm{dist}(x_t)=\sqrt{n}\sigma_{t-1}$$

The base case holds by assumption, proving the claim.

C.2. Proof of Lemma C.1

Letting $x_0=\mathrm{proj}_{\mathcal{K}}(x)$ and noting $\nabla f(x)=x-x_0$, we have

$$\begin{aligned}\mathrm{dist}_{\mathcal{K}}(x_+)&=\mathrm{dist}_{\mathcal{K}}(x+\beta(x_0-x))\\ &=\|x+\beta(x_0-x)-x_0\|\\ &=\|(x-x_0)(1-\beta)\|\\ &=(1-\beta)\mathrm{dist}_{\mathcal{K}}(x)\end{aligned}$$

C.3. Distance function bounds

The distance function admits the following upper and lower bounds.

Lemma C.2. *The distance function $\mathrm{dist}_{\mathcal{K}}:\mathbb{R}^n\to\mathbb{R}$ for $\mathcal{K}\subseteq\mathbb{R}^n$ satisfies*

$$\mathrm{dist}_{\mathcal{K}}(u)-\|u-v\|\le\mathrm{dist}_{\mathcal{K}}(v)\le\mathrm{dist}_{\mathcal{K}}(u)+\|u-v\|$$

for all $u,v\in\mathbb{R}^n$.

Proof. By (Delfour & Zolésio, 2011, Chapter 6, Theorem 2.1), $|\text{dist}_{\mathcal{K}}(u) - \text{dist}_{\mathcal{K}}(v)| \leq \|u - v\|$, which is equivalent to

$$\text{dist}_{\mathcal{K}}(u) - \text{dist}_{\mathcal{K}}(v) \leq \|u - v\|, \text{dist}_{\mathcal{K}}(v) - \text{dist}_{\mathcal{K}}(u) \leq \|u - v\|.$$

Rearranging proves the claim. □

C.4. Proof of Lemma 4.1

We first restate the full version of Lemma 4.1.

Lemma C.3. *For $\mathcal{K} \subseteq \mathbb{R}^n$, let $f(x) := \frac{1}{2}\text{dist}_{\mathcal{K}}(x)^2$. The following statements hold.*

(a) If $x_+ = x - \beta(\nabla f(x) + e)$ for e satisfying $\|e\| \leq \eta \text{dist}_{\mathcal{K}}(x)$ and $0 \leq \beta \leq 1$, then

$$(1 - \beta(\eta + 1))\text{dist}_{\mathcal{K}}(x) \leq \text{dist}_{\mathcal{K}}(x_+) \leq (1 + \beta(\eta - 1))\text{dist}_{\mathcal{K}}(x).$$

(b) If $x_{t-1} = x_t - \beta_t(\nabla f(x_t) + e_t)$ for e_t satisfying $\|e_t\| \leq \eta \text{dist}_{\mathcal{K}}(x_t)$ and $0 \leq \beta_t \leq 1$, then

$$\text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 - \beta_i(\eta + 1)) \leq \text{dist}_{\mathcal{K}}(x_{t-1}) \leq \text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 + \beta_i(\eta - 1).)$$

For Item (a) we apply Lemma C.2 at points $u = x_+$ and $v = x - \beta\nabla f(x)$. We also use $\text{dist}(v) = (1 - \beta)\text{dist}_{\mathcal{K}}(x)$, since $0 \leq \beta \leq 1$, to conclude that

$$(1 - \beta)\text{dist}_{\mathcal{K}}(x) - \beta\|e\| \leq \text{dist}_{\mathcal{K}}(x_+) \leq (1 - \beta)\text{dist}_{\mathcal{K}}(x) + \beta\|e\|.$$

Using the assumption that $\|e\| \leq \eta \text{dist}_{\mathcal{K}}(x)$ gives

$$(1 - \beta - \eta\beta)\text{dist}_{\mathcal{K}}(x) \leq \text{dist}_{\mathcal{K}}(x_+) \leq (1 - \beta + \eta\beta)\text{dist}_{\mathcal{K}}(x)$$

Simplifying completes the proof. Item (b) follows from Item (a) and induction.

C.5. Proof of Theorem 4.2

We first state and prove an auxiliary theorem:

Theorem C.1. *Suppose Assumption 2 holds for $\nu \geq 1$ and $\eta > 0$. Given x_N and $\{\beta_t, \sigma_t\}_{i=1}^N$, recursively define $x_{t-1} = x_t + \beta_t\sigma_t\epsilon_\theta(x_t, t)$ and suppose that $\text{proj}_{\mathcal{K}}(x_t)$ is a singleton for all t. Finally, suppose that $\{\beta_t, \sigma_t\}_{i=1}^N$ satisfies $\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x_N) \leq \sqrt{n}\sigma_N \leq \nu\text{dist}_{\mathcal{K}}(x_N)$ and*

$$\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 + \beta_i(\eta - 1)) \leq \sqrt{n}\sigma_{t-1} \leq \nu\text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 - \beta_i(\eta + 1)). \tag{22}$$

The following statements hold.

- $\text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 - \beta_i(\eta + 1)) \leq \text{dist}_{\mathcal{K}}(x_{t-1}) \leq \text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 + \beta_i(\eta - 1))$
- $\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x_{t-1}) \leq \sqrt{n}\sigma_{t-1} \leq \nu\text{dist}_{\mathcal{K}}(x_{t-1})$

Proof. Since $\text{proj}_{\mathcal{K}}(x_t)$ is a singleton, $\nabla f(x_t)$ exists. Hence, the result will follow from Item (b) of Lemma C.3 if we can show that $\|\beta_t\sigma_t\epsilon_\theta(x_t, t) - \nabla f(x_t)\| \leq \eta\text{dist}_{\mathcal{K}}(x_t)$. Under Assumption 2, it suffices to show that

$$\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x_t) \leq \sqrt{n}\sigma_t \leq \nu\text{dist}_{\mathcal{K}}(x_t) \tag{23}$$

holds for all t. We use induction, noting that the base case ($t = N$) holds by assumption. Suppose then that (23) holds for all $t, t + 1, \ldots, N$. By Lemma 4.1 and Assumption 2, we have

$$\text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 - \beta_i(\eta + 1)) \leq \text{dist}_{\mathcal{K}}(x_{t-1}) \leq \text{dist}_{\mathcal{K}}(x_N)\prod_{i=t}^{N}(1 + (\eta - 1)\beta_i)$$

Combined with (22) shows

$$\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x_{t-1}) \leq \sqrt{n}\sigma_{t-1} \leq \nu\text{dist}_{\mathcal{K}}(x_{t-1}),$$

proving the claim. □

Theorem 4.2 follows by observing the admissibility assumption and the DDIM step-size rule, which satisfies $\sigma_{t-1} = (1-\beta_t)\sigma_t$, implies (22).

C.6. Proof of Theorem 4.3

Assuming constant step-size $\beta_i = \beta$ and dividing (8) by $\prod_{i=1}^{N}(1-\beta)$ gives the conditions

$$\left(1+\eta\frac{\beta}{1-\beta}\right)^N \leq \nu, \qquad \left(1-\eta\frac{\beta}{1-\beta}\right)^N \geq \frac{1}{\nu}.$$

Rearranging and defining $a = \eta\frac{\beta}{1-\beta}$ and $b = \nu^{\frac{1}{N}}$ gives

$$a \leq b-1, \qquad a \leq 1-b^{-1}.$$

Since $b-1-(1-b^{-1}) = b+b^{-1}-2 \geq 0$ for all $b>0$, we conclude $a \leq b-1$ holds if $a \leq 1-b^{-1}$ holds. We therefore consider the second inequality $\eta\frac{\beta}{1-\beta} \leq 1-\nu^{-1/N}$, noting that it holds for all $0 \leq \beta < 1$ if and only if $0 \leq \beta \leq \frac{k}{1+k}$ for $k = \frac{1}{\eta}(1-\nu^{-1/N})$, proving the claim.

C.7. Proof of Theorem 4.4

The value of σ_0/σ_N follows from the definition of σ_t and and the upper bound for $\text{dist}_{\mathcal{K}}(x_0)/\text{dist}_{\mathcal{K}}(x_N)$ follows from Theorem 4.3. We introduce the parameter μ to get a general form of the expression inside the limit:

$$(1-\mu\beta_{*,N})^N = \left(1-\mu\frac{1-\nu^{-1/N}}{\eta+1-\nu^{-1/N}}\right)^N.$$

Next we take the limit using L'Hôpital's rule:

$$\begin{aligned}
\lim_{N\to\infty}\left(1-\mu\frac{1-\nu^{-1/N}}{\eta+1-\nu^{-1/N}}\right)^N &= \exp\left(\lim_{N\to\infty}\log\left(1-\mu\frac{1-\nu^{-1/N}}{\eta+1-\nu^{-1/N}}\right)/(1/N)\right) \\
&= \exp\left(\lim_{N\to\infty}\frac{\eta\mu\log(\nu)}{(\nu^{-1/N}-\eta-1)(\nu^{1/N}(\eta-\mu+1)+\mu-1)}\right) \\
&= \exp\left(-\frac{\mu\log(\nu)}{\eta}\right) \\
&= (1/\nu)^{\mu/\eta}.
\end{aligned}$$

For the first limit, we set $\mu = 1$ to get

$$\lim_{N\to\infty}(1-\beta_{*,N})^N = (1/\nu)^{1/\eta}.$$

For the second limit, we set $\mu = 1-\eta$ to get

$$\lim_{N\to\infty}(1+(\eta-1)\beta_{*,N})^N = (1/\nu)^{\frac{1-\eta}{\eta}}.$$

C.8. Denoiser Error

Assumption 2 places a condition directly on the approximation of $\nabla f(x)$, where $f(x) := \frac{1}{2}\text{dist}_{\mathcal{K}}(x)$, that is jointly obtained from σ_t and the denoiser ϵ_θ. We prove this assumption holds under a direct assumption on $\nabla\text{dist}_{\mathcal{K}}(x)$, which is easier to verify in practice.

Assumption 3. *There exists $\nu \geq 1$ and $\eta > 0$ such that if $\frac{1}{\nu}\text{dist}_{\mathcal{K}}(x) \leq \sqrt{n}\sigma_t \leq \nu\text{dist}_{\mathcal{K}}(x)$ then $\|\epsilon_\theta(x,t) - \sqrt{n}\nabla\text{dist}_{\mathcal{K}}(x)\| \leq \eta$*

Lemma C.4. *If Assumption 3 holds with (ν, η), then Assumption 2 holds with $(\hat{\nu}, \hat{\eta})$, where $\hat{\eta} = \frac{1}{\sqrt{n}}\eta\nu + \max(\nu - 1, 1 - \frac{1}{\nu})$ and $\hat{\nu} = \nu$.*

Proof. Multiplying the error-bound on ϵ_θ by σ_t and using $\sqrt{n}\sigma_t \leq \nu\text{dist}_{\mathcal{K}}(x)$ gives

$$\|\sigma_t\epsilon_\theta(x,t) - \sqrt{n}\sigma_t\nabla\text{dist}_{\mathcal{K}}(x)\| \leq \eta\sigma_t \leq \eta\nu\frac{1}{\sqrt{n}}\text{dist}_{\mathcal{K}}(x)$$

Defining $C = \sqrt{n}\sigma_t - \text{dist}_{\mathcal{K}}(x)$ and simplifying gives

$$\begin{aligned}\eta\nu\frac{1}{\sqrt{n}}\text{dist}_{\mathcal{K}}(x) &\geq \|\sigma_t\epsilon_\theta(x,t) - \sqrt{n}\sigma_t\nabla\text{dist}_{\mathcal{K}}(x)\| \\ &= \|\sigma_t\epsilon_\theta(x,t) - \nabla f(x) - C\nabla\text{dist}_{\mathcal{K}}(x)\| \\ &\geq \|\sigma_t\epsilon_\theta(x,t) - \nabla f(x)\| - \|C\nabla\text{dist}_{\mathcal{K}}(x)\| \\ &= \|\sigma_t\epsilon_\theta(x,t) - \nabla f(x)\| - |C|\end{aligned}$$

Since $(\frac{1}{\nu} - 1)\text{dist}_{\mathcal{K}}(x) \leq C \leq (\nu - 1)\text{dist}_{\mathcal{K}}(x)$ and $\nu \geq 1$, the Assumption 2 error bound holds for the claimed $\hat{\eta}$. □

D. Derivation of Gradient Estimation Sampler

To choose W, we make two assumptions on the denoising error: the coordinates $e_t(\epsilon)_i$ and $e_t(\epsilon)_j$ are uncorrelated for all $i \neq j$, and $e_t(\epsilon)_i$ is only correlated with $e_{t+1}(\epsilon)_i$ for all i. In other words, we consider W of the form

$$W = \begin{bmatrix} aI & bI \\ bI & cI \end{bmatrix} \tag{24}$$

and next show that this choice leads to a simple rule for selecting $\bar{\epsilon}$. From the optimality conditions of the quadratic optimization problem (11), we get that

$$\bar{\epsilon}_t = \frac{a+b}{a+c+2b}\epsilon_\theta(x_t, \sigma_t) + \frac{c+b}{a+c+2b}\epsilon_\theta(x_{t+1}, \sigma_{t+1}).$$

Setting $\gamma = \frac{a+b}{a+c+2b}$, we get the update rule (12). When $b \geq 0$, the minimizer $\bar{\epsilon}_t$ is a simple convex combination of denoiser outputs. When $b < 0$, we can have $\gamma < 0$ or $\gamma > 1$, i.e., the weights in (12) can be negative (but still sum to 1). Negativity of the weights can be interpreted as cancelling positively correlated error ($b < 0$) in the denoiser outputs. Also note we can implicitly search over W by directly searching for γ.

E. Further Experiments

E.1. Denoising Approximates Projection

We test our interpretation that denoising approximates projection on pretrained diffusion models on the CIFAR-10 dataset. In these experiments, we take a 50-step DDIM sampling trajectory, extract $\epsilon(x_t, \sigma_t)$ for each t and compute the cosine similarity for every pair of $t, t' \in [1, 50]$. The results are plotted in Figure 7. They show that the direction of $\epsilon(x_t, \sigma_t)$ over the entire sampling trajectory is close to the first step's output $\epsilon(x_N, \sigma_N)$. On average over 1000 trajectories, the minimum similarity (typically between the first step when $t = 50$ and last step when $t' = 1$) is 0.85, and for the vast majority (over 80%) of pairs the similarity is > 0.99, showing that the denoiser outputs approximately align in the same direction, validating our intuitive picture in Figure 1.

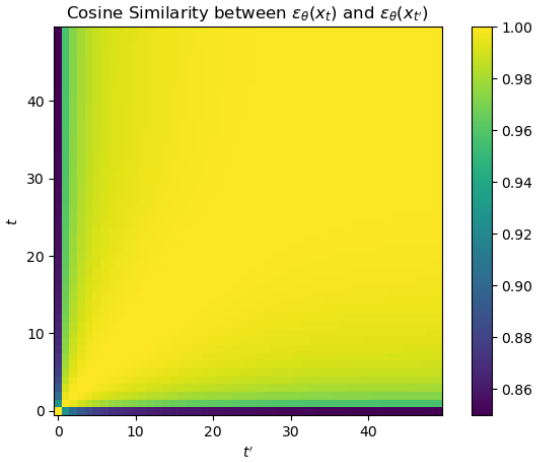


Figure 7: Plot of the cosine similarity between $\epsilon_\theta(x_t, t)$ and $\epsilon_\theta(x_{t'}, t')$ over $N = 50$ steps of DDIM denoising on the CIFAR-10 dataset. Each cell is the average result of 1000 runs.

(a) Plot of $\|\epsilon_\theta(x_t, \sigma_t)\| / \sqrt{n}$ against t.

(b) Plot of $\|\epsilon_\theta(x_0 + \sigma_t \epsilon, \sigma_t) - \epsilon\| / \sqrt{n}$ against t.

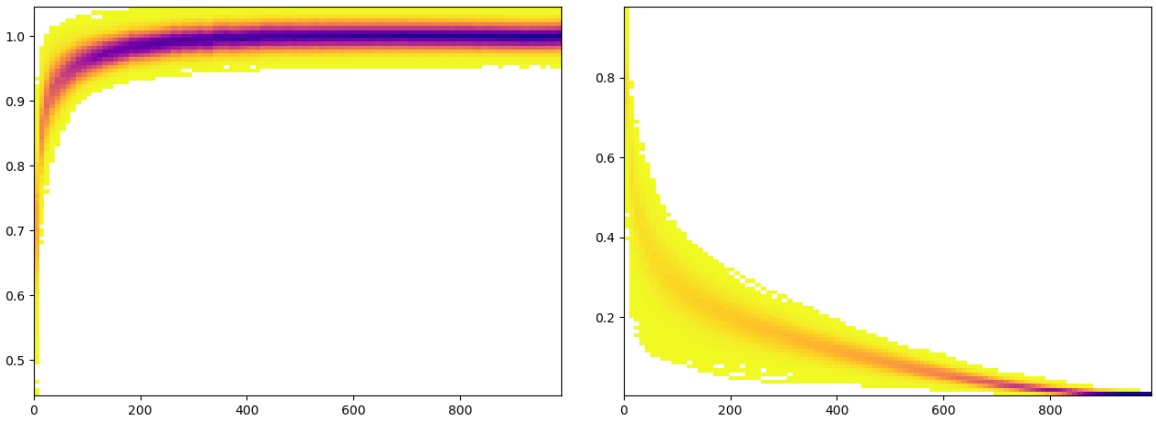


Figure 8: Plots of the norm of the denoiser at different stages of denoising, as well as the ability of the denoiser to accurately predict the added noise as a function of noise added.

	CIFAR-10	CelebA
DDIM (w/o both)	16.86	18.08
Ours (w/o sampler)	13.25	13.55
Ours (w/o schedule)	8.30	11.87
Ours (with both)	**3.85**	**4.30**

Table 3: Ablation study of the effects of the schedule improvements in Section 4.3 and the sampler improvements in Section 5, for $N = 10$ steps.

	CIFAR-10 FID				CelebA FID			
DDIM Sampler	$N = 5$	$N = 10$	$N = 20$	$N = 50$	$N = 5$	$N = 10$	$N = 20$	$N = 50$
DDIM	47.21	16.86	8.28	4.81	32.21	18.08	11.81	7.39
DDIM Offset	36.09	14.19	7.51	4.69	27.79	15.38	10.05	6.80
EDM	61.63	20.85	9.25	5.39	33.00	16.72	9.78	6.53
Ours (Log-linear)	40.33	13.37	6.88	4.71	28.07	13.63	8.80	6.79
	CIFAR-10 FID				CelebA FID			
Our Sampler	$N = 5$	$N = 10$	$N = 20$	$N = 50$	$N = 5$	$N = 10$	$N = 20$	$N = 50$
DDIM	51.65	8.30	4.96	3.33	28.64	11.87	8.00	4.33
DDIM Offset	14.95	7.50	4.58	3.29	12.19	9.49	6.58	4.80
EDM	34.26	4.87	3.64	3.67	18.68	5.30	3.95	4.11
Ours (Log-linear)	12.57	3.79	3.32	3.41	10.76	4.79	4.57	5.01

Table 4: Ablation of σ_t schedules for both the DDIM and GE sampler.

We perform an ablation study on different sampling schedules. We use the four different schedules as shown in Table 1:

- **DDIM** Default DDIM schedule with $\sigma_N = 157, \sigma_0 = 0.002$
- **DDIM Offset** Truncated DDIM schedule starting with a smaller σ, with $\sigma_N = 40, \sigma_0 = 0.002$.
- **EDM** Schedule used in (Karras et al., 2022) with $\sigma_N = 80, \sigma_0 = 0.002$.
- **Linear** Log-linear schedule with $\sigma_N = 40$, σ_1, σ_0 selected based on Appendix F.3.

Our results are reported in Table 4. Our gradient-estimation sampler consistently outperforms the DDIM sampler for all schedules and N. The *DDIM Offset* schedule that starts at $\sigma_N = 40$ offers an improvement over the *DDIM* schedule for $N = 5, 10, 20$, but performs worse for $N = 50$. This suggests starting from a higher σ_N for larger N, which we have done in our final evaluations.

E.2. Distance Function Properties

We test Assumption 1 and Assumption 2 on pretrained networks. If Assumption 1 is true, then $\|\epsilon_\theta(x_t, \sigma_t)\| \sqrt{n} = \|\nabla \text{dist}_{\mathcal{K}}(x_t)\| = 1$ for every x_t along the DDIM trajectory. In Figure 8a, we plot the distribution of norm of the denoiser $\epsilon_\theta(x_t, \sigma_t)$ over the course of many runs of the DDIM sampler on the CIFAR-10 model for $N = 100$ steps ($t = 1000, 990, \ldots, 20, 10, 0$). This plot shows that $\|\epsilon_\theta(x_t, \sigma_t)\| / \sqrt{n}$ stays approximately constant and is close to 1 until the end of the sampling process. We next test Assumption 3, which implies Assumption 2 by Lemma C.4. We do this by first sampling a fixed noise vector ϵ, next adding different levels of noise σ_t, then using the denoiser to predict $\epsilon_\theta(x_0 + \sigma_t \epsilon, \sigma_t)$. In Figure 8b, we plot the distribution of $\|\epsilon_\theta(x_0 + \sigma_t \epsilon, \sigma_t) - \epsilon\| / \sqrt{n}$ over different levels of t, as a measure of how well the denoiser predicts the added noise.

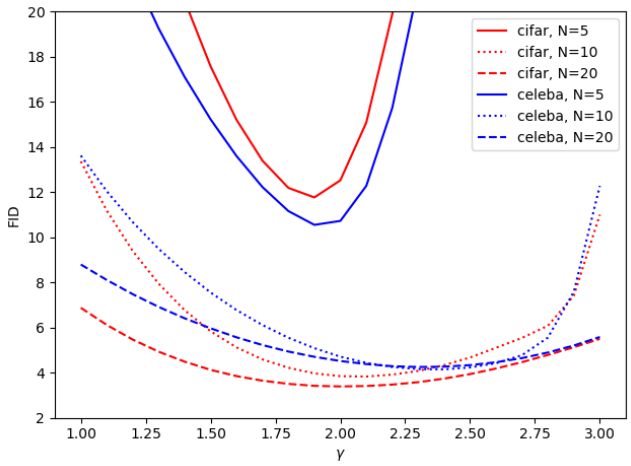


Figure 9: Plot of FID score against γ for our second-order sampling algorithm on the CIFAR-10 and CelebA datasets for $N = 5, 10, 20$ steps.

E.3. Choice of γ

We motivate our choice of $\gamma = 2$ in Algorithm 2 with the following experiment. For varying γ, Figure 9 reports FID scores of our sampler on the CIFAR-10 and CelebA models for $N = 5, 10, 20$ timesteps using the σ_t schedule described in Appendix F.3. As shown, $\gamma \approx 2$ achieves the optimal FID score over different datasets for $N < 20$. For sampling from the CelebA dataset, we found that setting $\gamma = 2.4$ for $N = 20$ and $\gamma = 2.8$ for $N = 50$ achieves the best FID results.

F. Experiment Details

F.1. Pretrained Models

The CIFAR-10 model and architecture were based on that in (Ho et al., 2020), and the CelebA model and architecture were based on that in (Song et al., 2020a). The specific checkpoints we use are provided by (Liu et al., 2022). We also use Stable Diffusion 2.1 provided in `https://huggingface.co/stabilityai/stable-diffusion-2-1`. For the comparison experiments in Figure 3, we implemented our gradient estimation sampler to interface with the HuggingFace diffusers library and use the corresponding implementations of UniPC, DPM++, PNDM and DDIM samplers with default parameters.

F.2. FID Score Calculation

For the CIFAR-10 and CelebA experiments, we generate 50000 images using our sampler and calculate the FID score using the library in `https://github.com/mseitzer/pytorch-fid`. The statistics on the training dataset were obtained from the files provided by (Liu et al., 2022). For the MS-COCO experiments, we generated images from 30k text captions drawn from the validation set, and computed FID with respect to the 30k corresponding images.

F.3. Our Selection of σ_t

Let $\sigma_1^{\text{DDIM}(N)}$ be the noise level at $t = 1$ for the DDIM sampler with N steps. For the CIFAR-10 and CelebA models, we choose $\sigma_1 = \sqrt{\sigma_1^{\text{DDIM}(N)}}$ and $\sigma_0 = 0.01$. For CIFAR-10 $N = 5, 10, 20, 50$ we choose $\sigma_N = 40$ and for CelebA $N = 5, 10, 20, 50$ we choose $\sigma_N = 40, 80, 100, 120$ respectively. For Stable Diffusion, we use the same sigma schedule as that in DDIM.

F.4. Text Prompts

For the text to image generation in Figure 3, the text prompts used are:

- “A digital Illustration of the Babel tower, 4k, detailed, trending in artstation, fantasy vivid colors”
- “London luxurious interior living-room, light walls”
- “Cluttered house in the woods, anime, oil painting, high resolution, cottagecore, ghibli inspired, 4k”