1.5em
UNDERSTANDING DETERMINISTIC DIFFUSION THROUGH REVERSE TRANSITION KERNELS

Adrita Das, Peiran Jiang, Barnabás Póczos & Jose Lugo-Martinez
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{adritad,peiranj,bapoczos,jlugomar}@cs.cmu.edu

ABSTRACT

Diffusion models have become a leading paradigm for generative modeling across visual and scientific domains, but their deployment is limited by the high computational cost of sampling, which typically requires long stochastic trajectories and sequential updates. This challenge is particularly severe in high-dimensional settings such as molecular generation and image synthesis. In this work, we provide a principled reinterpretation of deterministic diffusion models through the Reverse Transition Kernel (RTK) framework Huang et al. (2024), showing that the reverse process induces structured and well-conditioned subproblems. We demonstrate that these can be viewed as approximately strongly log-concave optimization problems, enabling stable updates and constant step sizes. This perspective unifies deterministic and stochastic diffusion while providing a pathway toward efficient and scalable sampling. Empirically, we validate our theoretical insights by measuring regularity properties of the learned denoiser. Our method achieves faster convergence and improved structural fidelity on molecular benchmarks while preserving chemical validity, and yields consistent gains in stability and sample quality on image generation tasks.

1 INTRODUCTION

Diffusion models and flow-based methods (Albergo et al. (2025); Albergo & Vanden-Eijnden (2023); Lipman et al. (2023)) have emerged as a powerful framework for generative modeling, achieving state-of-the-art performance across domains such as image, molecular, and text generation (Dhariwal & Nichol (2021); Austin et al. (2021); Ramesh et al. (2022)) as well as scientific data modeling (Trippe et al. (2023)). By learning to reverse a gradual noising process, these models capture complex high-dimensional distributions with high fidelity. However, their practical deployment is limited by computationally expensive sampling procedures, often requiring hundreds to thousands of sequential denoising steps. Despite diverse formulations—including DDPMs (Ho et al. (2020)), score-based models (Song & Ermon (2020a)), Schrödinger bridges (Bortoli et al. (2023)), and flow matching (Lipman et al. (2023))—these approaches share a common principle: constructing a stochastic or deterministic trajectory that transforms a simple reference distribution into the target distribution via learned reverse dynamics.

To address sampling inefficiency, prior work has explored deterministic formulations such as probability flow ODEs (Chen et al. (2023b)) and DDIMs (Song et al. (2022)), along with distillation (Luhman & Luhman (2021); Salimans & Ho (2022)) and consistency-based methods (Heek et al. (2024); Lu & Song (2025)). While effective in reducing sampling cost, these approaches are often based on heuristic approximations and lack a discrete-time perspective explaining their stability under finite step sizes. *In particular, it remains unclear why deterministic denoising updates, even without stochasticity, can yield stable and accurate samples. In this work, we reinterpret DDDM (Zhang et al. (2024)) through the Reverse Transition Kernel (RTK) framework (Huang et al. (2024)), providing a unified probabilistic view of deterministic and stochastic diffusion. We show that, under mild regularity conditions on the denoiser, the induced proxy energies are strongly log-concave, leading to well-conditioned and stable reverse dynamics.* We support our analysis with empirical evaluation on trained models, measuring Jacobian spectral norms, residuals, and curvature properties.

1.1 CONTRIBUTIONS

Our main contributions are as follows:

- We introduce a novel reinterpretation of deterministic diffusion models through the RTK framework, unifying stochastic and deterministic formulations.
- We provide a second-order analysis of the reverse process, establishing strong log-concavity of the induced subproblems under mild conditions on the denoiser.
- We show that this structure enables stable, constant step-size sampling, thereby reducing the need for long stochastic trajectories.
- We empirically validate the theoretical assumptions by analyzing trained models and measuring curvature-related quantities.

2 RELATED WORK

In terms of dimensional dependence, the complexity of diffusion-based sampling has been progressively improved in recent work. For SDE-based formulations, the current state-of-the-art achieves a complexity of $\tilde{O}(d)$ (Benton et al. (2024)), improving upon earlier $\tilde{O}(d^2)$ bounds (Chen et al. (2023a)). For probability flow ODE formulations, a sharper $\tilde{O}(\sqrt{d})$ complexity has been established using predictor–corrector schemes combined with underdamped Langevin Monte Carlo (UMLC) (Chen et al. (2023c)). Chen et al. (2024) proposed a principled framework for accelerating diffusion model inference via parallelization and rigorous analysis. They identify discretization error as a key bottleneck, which forces step sizes to scale as $\tilde{O}(1/d)$ for SDEs and $\tilde{O}(1/\sqrt{d})$ for probability flow ODEs, leading to polynomial complexity in d. To overcome this, they partition the reverse process into a constant number of blocks ($O(1)$ in d) with parallelizable score evaluations, enabling constant step sizes while preserving accuracy. This results in an overall complexity of $\tilde{O}(\mathrm{poly}\log d)$, significantly improving scalability for high-dimensional sampling. Huang et al. (2024) reinterpreted diffusion inference as a sequence of reverse transition kernel (RTK) subproblems, where conventional methods such as DDPM and DDIM correspond to fine-grained decompositions with many steps. They propose a generalized RTK framework that reduces the number of subproblems to $\tilde{O}(1)$ by constructing strongly log-concave targets, enabling efficient sampling via MALA and underdamped Langevin dynamics (ULD). Their approach provides improved convergence guarantees, achieving faster rates in total variation distance compared to prior diffusion-based methods.

3 NOTATIONS & PRELIMINARIES

3.1 DIFFUSION MODELS

DDPMs Diffusion models (DDPMs) (Ho et al. (2020); Sohl-Dickstein et al. (2015)) are generative models that learn data distributions via a forward process (adding noise) and a reverse process (removing noise). Let $\{\beta_i\}_{i=1}^T$ denote a sequence of positive noise scales such that $0 < \beta_1, \beta_2, \ldots, \beta_T < 1$. The forward process gradually adds Gaussian noise to the data x_0 over time steps t in a Markov chain: $p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative noise schedule and $x_0 \sim p^* \propto \exp(-f^*)$ is a sample from the true data distribution and $f^*(x)$ is the corresponding energy function. We consider a discrete-time Markov chain $x_0, x_1, \ldots, x_T$. The perturbed data distribution is denoted as $p_\alpha(\hat{\mathbf{x}}) := \int p^*(\mathbf{x})\, p_\alpha(\hat{\mathbf{x}} \mid \mathbf{x})\, d\mathbf{x}$. A variational Markov chain in the reverse direction is parameterized as $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\Big(\mathbf{x}_{t-1}; \frac{1}{\sqrt{1-\beta_t}}\big(\mathbf{x}_t + \beta_t\, s_\theta(\mathbf{x}_t, t)\big),\, \beta_t \mathbf{I}\Big)$. Samples are produced via an ancestral sampling procedure, following $\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_i}}\left(\mathbf{x}_t + \beta_t\, s_{\theta^*}(\mathbf{x}_t, t)\right) + \sqrt{\beta_t}\,\mathbf{z}_t$. Here, $s_\theta(x_t, t)$ is the score function parameterized by θ, which predicts the noise component in x_t. The optimal parameters θ^* are learned by minimizing the expected denoising error between the predicted and true noise over all timesteps.

Implicit Models The induced joint distribution over latent variables is given by $p(x_{1:T} \mid x_0) := p(x_T \mid x_0) \prod_{t=2}^{T} p(x_{t-1} \mid x_t, x_0)$. The terminal marginal distribution is specified as $p(x_T \mid x_0) =$

$\mathcal{N}(\sqrt{\bar{\alpha}_T}\, x_0, (1-\bar{\alpha}_T)\mathbf{I})$, where $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$. For all $t > 1$, the backward kernels are defined as $p(x_{t-1} \mid x_t, x_0) = \mathcal{N}\Big(\sqrt{\bar{\alpha}_{t-1}}\, x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\beta_t}\, \frac{x_t - \sqrt{\bar{\alpha}_t}\, x_0}{\sqrt{1-\bar{\alpha}_t}},\ \beta_t\, \mathbf{I}\Big)$. The mean functions are constructed such that the induced marginals satisfy $p(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\mathbf{I})$ for all t, ensuring consistency across time steps. Applying Bayes' rule yields the induced forward transition $p(x_t \mid x_{t-1}, x_0) = \frac{p(x_{t-1}|x_t, x_0)\, p(x_t|x_0)}{p(x_{t-1}|x_0)}$. Song et al. (2022) defined an oracle reverse-time inference process conditioned on the clean data sample x_0, which serves as a theoretical reference distribution rather than a directly implementable sampling algorithm.

Remark (Implicit Forward Process). *The forward transitions in implicit diffusion models do not correspond to the Markovian noising process used in DDPMs. Instead, they arise from an oracle-based inference construction and are therefore generally non-Markovian, with explicit dependence on the clean sample x_0. A tractable generative procedure is obtained by replacing this oracle dependence with a learned prediction $\hat{x}_0 = f_\theta(x_t)$, yielding a reverse-time process consistent with the DDIM parameterization. In this setting, choosing $\sigma_t^2 = \beta_t$ corresponds to the stochastic formulation, while the limit $\sigma_t \to 0$ recovers the deterministic (implicit) case.*

Directly Denoising Diffusion. From (Song & Ermon (2020b)) reformulation of DDPMs as Stochastic Differential Equations (SDEs) to generalize the forward process: $d\mathbf{X}_t = -\frac{1}{2}\beta(t)\mathbf{X}_t\, dt + \sqrt{\beta(t)}\, d\mathbf{B}_t$, where $\mathbf{B}_t$ represents Brownian motion. The reverse-time VP SDE is: $d\mathbf{X}_t = \big[-\frac{1}{2}\beta(t)(\mathbf{X}_t + \nabla \log p_t(\mathbf{X}_t))\big]\, dt + \sqrt{\beta(t)}\, d\mathbf{B}_t$, where $\nabla \log p_t$ is estimated by a time-dependent score network $s_\theta(\mathbf{x}, t)$. Directly Denoising Diffusion Models (DDDMs) (Zhang et al. (2024)) integrate DDPMs with the Probability Flow (PF) ODE framework, enabling faster denoising without complex solvers. The solution of the PF ODE is obtained by evaluating the integral expression $\mathbf{x}_0 = \mathbf{x}_T + \int_0^T -\frac{1}{2}\beta(t)[\mathbf{x}_t + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)]\, dt$, where $\mathbf{x}_T \sim \mathcal{N}(0, I)$. Directly Denoising Diffusion Models (DDDM) refine the estimate of the clean state $\mathbf{x}_0$ by leveraging the probability flow ODE. Specifically, the mapping $f(\mathbf{x}_0, \mathbf{x}_t, t) = \mathbf{x}_t - F(\mathbf{x}_0, \mathbf{x}_t, t)$ is defined, where F involves an integral of the drift term parameterized by the noise schedule $\beta(t)$. A neural approximation $f_\theta(\mathbf{x}_0, \mathbf{x}_t, t) = \mathbf{x}_t - F_\theta(\mathbf{x}_0, \mathbf{x}_t, t)$ is trained such that $f_\theta \approx f$.

$$\Theta := \arg\min_\theta\ \mathbb{E}_{t\sim\mathcal{U}[1,T]}\left[\mathbb{E}_{\mathbf{x}_0\sim p_{\text{data}}(\mathbf{x}_0)}\left[\mathbb{E}_{\mathbf{x}_t\sim\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0,(1-\bar{\alpha}_t)\mathbf{I}\right)}\left[d\big(f_\theta(\mathbf{x}_0^{(n)}, \mathbf{x}_t, t), \mathbf{x}_0\big)\right]\right]\right] \quad (1)$$

where, $d(\cdot,\cdot)$ is a suitable distance metric. We define Θ as the set of optimal neural parameters for the denoising network.

Definition 3.1 (Time Complexity). *In diffusion models, the primary computational bottleneck lies in the inference process due to sequential evaluations of the learned score (or denoising) function $s_\theta(x_t, t)$ (or F_θ). We define the approximate time complexity as the number T of non-parallelizable evaluations of this function along the reverse trajectory $\{x_t\}_{t=0}^T$, i.e., the total number of sequential steps required for sampling. This notion aligns with standard measures such as iteration complexity, where T denotes the number of iterations needed to approximate the target distribution, and typically scales with the data dimension d, e.g., $T = \tilde{O}(d^\alpha)$ for some $\alpha > 0$.*

OR

Definition 3.2 (Computational Efficiency of Diffusion Sampling). *The computational efficiency of diffusion models is fundamentally tied to the problem of determining how many discretization steps T and score function evaluations $s_\theta(x_t, t)$ are required to approximate a target data distribution $p_{data} \subset \mathbb{R}^d$ up to a prescribed accuracy $\delta > 0$. This question has been extensively studied in the literature, where convergence guarantees are typically characterized in terms of the number of iterations T needed such that the generated distribution p_T satisfies $D(p_T, p_{data}) \le \delta$ under an appropriate probability metric $D(\cdot,\cdot)$ (e.g., total variation or Wasserstein distance). Existing results show that this complexity often scales polynomially with the dimension d, i.e., $T = \tilde{O}(d^\alpha)$ for some $\alpha > 0$, reflecting the intrinsic difficulty of simulating high-dimensional stochastic dynamics with controlled error.*

Definition 3.3 (Reverse Transition Kernel (RTK) Framework): RTK is a flexible approach for accelerating diffusion inference by decomposing the reverse diffusion process into a small number of reverse transition kernel (RTK) sampling subproblems. Under an appropriate decomposition, the number of such subproblems can be reduced to approximately $\tilde{O}(1)$, independent of the diffusion discretization resolution, while ensuring that each RTK target distribution remains strongly log-concave. **Strong log-concavity** implies that each target density $p(x) \propto \exp(-U(x))$ admits a strongly convex potential $U(x)$, yielding a unique global mode and well-conditioned energy landscape. This enables fast mixing and provable convergence guarantees for sampling algorithms such as Langevin dynamics and Metropolis–Hastings.

Algorithm 1: Reverse Diffusion via RTKs

Setup: Step size $\eta > 0$, horizon $T = K\eta$.
for $k = 0$ **to** $K - 1$ **do**
  Sample $\hat{x}_{(k+1)\eta} \sim \hat{p}_{(k+1)\eta|k\eta}(\cdot \mid \hat{x}_{k\eta})$.
  Assume marginal consistency: $\hat{p}_{k\eta} \approx p_{(K-k)\eta}$.
  Update marginal:

$$\hat{p}_{(k+1)\eta}(z) \approx \int p^{\leftarrow}_{(k+1)\eta|k\eta}(z \mid x)\, \hat{p}_{k\eta}(x)\, dx.$$

end for
return $\hat{X}_{K\eta}$

This perspective implies that each reverse step admits a unique solution, enables stable constant step-size updates, and recasts sampling as a sequence of well-structured deterministic subproblems, improving efficiency over traditional stochastic diffusion.

Theorem 3.4 (Prékopa's theorem). If the joint density $q(z, x)$ is log-concave in the joint variables (z, x), then the marginal density $q(z) = \int q(z, x)\, dx$, obtained by integrating out x, is also log-concave in z.

Definition 3.5 (Log-concavity). A function $f : \mathbb{R}^d \to \mathbb{R}_+$ is *log-concave* if $\log f$ is concave. Equivalently, for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \;\geq\; f(x)^{\lambda} f(y)^{1-\lambda}.$$

Definition 3.6 (Strong convexity). A twice-differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ is m-*strongly convex* if its Hessian satisfies

$$\nabla^2 g(z) \succeq m I_d \qquad (m > 0).$$

If $p(z) \propto \exp(-g(z))$, then p is m-strongly log-concave.

Definition 3.7 (Jacobian) A differentiable map $F : \mathbb{R}^d \to \mathbb{R}^d$ has Jacobian $J_F(z) = \nabla_z F(z)$ with entries $[J_F(z)]_{ij} = \partial F_i(z)/\partial z_j$.

Definition 3.8 (Hessian) A twice-differentiable scalar function $g : \mathbb{R}^d \to \mathbb{R}$ has Hessian $\nabla^2 g(z) = [\partial^2 g(z)/(\partial z_i \partial z_j)]_{i,j=1}^d$. For a vector-valued map F, the Hessian of its ith coordinate is $H_{F,i}(z) = \nabla^2 F_i(z)$.

Definition 3.9 (Spectral and operator norms). For a matrix $A \in \mathbb{R}^{d\times d}$, the spectral norm is $\|A\|_2 := \sup_{\|v\|=1} \|Av\|$. For a second-derivative tensor $D^2 F(z)$, viewed as a symmetric bilinear map, the operator norm is defined as $\|D^2 F(z)\|_{\mathrm{op}} := \sup_{\|u\|=1,\ \|v\|=1} \|D^2 F(z)[u, v]\|$.

Definition 4.0 (Local Lipschitz constant). The local Lipschitz constant of F at a point z is $L(z) = \|J_F(z)\|_2$.

Definition 4.1 (Residual). The residual in the reverse update is $r(z) = z - (x - F(z))$ with magnitude $R(z) = \|r(z)\|$.

3.2 Diffusion as Approximate Sampling of Transition Kernels

Under this perspective, the denoising diffusion process is viewed as a sequence of sampling subproblems, where computational efficiency is governed by the trade-off between the number of subproblems and the difficulty of solving each one. DDPMs approximate each reverse transition using simple Gaussian kernels with $\mathcal{O}(1)$ per-step cost, but require $\tilde{\mathcal{O}}(d\,\varepsilon^{-2})$ such steps or subproblems to achieve ε-accuracy in total variation (TV) distance shown by (Chen et al. (2023c); Benton et al. (2024)). In contrast, the Directly Denoising Diffusion Model (DDDM) can be interpreted within the Reverse Transition Kernel (RTK) framework as collapsing many small stochastic subproblems into a small number of deterministic kernel approximations. Rather than sampling from each intermediate reverse kernel, DDDM replaces stochastic transitions with a learned deterministic map that directly

approximates the conditional mean of the RTK target distribution. From this viewpoint, DDDM performs approximate sampling by implicitly optimizing a transport map between successive noisy states, effectively solving fewer but harder subproblems. The RTK formulation clarifies that this deterministic denoising corresponds to the zero-noise limit of a valid reverse transition kernel, explaining how DDDM achieves efficient inference while remaining consistent with an underlying probabilistic model. Unlike stochastic diffusion models such as DDPMs and DDIMs, DDDM does not explicitly construct a reverse-time trajectory over intermediate noise levels. Instead, it performs iterative refinement at a fixed noise level.

4 Problem Setup

In this section, we reformulate diffusion inference through the lens of reverse transition kernels (RTKs) and provide a principled connection between probability flow ODEs, Directly Denoising Diffusion Models (DDDMs), and kernel-based reverse diffusion. We begin by reviewing the probability flow ODE formulation, which yields a deterministic reverse-time dynamics governed by the score function. We show that an Euler discretization of this ODE naturally induces a reverse-time subproblem parameterized by a step size η.

4.1 Probability Flow ODE and Deterministic Reverse Diffusion

Diffusion-based generative models are commonly formulated through a stochastic forward process that gradually perturbs data with noise, together with a reverse-time process that removes noise to recover samples from the data distribution. Following the stochastic differential equation (SDE) formulation of diffusion models introduced by (Song et al. (2021)), the variance-preserving (VP) forward diffusion process is given by

$$dX_t = -\frac{1}{2}\beta(t)X_t\,dt + \sqrt{\beta(t)}\,dB_t, \tag{2}$$

where $\beta(t)$ is a time-dependent noise schedule and B_t denotes standard Brownian motion. The corresponding reverse-time SDE takes the form

$$dX_t = \left[-\frac{1}{2}\beta(t)\left(X_t + \nabla_x \log p_t(X_t)\right)\right] dt + \sqrt{\beta(t)}\,dB_t, \tag{3}$$

where $\nabla_x \log p_t(x)$ denotes the score function of the marginal distribution at time t.

(Song et al., 2021) further showed that the reverse-time SDE admits an equivalent *deterministic* formulation, known as the *probability flow ordinary differential equation (PF ODE)*, which shares the same time-marginal distributions as the stochastic reverse process. The PF ODE is given by

$$d\overleftarrow{x}_t = -\frac{1}{2}\beta(t)\left[\overleftarrow{x}_t - \nabla_x \log p_t(\overleftarrow{x}_t)\right] dt. \tag{4}$$

Let $\{\hat{p}_{k\eta}\}_{k=0}^{K}$ denote the marginal distributions of the reverse-time process. The process is initialized from a simple prior $\hat{p}_0 = \mathcal{N}(0, I)$ and evolves such that the terminal distribution $\hat{p}_{K\eta}$ approximates the target data distribution p^*. Accordingly, the initial state is sampled as $\hat{x}_0 \sim \mathcal{N}(0, I)$. The marginal distributions of the forward and reverse processes are related through the identity

$$p_{T-t} = \overleftarrow{p}_t. \tag{5}$$

In practice, the score function $\nabla_x \log p_t(x)$ is unknown and is approximated using a time-dependent neural network $s_\theta(x, t)$. Substituting this approximation into equation 4 yields a neural ordinary differential equation that governs deterministic reverse-time inference. Over a discrete time interval $t \in [k\eta, (k+1)\eta]$, the neural ODE can be written as

$$d\bar{x}_t = -\frac{1}{2}\beta(t)\left[\bar{x}_t + s_{\theta,\,T-k\eta}(\bar{x}_{k\eta})\right] dt. \tag{6}$$

4.2 Euler Discretization and Directly Denoising Diffusion Models

Let T denote the total diffusion horizon and let $\eta > 0$ be a step size such that $K = T/\eta$. Applying a forward Euler discretization to the probability flow ODE yields the approximation

$$\hat{x}_0 \approx x_{K\eta} - \sum_{k=0}^{K-1} \eta \cdot \frac{1}{2}\,\beta(k\eta)\left[x_{k\eta} - \nabla_x \log p_{k\eta}(x_{k\eta})\right] \tag{7}$$

Smaller step sizes η lead to simpler local subproblems but require a larger number of steps, whereas larger step sizes reduce the number of steps at the cost of more challenging local approximations.

Directly Denoising Diffusion Models (DDDMs) (Zhang et al., 2024) are built on the discretized probability flow formulation. Instead of simulating the stochastic reverse-time SDE as in DDPMs, DDDMs perform deterministic denoising by iteratively refining an estimate of the clean sample.

At a given reverse iteration k, starting from the noisy state $x_{k\eta}$, DDDM introduces an inner iterative refinement procedure to estimate x_0:

$$\hat{x}_0^{(m+1)} = x_{k\eta} - F_{\theta,k\eta}\Big(\hat{x}_0^{(m)}, x_{k\eta}\Big), \quad m = 0, \dots, M-1, \tag{8}$$

where $\hat{x}_0^{(m)}$ denotes the estimate of the clean sample at iteration m, and $F_{\theta,k\eta}$ is a neural approximation of the integrated drift term induced by the probability flow ODE.

Proposition 1 (Deterministic Reverse Kernel and Energy-Based Formulation of DDDM)**.** *Consider the DDDM reverse update at step k, where an estimate of the clean sample is obtained via the iterative refinement according to Eq. 8 The resulting update induces a deterministic reverse transition kernel that can be expressed as a Dirac measure concentrated at the output of the denoising map, i.e., $\overleftarrow{p}_{(k+1)\eta|k\eta}(z \mid \hat{x}_{k\eta}) = \delta\big(z - (x_{k\eta} - F_{\theta,k\eta}(\hat{x}_{k\eta}, x_{k\eta}))\big)$, where $\hat{x}_{k\eta}$ denotes the current estimate of the clean sample, i.e., $\hat{x}_{k\eta} \equiv \hat{x}_0^{(m)}$. Moreover, defining the proxy energy function*

$$g_{k\eta}(z) = \frac{1}{2\sigma^2}\left\|z - (x_{k\eta} - F_{\theta,k\eta}(z, x_{k\eta}))\right\|^2, \tag{9}$$

the corresponding reverse update admits an energy-based interpretation, where the induced proxy kernel satisfies $\hat{p}_{(k+1)\eta|k\eta}(z \mid \hat{x}_{k\eta}) \propto \exp(-g_{k\eta}(z))$. This formulation replaces an intractable sampling problem with a deterministic optimization objective, allowing each reverse step to be interpreted as mode-seeking in a proxy energy landscape. In particular, the DDDM update can be viewed as approximately computing a fixed point of the denoising map, or equivalently, as seeking a mode of the induced energy.

Note. The RTK-style indexing $(k, k+1)$ is used for interpretational convenience. DDDM does not explicitly construct a reverse diffusion trajectory; instead, it applies a deterministic denoising map iteratively at a fixed noise level.

4.3 Implications for Reverse-Time Inference

This reinterpretation of DDDM within the RTK framework provides a useful lens for analyzing the structure of the reverse-time inference process. In the following section, we investigate the extent to which, under mild regularity assumptions on the learned denoising maps $F_{\theta,k\eta}$, the proxy energy function $g(z)$ induces well-behaved subproblems. In particular, we examine conditions under which these subproblems exhibit approximate log-concavity, which is often associated with improved numerical stability and more favorable optimization landscapes.

From this perspective, the reverse diffusion process may be viewed as a sequence of structured subproblems that are potentially better conditioned than their stochastic counterparts. This viewpoint suggests the possibility of performing inference with relatively stable updates and reduced reliance on very small step sizes or long stochastic trajectories. However, the extent to which these properties hold in practice depends on the behavior of the learned denoiser.

Proposition 2 (Log-Concavity of Reverse Subproblems)**.** *DDDM reformulates the reverse process as a deterministic iterative refinement procedure, where each iteration corresponds to solving a structured reconstruction problem.*

Let $p(z) \propto \exp(-g(z))$ denote the target distribution of a reverse subproblem, where $g : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Consider the proxy energy

$$g(z) = \frac{1}{2\sigma^2}\left\|z - (x - F(z))\right\|^2, \tag{10}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable, and define the residual $r(z) := z - x + F(z)$. [1]

Then the gradient and Hessian of g are given by $\nabla g(z) = \frac{1}{\sigma^2}(I + J_F(z))^\top r(z)$ and $\nabla^2 g(z) = \frac{1}{\sigma^2}\Big[(I+J_F(z))^\top(I+J_F(z)) + \sum_{i=1}^d r_i(z)\, H_{F,i}(z)\Big]$, where $J_F(z)$ is the Jacobian of F and $H_{F,i}(z)$ is the Hessian of its i-th coordinate.

To characterize the local geometry, define $L(z) := \|J_F(z)\|_2$, $R(z) := \|r(z)\|$, and $B_{\mathrm{sq}}(z) := \sum_{i=1}^d \|H_{F,i}(z)\|_{\mathrm{op}}^2$. Then, for any $v \in \mathbb{R}^d$,

$$v^\top \nabla^2 g(z)\, v \;\geq\; \frac{1}{\sigma^2}\Big((1 - L(z))^2 - R(z)\, B_{\mathrm{sq}}(z)\Big)\, \|v\|^2. \tag{11}$$

Consequently, if $(1-L(z))^2 > R(z)\, B_{\mathrm{sq}}(z)$, then g is locally m-strongly convex with $m = \frac{1}{\sigma^2}\big((1-L(z))^2 - R(z)\, B_{\mathrm{sq}}(z)\big) > 0$, and hence the induced density $p(z) \propto \exp(-g(z))$ is locally strongly log-concave.

Theorem 1 (Admissibility of Constant Step-Size for DDDM)**.** *Consider the DDDM update at reverse-time step k with step-size $\eta > 0$, from Eq. 8 and the corresponding proxy energy given by Eq. 9, where $F_{\theta,k\eta} : \mathbb{R}^d \to \mathbb{R}^d$ is twice continuously differentiable on the region of interest. If there $\exists \kappa \in [0, 1)$ and $B \geq 0$ such that $\|J_F(z)\|_2 \leq \kappa$ and $\|D^2 F_{\theta,k\eta}(z)\|_{\mathrm{op}} \leq B$ for all $z \in \mathcal{Z}$, where $J_F(z) = \nabla_z F_{\theta,k\eta}(z)$ and $\|D^2 F\|_{\mathrm{op}}$ denotes an operator norm bound on the second derivative (tensor) action on unit vectors, then under the condition*

$$\frac{(1-\kappa)^2}{\sigma^2} > C_d B, \tag{12}$$

where d denotes the dimensionality of the diffusion state $z \in \mathbb{R}^d$, and the constant C_d scales polynomially with both d and the smoothness bound B, the function g is m-strongly convex on the region for some $m > 0$.

Consequently, the proxy density $\propto \exp(-g(z))$ is sharply concentrated about a unique minimizer, and the transition kernel is (exactly or asymptotically) degenerate/Dirac. The admissibility of a constant step-size $\eta = \Theta(1)$ remains valid as long as the parameterized mappings $F_{\theta,k\eta}$ obey the aforementioned uniform smoothness and boundedness conditions for the given η. We relegate the proof of this proposition to Appendix C.

Lemma 3.3 (Log-Concavity Preservation of Reverse Marginals). Let $p^{\leftarrow}_{(k+1)\eta|k\eta}(z \mid x)$ denote the reverse transition kernel and $p^{\leftarrow}_{k\eta}(x)$ the marginal at step k.

If:

- $p^{\leftarrow}_{(k+1)\eta|k\eta}(z \mid x)$ is jointly log-concave in (z, x), and
- $p^{\leftarrow}_{k\eta}(x)$ is log-concave in x,

then the next marginal

$$\hat{p}_{(k+1)\eta}(z) \;\approx\; \int p^{\leftarrow}_{(k+1)\eta|k\eta}(z \mid x)\, p^{\leftarrow}_{k\eta}(x)\, dx \tag{13}$$

is log-concave in z.

Remark. This result follows directly from Prékopa's theorem, which states that marginalization preserves log-concavity. As a consequence, log-concavity is propagated along the reverse trajectory, ensuring that each induced subproblem retains favorable geometric structure.

5 Discussion

The admissibility of a constant step size $\eta = \Theta(1)$ in DDDM plays a crucial role in enabling efficient diffusion inference. When the parameterized mappings $F_{\theta,k\eta}$ satisfy the uniform smoothness and

[1] For notational simplicity, we suppress the explicit dependence on the time index $k\eta$ and write $x := x_{k\eta}$ and $F(z) := F_{\theta,k\eta}(z, x_{k\eta})$. Under this shorthand, the proxy energy in Eq. (9) can be written as $g(z) = \frac{1}{2\sigma^2}\big\|z - \big(x - F(z)\big)\big\|^2$.

boundedness conditions, the associated proxy energy $g(z)$ becomes m-strongly convex, ensuring that the corresponding proxy density $\propto e^{-g(z)}$ is sharply concentrated around a unique minimizer. This strong convexity guarantees that each reverse-time update converges stably and effectively without requiring excessively small step sizes. Consequently, the denoising trajectory can be decomposed into only $\tilde{\mathcal{O}}(1)$ well-conditioned subproblems, as opposed to the $\tilde{\mathcal{O}}(1/\eta) = \tilde{\mathcal{O}}(1/\epsilon^2)$ subproblems required in standard DDPMs. This constant-step-size regime eliminates the need for finely discretized reverse diffusion schedules, substantially reducing the number of iterative updates while preserving numerical stability and accuracy. As a result, DDDM attains a provably faster inference process, enabling efficient sampling through a small number of strongly log-concave subproblems. In practice, depending on the smoothness of the learned field $F_{\theta,k\eta}$ and the quality of its score approximation, each subproblem $g_k(z)$ can typically be solved in a single deterministic update, or refined through a few inner iterations. DDIM achieves faster generation achieves faster generation by reducing the number of reverse transition kernels that must be sampled.Instead of solving T Gaussian RTK subproblems (as in DDPM), DDIM constructs a sparse reverse trajectory and computes larger RTK jumps using closed-form, noise-free transitions. Because DDIM preserves the correct marginals $q(x_t \mid x_0)$, these coarse RTK steps remain consistent with the original diffusion model. Thus, DDIM is an example of RTK acceleration achieved through trajectory subsampling, analogous in spirit to the RTK–MALA (Huang et al. (2024)) and RTK–ULD (Huang et al. (2024)) approaches, which reduce the number of subproblems by employing stronger, more expressive kernels.

6 Conclusion and Future Work

The admissibility of a constant step size $\eta = \Theta(1)$ in DDDM plays a crucial role in enabling efficient diffusion inference. When the parameterized mappings $F_{\theta,k\eta}$ satisfy uniform smoothness and boundedness conditions, the associated proxy energy $g(z)$ becomes m-strongly convex, ensuring that the corresponding proxy density $\hat{p}(z \mid x) \propto \exp\big(-g(z)\big)$ is sharply concentrated around a unique minimizer. This strong convexity guarantees that each reverse-time update converges stably without requiring excessively small step sizes. Consequently, the denoising trajectory decomposes into only $\tilde{O}(1)$ well-conditioned subproblems, as opposed to the $\tilde{O}(1/\eta) = \tilde{O}(1/\epsilon^2)$ subproblems required in standard DDPMs. This constant-step-size regime removes the need for finely discretized reverse diffusion schedules, substantially reducing iterative updates while preserving numerical stability and accuracy. In ongoing and extended work, we aim to further substantiate these observations through empirical analysis, including measurements of curvature, smoothness, and stability of the learned reverse dynamics.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2023. URL https://arxiv.org/abs/2209.15571.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. URL https://arxiv.org/abs/2303.08797.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/958c530554f78bcd8e97125b70e6973d-Paper.pdf.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization, 2024. URL https://arxiv.org/abs/2308.03686.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling, 2023. URL https://arxiv.org/abs/2106.01357.

Haoxuan Chen, Yinuo Ren, Lexing Ying, and Grant M. Rotskoff. Accelerating diffusion models with parallel sampling: Inference at sub-linear time complexity. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 133661–133709. Curran Associates, Inc., 2024. doi: 10.52202/079017-4248. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f162fa05675e3db4a733aafc081653cf-Paper-Conference.pdf.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions, 2023a. URL https://arxiv.org/abs/2211.01916.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast, 2023b. URL https://arxiv.org/abs/2305.11798.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023c. URL https://arxiv.org/abs/2209.11215.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Jonathan Heek, Emiel Hoogeboom, and Tim Salimans. Multistep consistency models, 2024. URL https://arxiv.org/abs/2403.06807.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Xunpeng Huang, Difan Zou, Hanze Dong, Yi Zhang, Yi-An Ma, and Tong Zhang. Reverse transition kernel: A flexible framework to accelerate diffusion inference, 2024. URL https://arxiv.org/abs/2405.16387.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models, 2025. URL https://arxiv.org/abs/2410.11081.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed, 2021. URL https://arxiv.org/abs/2101.02388.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. URL https://arxiv.org/abs/2202.00512.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020a. URL https://arxiv.org/abs/1907.05600.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models, 2020b. URL https://arxiv.org/abs/2006.09011.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem, 2023. URL https://arxiv.org/abs/2206.04119.

Dan Zhang, Jingjing Wang, and Feng Luo. Directly denoising diffusion models, 2024. URL https://arxiv.org/abs/2405.13540.

A Further Discussions

Although DDIM employs the same Gaussian RTK approximation as DDPM, its key difference arises in the *deterministic* limit $\sigma_t \to 0$. In this limit, the Gaussian kernel degenerates into a Dirac measure, yielding an implicit deterministic mapping

$$x_{t-1} = \mu_{\mathrm{DDIM}}(x_t, x_0). \tag{14}$$

Algorithm 2: DDIM Sampling as a Reverse Transition Kernel (RTK)

Input: Noise predictor ϵ_θ; noise schedule $\{\bar{\alpha}_{k\eta}\}_{k=0}^{K}$; coarse grid $\tau = \{k_0 = 0 < k_1 < \cdots < k_S = K\}$ with $S \ll T$

Output: Sample $\hat{x}_{K\eta} \sim p^*$

Initialization: Sample $\hat{x}_{k_0\eta} \sim \mathcal{N}(0, I)$;

for $i = 0$ **to** $S - 1$ **do**

Clean sample estimation (RTK mean estimator):

$$\hat{x}_0^{(k)} := f_\theta^{(k\eta)}(\hat{x}_{k\eta}) = (\hat{x}_{k\eta} - \sqrt{1 - \bar{\alpha}_{k\eta}}\, \epsilon_\theta^{(k\eta)}(\hat{x}_{k\eta}))/\sqrt{\bar{\alpha}_{k\eta}}$$

RTK transition (deterministic DDIM limit):

$$\hat{x}_{k_{i+1}\eta} = \sqrt{\bar{\alpha}_{k_{i+1}\eta}}\, \hat{x}_0^{(k_i)} + \sqrt{\frac{1 - \bar{\alpha}_{k_{i+1}\eta}}{1 - \bar{\alpha}_{k_i\eta}}} \left(\hat{x}_{k_i\eta} - \sqrt{\bar{\alpha}_{k_i\eta}}\, \hat{x}_0^{(k_i)} \right)$$

return $\hat{X}_{K\eta}$

Since this mapping is closed-form and deterministic, it composes across time indices. Consequently, instead of solving T RTK subproblems (as in DDPM), one may select a coarse sampling trajectory $\tau = \{k_0 < k_1 < \cdots < k_S\}$ with $S \ll T$ and apply the RTK map only along this subsequence:

$$x_{k_{i+1}\eta} = \mu_{\mathrm{DDIM}}\left(x_{k_i\eta},\, x_0^{(k_i)} \right). \tag{15}$$

Importantly, DDIM preserves the exact diffusion marginals $p(x_t \mid x_0)$, ensuring that these coarse RTK transitions remain consistent with the original diffusion process. Thus, DDIM accelerates generation by reducing the number of RTK subproblems while retaining model fidelity, mirroring the acceleration philosophy of the general RTK framework.

B Proof of Proposition 1

Proof. In the subsequent analysis, we omit the subscripts $\theta, k\eta$ for notational simplicity. Let

$$g(z) = \frac{1}{2\sigma^2}\|z - (x - F(z))\|^2 = \frac{1}{2\sigma^2}\sum_{i=1}^{d} r_i(z)^2, \qquad r_i(z) = z_i - x_i + F_i(z).$$

Differentiating gives

$$\partial_j g(z) = \frac{1}{\sigma^2}\sum_{i=1}^{d} r_i(z)\, \partial_j r_i(z), \qquad \nabla g(z) = \frac{1}{\sigma^2} J_r(z)^\top r(z),$$

where $(J_r)_{ij} = \partial_j r_i(z)$. A second differentiation and substitution $\partial_j r_i(z) = \delta_{ij} + \partial_j F_i(z)$, $\partial_k \partial_j r_i(z) = \partial_k \partial_j F_i(z)$ yields the exact Hessian

$$\nabla^2 g(z) = \frac{1}{\sigma^2}\Big((I + J_F(z))^\top (I + J_F(z)) + \sum_{i=1}^{d} r_i(z)\, H_{F,i}(z)\Big). \tag{16}$$

where $J_F(z) = \nabla_z F(z)$ and $H_{F,i}(z) = \nabla^2 F_i(z)$ is the Hessian of the i-th output coordinate.

Consider an arbitrary direction $v \in \mathbb{R}^d$. Then

$$v^\top \nabla^2 g(z)\, v = \frac{1}{\sigma^2}\Big(\|(I + J_F(z))v\|^2 + \sum_{i=1}^{d} r_i(z)\, v^\top H_{F,i}(z)\, v\Big). \tag{17}$$

We bound the two terms separately.

(I) Lower bound for the quadratic term. Using the reverse triangle inequality and the operator norm of J_F,

$$\|(I + J_F)v\| = \|v + J_F v\| \geq \|v\| - \|J_F v\| \geq (1 - \|J_F\|_{\mathrm{op}})\|v\|.$$

Define $L(z) := \|J_F(z)\|_{\mathrm{op}}$; then

$$\|(I + J_F)v\|^2 \geq (1 - L(z))^2 \|v\|^2.$$

(II) Upper bound for the correction (Hessian) term. Set $a_i := v^\top H_{F,i}(z)\, v$ (scalars). By the operator-norm bound on $H_{F,i}$,

$$|a_i| \leq \|H_{F,i}(z)\|_{\mathrm{op}}\, \|v\|^2. \tag{18}$$

Thus

$$\sum_{i=1}^{d} a_i^2 \leq \|v\|^4 \sum_{i=1}^{d} \|H_{F,i}(z)\|_{\mathrm{op}}^2 = B_{\mathrm{sq}}(z)\, \|v\|^4. \tag{19}$$

where we define $B_{\mathrm{sq}}(z) := \sum_{i=1}^{d} \|H_{F,i}(z)\|_{\mathrm{op}}^2$. By Cauchy–Schwarz over the index i,

$$\Big|\sum_{i=1}^{d} r_i(z)\, a_i\Big| \leq \|r(z)\|_2 \sqrt{\sum_{i=1}^{d} a_i^2} \leq R(z)\, B_{\mathrm{sq}}(z)\, \|v\|^2. \tag{20}$$

where $R(z) := \|r(z)\|_2$.

(III) Combine the two bounds. Inserting (I) and (II) into the expression for $v^\top \nabla^2 g(z) v$ gives

$$v^\top \nabla^2 g(z)\, v \geq \frac{1}{\sigma^2}\Big((1 - L(z))^2 - R(z)\, B_{\mathrm{sq}}(z)\Big)\|v\|^2.$$

Therefore if $(1 - L(z))^2 > R(z) B_{\mathrm{sq}}(z)$ we obtain a positive lower bound; defining

$$m := \frac{1}{\sigma^2}\Big((1 - L(z))^2 - R(z) B_{\mathrm{sq}}(z)\Big) > 0. \tag{21}$$

yields $v^\top \nabla^2 g(z) v \geq m\|v\|^2$ for all v, i.e. $\nabla^2 g(z) \succeq m I_d$ and g is m-strongly convex at z. Consequently $p(z) \propto e^{-g(z)}$ is m-strongly log-concave in a neighborhood of z.

Affine special case. If F is affine, $F(z) = \Lambda z + b$ (so $J_F = \Lambda$ and $H_{F,i} = 0$), then

$$\nabla^2 g(z) = \frac{1}{\sigma^2}(I + \Lambda)^\top (I + \Lambda), \qquad v^\top \nabla^2 g(z) v = \frac{1}{\sigma^2}\|(I + \Lambda)v\|^2 \geq 0.$$

Hence g is convex. Moreover:

(a) If $(I + \Lambda)$ is full-rank then $\|(I + \Lambda)v\| > 0$ for all $v \neq 0$, so g is strictly (and hence strongly) convex.

(b) If $(I + \Lambda)$ is not of full rank, then there exists a nonzero vector v such that $(I + \Lambda)v = 0$. Consequently, the Hessian $\nabla^2 g(z)$ admits a zero eigenvalue corresponding to this direction, implying that g is not strictly convex along v.

Corollary 1 (One-hidden-layer explicit bounds)**.** *Consider the one-hidden-layer network*

$$J_F(z) = W_2 \operatorname{diag}\big(\phi'(u)\big) W_1, \qquad u = W_1 z + b_1. \tag{22}$$

such that the Jacobian operator norm admits the bound

$$\|J_F(z)\|_{\mathrm{op}} \le \|W_2\|_{\mathrm{op}} \|W_1\|_{\mathrm{op}} \max_j |\phi'(u_j)|. \tag{23}$$

The i-th component Hessian is given by

$$H_{F,i}(z) = W_1^\top \operatorname{diag}\big(w_{2,i} \odot \phi''(u)\big) W_1. \tag{24}$$

and therefore satisfies

$$\|H_{F,i}(z)\|_{\mathrm{op}} \le \|W_1\|_{\mathrm{op}}^2 \max_j |w_{2,i,j}\, \phi''(u_j)|. \tag{25}$$

A uniform scalar curvature bound can thus be taken as

$$B = \|W_1\|_{\mathrm{op}}^2 \max_{i,j} |w_{2,i,j}| \max_j |\phi''(u_j)|. \tag{26}$$

Substituting into Eq. 21 gives the explicit sufficient condition

$$\Big(1 - \|W_2\|_{\mathrm{op}} \|W_1\|_{\mathrm{op}} \max_j |\phi'(u_j)|\Big)^2 > R\left(\|W_1\|_{\mathrm{op}}^2 \max_{i,j} |w_{2,i,j}| \max_j |\phi''(u_j)|\right). \tag{27}$$

[2]

□

C Proof of Proposition 2

Proof. From the computation in the main text we have the exact Hessian

$$\nabla^2 g(z) = \frac{1}{\sigma^2}(I + J_F(z))^\top (I + J_F(z)) + R(z), \qquad R(z) := \frac{1}{\sigma^2} \sum_{i=1}^{d} r_i(z)\, H_{F,i}(z),$$

where $J_F(z) = \nabla_z F_{\theta,k\eta}(z)$ and $H_{F,i}(z) = \nabla^2 F_i(z)$. The first matrix is a Gram matrix and therefore positive semidefinite:

$$\frac{1}{\sigma^2}(I + J_F)^\top (I + J_F) \succeq 0.$$

By assumption $\|J_F(z)\|_2 \le \kappa < 1$. For any unit vector v the reverse triangle inequality yields

$$\|(I + J_F)v\| = \|v + J_F v\| \ge \big|\|v\| - \|J_F v\|\big| \ge 1 - \|J_F\|_2 \ge 1 - \kappa,$$

hence the smallest singular value of $I + J_F$ satisfies $\sigma_{\min}(I + J_F) \ge 1 - \kappa$.

Using the definition of singular values of a matrix A (spectral inequality from SVD),

$$A^\top A \succeq \sigma_{\min}(A)^2 I.$$

we obtain the spectral inequality

$$(I + J_F)^\top (I + J_F) \succeq (1 - \kappa)^2 I,$$

[2]For piecewise-linear nonlinearities such as ReLU or leaky ReLU, $\phi'' = 0$ almost everywhere, implying $B = 0$. The sufficient condition then simplifies to $\|J_F\|_{\mathrm{op}} < 1$, which is a global contraction condition ensuring convexity everywhere.

Finally, scaling by $\frac{1}{\sigma^2}$

$$\frac{1}{\sigma^2}(I + J_F)^\top (I + J_F) \succeq \frac{(1-\kappa)^2}{\sigma^2} I.$$

We now establish an operator-norm bound for the remainder term $R(z)$. We assume that, over the region of interest, the following uniform bounds hold.

$$\|H_{F,i}(z)\|_{\mathrm{op}} \le B \qquad \text{for } i = 1, \dots, d, \qquad \text{and} \qquad \|r(z)\|_2 \le R_{\max}.$$

Then, using triangle inequality and Cauchy–Schwarz,

$$\Big\| \sum_{i=1}^{d} r_i(z)\, H_{F,i}(z) \Big\|_{\mathrm{op}} \le \sum_{i=1}^{d} |r_i(z)|\, \|H_{F,i}(z)\|_{\mathrm{op}} \le \|r(z)\|_2 \sqrt{\sum_{i=1}^{d} \|H_{F,i}(z)\|_{\mathrm{op}}^2} \le R_{\max} \sqrt{d}\, B.$$

Including the factor $1/\sigma^2$ from the definition of $R(z)$ gives

$$\|R(z)\|_{\mathrm{op}} \le \frac{R_{\max} \sqrt{d}\, B}{\sigma^2}.$$

Hence, we can define

$$C_B \;=\; \frac{R_{\max} \sqrt{d}\, B}{\sigma^2},$$

such that $\|R(z)\|_{\mathrm{op}} \le C_B$ uniformly on the region. Combining the two bounds yields the matrix inequality

$$\nabla^2 g(z) \succeq \left(\frac{(1-\kappa)^2}{\sigma^2} - C_B \right) I.$$

Therefore, if

$$\frac{(1-\kappa)^2}{\sigma^2} > C_B,$$

then $\nabla^2 g(z)$ is uniformly positive definite on the region and g is m-strongly convex with modulus

$$m \;=\; \frac{(1-\kappa)^2}{\sigma^2} - C_B \;>\; 0.$$

□