Non-Normal Diffusion Models

Henry Li [1]

Abstract

Diffusion models generate samples by incrementally reversing a process that turns data into noise. We show that when the step size goes to zero, the reversed process is invariant to the distribution of these increments. This reveals a previously unconsidered parameter in the design of diffusion models: the distribution of the diffusion step $\boldsymbol{\Delta}\mathbf{x}_k := \mathbf{x}_k - \mathbf{x}_{k+1}$. This parameter is implicitly set by default to be normally distributed in most diffusion models. By lifting this assumption, we generalize the framework for designing diffusion models and establish an expanded class of diffusion processes with greater flexibility in the choice of loss function used during training. We demonstrate the effectiveness of these models on density estimation and generative modeling tasks on standard image datasets, and show that different choices of the distribution of $\boldsymbol{\Delta}\mathbf{x}_k$ result in qualitatively different generated samples.

1. Introduction

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b; Vahdat & Kautz, 2020; Dhariwal & Nichol, 2021) have quickly established themselves as one of the most powerful classes of generative models in an already crowded and competitive space — one which also includes GANs (Goodfellow et al., 2020; Brock et al., 2018; Karras et al., 2019), VAEs (Kingma & Welling, 2013; Vahdat & Kautz, 2020; Child, 2020), flows (Dinh et al., 2014; Kingma & Dhariwal, 2018; Dinh et al., 2016), and autoregressive models (Salimans et al., 2017; Oord et al., 2016; Child et al., 2019), among others.

A standard assumption for diffusion models is that $\boldsymbol{\Delta}_{\mathbf{x}_k} := \mathbf{x}_k - \mathbf{x}_{k+1}$ are normally distributed (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b; Ho et al., 2022). However, there are many known cases in physical and biological systems where the random incremental behavior of particles colliding in a space does not follow the standard Gaussian distribution (Hidalgo-Soria & Barkai, 2020; Cugliandolo, 2002). These examples are also called anomalous diffusions (Gefen et al., 1983; Bouchaud & Georges, 1990). In this work, we consider such a scenario, and propose a generalized framework for modeling diffusion models with minimal assumptions on the distribution of the $\boldsymbol{\Delta}_{\mathbf{x}_k}$. To develop this framework, we prove a novel result on the convergence of non-time homogeneous random walks to stochastic processes in the limit of small time steps. Finally, we demonstrate that our framework allows for greater freedom in the design of the model and its training dynamics, while retaining competitive generative modeling capabilities in terms of both model likelihood and sample quality.

[1]Department of Computer Science and Applied Mathematics, Yale University, New Haven, CT. Correspondence to: Henry Li <henry.li@yale.edu>.

2. Background

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b) take the form $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}_{0:T})\,\mathrm{d}\mathbf{x}_{1:T}$ where data $\mathbf{x}_0 := \mathbf{x}$ are related to a set of latent variables $\mathbf{x}_{1:T} := (\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T))$ distributed as marginals of a diffusion process governed by an Itô stochastic differential equation (SDE)

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w} \tag{1}$$

with respect to time points $\{t_k\}_{k=1}^T$. $\mathbf{f}$ and g are typically called *drift* and *diffusion* functions, and $\mathbf{w}$ is the standard Wiener process. Samples can then be generated by modeling the reverse diffusion, which has a simple form given by (Anderson, 1982)

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \underbrace{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)}_{\approx \mathbf{s}_\theta(\mathbf{x}, t)}]\,\mathrm{d}t + g(t)\,\mathrm{d}\overline{\mathbf{w}}, \tag{2}$$

where $\overline{\mathbf{w}}$ is a reverse-time Wiener process. Note that Eq. (2) is itself an Itô SDE of the form Eq. (1). Training the diffusion model involves approximating the true score function $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ with a neural network $\mathbf{s}_\theta(\mathbf{x}, t)$ in Eq. (2). This can be achieved directly via score matching (Hyvärinen & Dayan, 2005; Song & Ermon, 2019; Song et al., 2020b), or by modeling the sampling process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Kingma et al., 2021), which is obtained by discretizing the reverse-time SDE into a Markov

chain with joint likelihood

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{k=0}^{T-1} \nu_\theta(\mathbf{x}_k|\mathbf{x}_{k+1}) \tag{3}$$

or equivalently

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{k=0}^{T-1} \rho_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1}), \tag{4}$$

where ν_θ, ρ_θ are Markov models and $\boldsymbol{\Delta}_{\mathbf{x}_k} := \mathbf{x}_{k+1} - \mathbf{x}_k$. While most works e.g. (Song et al., 2020b; Ho et al., 2020; Kingma et al., 2021) model Eq. (3), we shall turn our attention to the equivalent formulation Eq. (4), which focuses on the *increments*, rather than the *marginals* of the diffusion. Letting q be the density of the Gaussian process Eq. 2, Eqs. (3) and (4) result in the same likelihood bound

$$\begin{aligned}\log p_\theta(\mathbf{x}) \geq \mathbb{E}_q \Bigg[& \underbrace{\log p(\mathbf{x}_0|\mathbf{x}_1)}_{\mathcal{L}_0} \\ & - \sum_{k=1}^{T} \underbrace{KL(q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1})||p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1}))}_{\mathcal{L}_k} \\ & - \underbrace{KL(q(\mathbf{x}_T)||p(\mathbf{x}_T))}_{\mathcal{L}_T} \Bigg]\end{aligned} \tag{5}$$

that reduces to a simple function of $\mathbf{s}_\theta(\mathbf{x}, t)$.

When forming approximations such as Eq. (4), it is important to consider the conditions under which they converge to Eq. (2). While this convergence is known for normally distributed $\boldsymbol{\Delta}_{\mathbf{x}_k}$ (Sohl-Dickstein et al., 2015; Song et al., 2020b; Särkkä & Solin, 2019), we shall extend this result to arbitrarily distributed $\boldsymbol{\Delta}_{\mathbf{x}_k}$ in Section 3.

Ultimately, either choice of learning $\mathbf{s}_\theta(\mathbf{x}, t)$ allows for unbiased estimates of $\log p_\theta(\mathbf{x})$ by modeling the probability flow ODE (PF-ODE) corresponding to Eq. (2), which can be derived via the Fokker-Planck equation (Song et al., 2020b)

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \right] \mathrm{d}t, \tag{6}$$

and substituting the score with $\mathbf{s}_\theta(\mathbf{x}, t)$.

3. Convergence of Non-Normal Random Walks to Diffusion Processes

A fundamental challenge in diffusion modeling is forming tractable approximations to Eq. (1). Our result is inspired by Donsker's classic Invariance Principle (Billingsley, 2013), which gives the functional convergence of an unbiased random walk to a standard Brownian motion. We now consider a time-inhomogeneous, biased random walk $\mathbf{x}_k$. Let $\mathbf{x}(t)$ be the solution to Eq. (1). Intuitively, one might expect a similar convergence of $\mathbf{x}_k$ to $\mathbf{x}(t)$ if we constrain the first and second moments of its increments $\boldsymbol{\Delta}_{\mathbf{x}_k} := \mathbf{x}_{k+1} - \mathbf{x}_k$ to be

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k] &= \mathbf{f}(\mathbf{x}_k, t_k)\Delta_{t_k} \\ \mathrm{Var}(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) &= g(t_k)^2 \Delta_{t_k}\end{aligned}. \tag{7}$$

This type of convergence has been previously explored for normally distributed $\boldsymbol{\Delta}_{\mathbf{x}_k}$ in diffusion modeling (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b), and is well known in general SDE literature (Särkkä & Solin, 2019; Øksendal & Øksendal, 2003; Kloeden & Platen, 1992). More generalized results also exist for time-homogeneous or equilibrium state processes (Ethier & Kurtz, 2009; Vidov & Romanovsky, 2009; Stroock, 2013). However, there does not exist to our knowledge a convergence result for the case of general $\boldsymbol{\Delta}_{\mathbf{x}_k}$ in our non-equilibrium case (Sohl-Dickstein et al., 2015). Here we shall provide such a result, and show that convergence occurs with surprisingly few assumptions. This inspires a generalized framework for designing diffusion probabilistic models where the distribution of $\boldsymbol{\Delta}_{\mathbf{x}_k}$ is left as a tunable free parameter. We leverage this framework in Section 4 to define a generalized class of diffusion probabilistic models.

3.1. Structured Random Walks

Let $\mathbf{x}_k$ be a random walk. We introduce the following notion of structure, which allows us to characterize a random walk entirely in terms of the drift and diffusion functions $\mathbf{f}$ and g, the time step Δ_{t_k}, and a sequence of independent variables $\mathbf{z}_k$.

Definition 1 (Structured Random Walks). We say that a random walk $\mathbf{x}_k$ is **structured** (with respect to an Itô SDE) when its increments $\boldsymbol{\Delta}_{\mathbf{x}_k} := \mathbf{x}_{k+1} - \mathbf{x}_k$ support the decomposition

$$\boldsymbol{\Delta}_{\mathbf{x}_k} = \mathbf{f}(\mathbf{x}_k, t_k)\Delta_{t_k} + g(t_k)\sqrt{\Delta_{t_k}}\mathbf{z}_k, \tag{8}$$

where $\mathbb{E}[\mathbf{z}_k] = 0$, $\mathrm{Var}(\mathbf{z}_k) = 1$, $\Delta_{t_k} := t_{k+1} - t_k$, and $\mathbf{f}$, g correspond to the drift and diffusion terms of the respective Itô SDE.

The structural property in Definition 1 is quite natural. In fact, it is how diffusion steps are usually computed, e.g., via the reparameterization trick (Kingma & Welling, 2013; Ho et al., 2020) or SDE solvers such as the Euler-Maruyama method (Song et al., 2020b). Moreover, it satisfies Eq. (7). If we additionally assume that $\mathbf{f}(\mathbf{x}, t)$ is linear in $\mathbf{x}$, as is the case with the forward diffusion process in standard diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b; Kingma et al., 2021), we have the following closed form representations of its first and second moments at all $k \in \{0, \ldots, T\}$.

p	q	$KL(p(\mathbf{x})\|\|q(\mathbf{x}))$	$\mathcal{L}_k$ (note: **)
$\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2)$	$\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2)$	$\frac{1}{2\sigma^2}\|\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|\|^2$	$\mathbb{E}\left[w_k\|\|\mathbf{r}_k\|\|^2\right]$
Laplace$(\boldsymbol{\mu}_1, \sigma^2)$	Laplace$(\boldsymbol{\mu}_2, \sigma^2)$	$\exp\left(-\frac{\|\mu_2-\mu_1\|}{\sigma}\right) + \frac{\|\mu_2-\mu_1\|}{\sigma} + 1$	$\mathbb{E}\big[\exp(-v_k\|\|\mathbf{r}_k\|\|_1) - 1 + v_k\|\|\mathbf{r}_k\|\|_1\big]$
Uniform$[\boldsymbol{\mu}_1 - \sqrt{3}\sigma, \boldsymbol{\mu}_1 + \sqrt{3}\sigma]$	$\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2)$	$\frac{1}{2}\left(\frac{1}{\sigma^2}(\mu_1 - \mu_2)^2 + \log\frac{\pi}{6} + 1\right)$	$w_k\mathbb{E}_{\boldsymbol{\epsilon}}\|\|\mathbf{r}_k\|\|^2 + \frac{1}{2}(1 + \log\sqrt{\frac{\pi}{6}})$
Uniform$[\boldsymbol{\mu}_1 - \sqrt{3}\sigma, \boldsymbol{\mu}_1 + \sqrt{3}\sigma]$	Laplace$(\boldsymbol{\mu}_2, \sigma^2)$	$\begin{cases}\frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2 + \frac{1}{2} & \mu_2 \in A^* \\ \frac{1}{\sigma}\|\mu_1 - \mu_2\| & \mu_2 \notin A\end{cases}$	$\begin{cases}w_k\mathbb{E}\|\|\mathbf{r}_k\|\|_2^2 + \frac{1}{2} & \text{if } \boldsymbol{\epsilon}_\theta(\mathbf{x}, t) \in A \\ v_k\mathbb{E}\|\|\mathbf{r}_k\|\|_1 & \text{if } \boldsymbol{\epsilon}_\theta(\mathbf{x}, t) \notin A\end{cases}$

Table 1. Summary of the diffusion models proposed in Section 4. $^*A = [\mu_1 - b_1, \mu_1 + b_1]$. $^{**}\mathbf{r}_k := \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)$.

Theorem 3.1 (Moments of Structured Random Walks). Let $\mathbf{x}_k$ be a structured random walk and $\mathbf{f}(\mathbf{x}, t_k) = \beta(t_k)\mathbf{x}$ be linear. Then

$$\boldsymbol{\mu}(t_k) := \mathbb{E}[\mathbf{x}_k] = \bar{\alpha}_k\mathbf{x}_0 \quad \text{and} \quad \boldsymbol{\sigma}(t_k)^2 := \mathrm{Var}(\mathbf{x}_k) = \bar{\gamma}_k,$$

where $\bar{\alpha}_k = \prod_{i=1}^k (1 + \beta_i)$ and $\bar{\gamma}_k = \sum_{i=1}^k \left(\frac{\bar{\alpha}_k}{\bar{\alpha}_{i+1}} g_i\right)^2$. For notational convenience, we let $\beta_i := \beta(t_i)\Delta_{t_k}$ and $g_i := g(t_i)\sqrt{\Delta_{t_k}}$.

In diffusion modeling, we are not just interested in computing the moments of $\mathbf{x}_k$ — we would like to sample from $p(\mathbf{x}_k)$[1]. This is a difficult task for generally distributed $\boldsymbol{\Delta}_{\mathbf{x}_k}$, since the distribution of $\mathbf{x}_k = \mathbf{x}_0 + \sum_{i=1}^k \boldsymbol{\Delta}_{\mathbf{x}_i}$ is usually intractable. To sidestep this issue, many works assume that $\boldsymbol{\Delta}_{\mathbf{x}_k}$ are normally distributed; since Gaussian random variables are closed under summation and specified by their first and second moments, we see below that Lemma 3.1 is sufficient for identifying the distribution of $\mathbf{x}_k$.

Corollary 3.1. Let $\mathbf{x}_k$ be a structured random walk, $\mathbf{z}_k$ be normally distributed, and $\mathbf{f}(\mathbf{x}, t) = \beta(t)\mathbf{x}$ where $\beta(t)$ is one of the noise schedules defined in (Sohl-Dickstein et al., 2015; Ho et al., 2020; Kingma et al., 2021). Then we recover their respective forward processes.

3.2. An Invariance Principle

Lifting the assumption of normally distributed increments $\boldsymbol{\Delta}_{\mathbf{x}_k}$, we show that we still ultimately obtain a Gaussian process in the limit as $\Delta_{t_k} \to 0$. Much like the aforementioned Donsker's theorem, this also gives rise to an invariance — in the distribution of $\boldsymbol{\Delta}_{\mathbf{x}_k}$. We once again leverage the notion of *structured* random walks to present a general theorem for the convergence of Markov chains with increments of the form Eq. (8).

Theorem 3.2 (Structured Invariance Principle). Suppose regularity conditions (B.3) hold and $\{\mathbf{x}_k\}_{k=1}^n$ is a structured random walk on $\mathbb{R}^d$. Let $\overline{\mathbf{x}}_T(t) = \mathbf{x}_0 + \sum_{k=1}^{n_t} \boldsymbol{\Delta}_{\mathbf{x}_k}$ be the continuous-time càdlàg extension of $\mathbf{x}_k$, where $n_t = \lfloor t * T \rfloor$. Then $\overline{\mathbf{x}}_T$ converges in distribution to $\mathbf{x}$, the solution to the Itô SDE (Eq. 1), as $\Delta_{t_k} \to 0$.

[1]Where $p = q$ or $p = p_\theta$.

Theorem 3.3 (Structured Invariance Principle). Suppose regularity conditions hold and $\{\mathbf{x}_k\}_{k=1}^n$ is a structured random walk on $\mathbb{R}^d$. Let $\overline{\mathbf{x}}_T(t) = \mathbf{x}_0 + \sum_{k=1}^{n_t} \boldsymbol{\Delta}_{\mathbf{x}_k}$ be the continuous-time càdlàg extension of $\mathbf{x}_k$, where $n_t = \lfloor t * T \rfloor$. Then $\overline{\mathbf{x}}_T$ converges in distribution to $\mathbf{x}(t)$, as $\Delta_{t_k} \to 0$.

Theorem 3.3 outlines the existence of a much larger class of increments $\boldsymbol{\Delta}_{\mathbf{x}_k}$ that converge to our desired limiting distribution $\mathbf{x}(t)$. The convergence to $\mathbf{x}(t)$ unlocks many of the essential properties for the tractability of diffusion models which we take for granted in Gaussian increments, such as fast sampling from the forward process and a closed form Eq. (5), without the need to assume Gaussian increments. Finally, we verify that we can recover Donsker's theorem when we let $\mathbf{f} = \mathbf{0}$ and $g = 1$.

4. Non-Normal Diffusion Models

Leveraging the framework established in Section 3, we introduce an expanded class of probabilistic diffusion models, centered around alternative distributional assumptions for $q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1})$ and $p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1})$. While the space of viable diffusion models allowed by Theorem 3.3 effectively contains all distributions of $\boldsymbol{\Delta}_{\mathbf{x}_k}$ with finite mean and variance, we restrict our study to the following examples and leave further exploration to future work. Detailed derivations can be found in Appendix A.4. A summary of all models can be found in Table 1.

4.1. Gaussian q and p_θ

First, we recover the default diffusion model loss term $\mathcal{L}_k$ (from Eq. 5) by making the standard assumption that $\boldsymbol{\Delta}_{\mathbf{x}_k}$ are normally distributed. Since the space of Gaussian-distributed random variables is closed under affine operations, we trivially obtain the convergence of the random walk (Eq. 4) to a Gaussian process. Using the closed form mean and variance terms of a linear ODE (Särkkä & Solin, 2019), we obtain

$$\mathcal{L}_k = w_k\mathbb{E}_{\boldsymbol{\epsilon}}\|\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)\|\|^2, \tag{9}$$

where $w_k = \frac{g(t_k)^2}{2\sigma(t_k)^2}\Delta_{t_k}$ and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k) = \sigma(t_k)\mathbf{s}_\theta(\mathbf{x}_k, t_k)$. Plugging Eq. 9 into the likelihood bound Eq. 5, we see that

maximizing the likelihood of a standard diffusion model with Gaussian increments minimizes a quadratic error term between the score function $\mathbf{s}_\theta(\mathbf{x}_k, t_k) = \frac{1}{\sigma(t_k)}\boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)$.

4.2. Laplace q and p_θ

We now consider the case of Laplace distributed $\boldsymbol{\Delta}_{\mathbf{x}_k}$. Invoking Theorem 3.3, we can derive the alternative loss

$$\mathcal{L}_k = \mathbb{E}_{\boldsymbol{\epsilon}}\big[\exp(-v_k||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_1) - 1 + v_k||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_1\big], \tag{10}$$

where $v_k := \sqrt{w_k}$.

While the term in the expectation $\mathbf{d}(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)) := \exp\left(-v_k||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_1\right) - 1 + v_k||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_1$ appears somewhat opaque, we can see that it converges to a weighted $L1$ norm of the error $\mathbf{r}_k := \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)$ under two conditions:

$$\lim_{t_k \to 0} \frac{v_k||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_1}{\mathbf{d}(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k))} = 1, \tag{11}$$

i.e., when t is small, and

$$\lim_{||\mathbf{r}_k||_1 \to \infty} \frac{v_k||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_1}{\mathbf{d}(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k))} = 1 \tag{12}$$

i.e., when $||\mathbf{r}_k||$ is large.

4.3. Uniform q, Gaussian p_θ

Next, we note that q and p_θ need not be the same family of distributions to apply our framework. To illustrate this, we let q be uniformly distributed on the interval $[\mu_1 - \sqrt{3}\sigma, \mu_2 + \sqrt{3}\sigma$, and p be Gaussian distributed. This results in the familiar form

$$\mathcal{L}_k = w_k \mathbb{E}_{\boldsymbol{\epsilon}}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||^2 + C, \tag{13}$$

where $C = \frac{1}{2}\left(1 + \log\frac{\pi}{6}\right) \approx 0.34$ may be seen as an additional distributional mismatch penalty incurred by the joint combination of the uniform and normal distributions. We note, however, that such a penalty does not always arise when p_θ and q are not from the same family of distributions.

4.4. Uniform q, Laplace p_θ

Finally, we demonstrate that the phase transition in Section 4.2 to an $L1$-based loss is made explicit in the case where q is uniform and p_θ is the Laplace distribution. This configuration of distributions produces the piecewise loss

$$\mathcal{L}_k = \begin{cases} w_k \mathbb{E}_{\boldsymbol{\epsilon}}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_2^2 + \frac{1}{2} & \text{if } \boldsymbol{\epsilon}_\theta(\mathbf{x}, t) \in A \\ v_k \mathbb{E}_{\boldsymbol{\epsilon}}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||_1 & \text{if } \boldsymbol{\epsilon}_\theta(\mathbf{x}, t) \notin A \end{cases}, \tag{14}$$

where $A = [\mu_1 - \sqrt{3}\sigma w_k, \mu_1 + \sqrt{3}\sigma w_k]$. Now, it is clear that $\mathcal{L}_k$ acts as a linear function in two cases. First, when $t_k \to 0$, as A becomes a vanishingly small set. And second, when $\mathbf{r}_k := \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)$ is large. Both imply that $\boldsymbol{\epsilon}_\theta(\mathbf{x}, t) \notin A$.

q	p_θ	BPD	FID
Gaussian	Gaussian	2.49	1.98
Laplace	Laplace	2.47	2.44
Uniform	Gaussian	2.82	1.99
Uniform	Laplace	2.66	2.39

Table 2. Comparison between the proposed diffusion models on the CIFAR10 dataset. We evaluate in terms of negative log-likelihood (BPD, lower is better) and sample quality (FID, lower is better). BPD and FID are computed with different architectures.

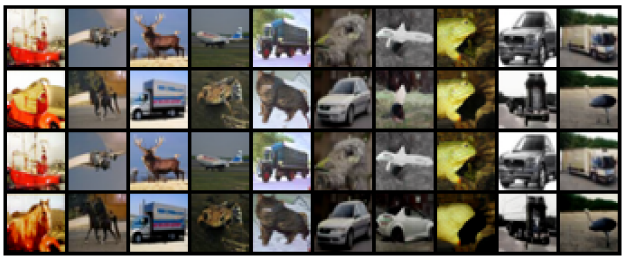

Figure 1. Images generated from the same seed via (in order from top to bottom) Gaussian-Gaussian, Laplace-Laplace, Uniform-Gaussian, and Uniform-Laplace diffusion increments. While the qualitative difference is somewhat subtle, Laplace diffusion appears to be biased towards smoother images with more saturated colors.

5. Experiments

For illustrative purposes, we evaluate the diffusion models proposed in Section 4 on the CIFAR10 (Krizhevsky et al., 2009) and down-sampled ImageNet (Van Den Oord et al., 2016) datasets. We quantify the performance of our models with the negative log-likelihood in terms of bits per dimension (BPD) and the Frechet Inception Distance (Heusel et al., 2017). Results are displayed in Table 2. We show that our model obtains competitive results in terms of both metrics.

More interestingly, some of the losses proposed in Section 4 result in generated samples with distinctly different visual characteristics. For example, images generated by the Laplace-based diffusion models exhibit markedly more saturated colors (Figure 1.

6. Conclusion and Limitations

We derived a probabilistic framework for designing more diverse diffusion models by showing an invariance to the distribution of the diffusion step $\boldsymbol{\Delta}_{\mathbf{x}_k} := \mathbf{x}_k - \mathbf{x}_{k+1}$. Freeing

up the distributional assumption on $\mathbf{\Delta}_{\mathbf{x}_k}$ allows the end-user greater control over the stylistic qualities of the generative model. An open question is whether score matching under an EMD norm enjoys the same statistical guarantees as the standard score matching objective, e.g., consistency, efficiency, and asymptotic normality (Hyvärinen, 2006; Song et al., 2020a). We hope that our theoretical framework opens the door for the further diversity and improvements in the design of diffusion models.

References

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

Billingsley, P. *Convergence of probability measures*. John Wiley & Sons, 2013.

Bouchaud, J.-P. and Georges, A. Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Physics reports*, 195(4-5):127–293, 1990.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Child, R. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Cugliandolo, L. F. Dynamics of glassy systems. *arXiv preprint cond-mat/0210312*, 2002.

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

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Ethier, S. N. and Kurtz, T. G. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

Gefen, Y., Aharony, A., and Alexander, S. Anomalous diffusion on percolating clusters. *Physical Review Letters*, 50(1):77, 1983.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Hidalgo-Soria, M. and Barkai, E. Hitchhiker model for laplace diffusion processes. *Physical Review E*, 102(1): 012109, 2020.

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

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022.

Hyvärinen, A. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

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

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kloeden, P. E. and Platen, E. *Stochastic differential equations*. Springer, 1992.

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

Øksendal, B. and Øksendal, B. *Stochastic differential equations*. Springer, 2003.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Särkkä, S. and Solin, A. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

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

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020a.

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.

Stroock, D. W. *An introduction to Markov processes*, volume 230. Springer Science & Business Media, 2013.

Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

Vidov, P. and Romanovsky, M. Y. Analytical representations of non-gaussian laws of random walks. *Physics of wave phenomena*, 17:218–227, 2009.

A. Derivations

A.1. KL Divergence Between Laplace Distributions

For completeness, we provide a derivation for the KL divergence between two Laplace distributions. Let p and q be density functions of distributions Laplace(μ_1, b_1) and Laplace(μ_2, b_2), i.e.,

$$p(x) = \frac{1}{2b_1} \exp\left(-\frac{|x-\mu_1|}{b_1}\right) \tag{15}$$

$$q(x) = \frac{1}{2b_2} \exp\left(-\frac{|x-\mu_2|}{b_2}\right). \tag{16}$$

Then the KL divergence between the two distributions can be written as

$$KL(p(x)||q(x)) = \underbrace{\int_{-\infty}^{\infty} p(x)\log p(x)dx}_{*} - \underbrace{\int_{-\infty}^{\infty} p(x)\log q(x)dx}_{**} \tag{17}$$

We will first approach $**$ as its solution will give us $*$. Plugging in p and q, we have

$$-\int_{-\infty}^{\infty} p(x)\log q(x)dx = \int_{-\infty}^{\infty} \frac{|x-\mu_2|}{2b_1b_2} \exp\left(-\frac{|x-\mu_1|}{b_1}\right) dx + \log(2b_2),$$

where in the case that $\mu_1 > \mu_2$, the integral can be written as

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{|x-\mu_2|}{2b_1b_2} \exp\left(-\frac{|x-\mu_1|}{b_1}\right) dx \\
&\quad= \int_{-\infty}^{\mu_2} \frac{\mu_2 - x}{2b_1b_2} \exp\left(-\frac{\mu_1 - x}{b_1}\right) dx + \int_{\mu_2}^{\mu_1} \frac{x-\mu_2}{2b_1b_2} \exp\left(-\frac{\mu_1 - x}{b_1}\right) dx \\
&\qquad+ \int_{\mu_1}^{\infty} \frac{x-\mu_2}{2b_1b_2} \exp\left(-\frac{x-\mu_1}{b_1}\right) dx \\
&\quad= \left[\frac{b_1}{2b_2} \exp\left(-\frac{\mu_1-\mu_2}{b_1}\right)\right] + \left[\frac{\mu_1-\mu_2-b_1}{2b_2} + \frac{b_1}{2b_2} \exp\left(-\frac{\mu_1-\mu_2}{b_1}\right)\right] \\
&\qquad+ \left[\frac{\mu_1-\mu_2+b_1}{2b_2}\right] \\
&\quad= \frac{\mu_1-\mu_2}{b_2} + \frac{b_1}{b_2} \exp\left(-\frac{\mu_1-\mu_2}{b_1}\right),
\end{aligned}$$

and similarly for the case $\mu_1 \leq \mu_2$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{|x-\mu_2|}{2b_1b_2} \exp\left(-\frac{|x-\mu_1|}{b_1}\right) dx \\
&\quad= \int_{-\infty}^{\mu_1} \frac{\mu_2 - x}{2b_1b_2} \exp\left(-\frac{\mu_1 - x}{b_1}\right) dx + \int_{\mu_1}^{\mu_2} \frac{\mu_2-x}{2b_1b_2} \exp\left(-\frac{x-\mu_1}{b_1}\right) dx \\
&\qquad+ \int_{\mu_2}^{\infty} \frac{x-\mu_2}{2b_1b_2} \exp\left(-\frac{x-\mu_1}{b_1}\right) dx \\
&\quad= \left[\frac{\mu_2-\mu_1+b_1}{2b_2}\right] + \left[\frac{b_1}{2b_2} \exp\left(-\frac{\mu_2-\mu_1}{b_1}\right) + \frac{\mu_2-\mu_1-b_1}{2b_2}\right] \\
&\qquad+ \left[\frac{b_1}{2b_2} \exp\left(-\frac{\mu_2-\mu_1}{b_1}\right)\right] \\
&\quad= \frac{\mu_2-\mu_1}{b_2} + \frac{b_1}{b_2} \exp\left(-\frac{\mu_2-\mu_1}{b_1}\right).
\end{aligned}$$

Combining both cases and returning to the original cross entropy term, we have

$$-\int_{-\infty}^{\infty} p(x)\log q(x)dx = \frac{|\mu_2-\mu_1|}{b_2} + \frac{b_1}{b_2}\exp\left(-\frac{|\mu_2-\mu_1|}{b_1}\right) + \log(2b_2). \tag{18}$$

Now, letting $p = q$ we can compute the entropy term as

$$\int_{-\infty}^{\infty} p(x)\log p(x)dx = -1 - \log(2b_1). \tag{19}$$

Thus, we can conclude that

$$KL(p(x)||q(x)) = \frac{b_1}{b_2}\exp\left(-\frac{|\mu_2-\mu_1|}{b_1}\right) + \frac{|\mu_2-\mu_1|}{b_2} + \log\frac{b_2}{b_1} - 1. \tag{20}$$

A.2. KL Divergence Between a Gaussian Distribution and a Bounded Uniform Distribution

Let p and q denote the density functions of the Uniform$([\mu_1 - b_1, \mu_1 + b_1])$ and $\mathcal{N}(\mu_2, \sigma_2)$ distributions, respectively. Then

$$p(x) = \mathbb{1}_{x\in[\mu_1-b_1,\mu_1+b_1]}\frac{1}{2b_1} \tag{21}$$

$$q(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\left(\frac{x-\mu_2}{b_2}\right)^2\right). \tag{22}$$

Again writing the KL decomposition between p and q as Eq. 17, we note that the entropy term $*$ is identical to that of Section **??**:

$$\int_{-\infty}^{\infty} p(x)\log p(x)dx = -\log(2b_1).$$

Turning to the cross-entropy term $**$:

$$\begin{aligned}
-\int_{-\infty}^{\infty} p(x)\log q(x)dx &= \int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1}\left(\log(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2}(x-\mu_2)^2\right)dx \\
&= \log(\sigma\sqrt{2\pi}) + \int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{4b_1\sigma^2}(x-\mu_2)^2dx \\
&= \log(\sigma\sqrt{2\pi}) + \frac{1}{4b_1\sigma^2}\left(\frac{1}{3}x^3 - \mu_2x^2 + \mu_2^2x\right)\Bigg|_{\mu_1-b_1}^{\mu_1+b_1} \\
&= \log(\sigma\sqrt{2\pi}) + \frac{1}{4b_1\sigma^2}\left[2b_1\left((\mu_1-\mu_2)^2 + \frac{1}{3}b_1^2\right)\right]
\end{aligned}$$

Combining terms, we obtain the KL divergence

$$KL(p(x)||q(x)) = \frac{1}{2}\left(\frac{1}{\sigma^2}(\mu_1-\mu_2)^2 + \log\frac{\pi}{6} + 1\right), \tag{23}$$

where we note that $b_1 = \sqrt{3}\sigma$.

A.3. KL Divergence Between a Laplace Distribution and a Bounded Uniform Distribution

Let p and q denote the density functions of the Uniform$([\mu_1 - b_1, \mu_1 + b_1])$ and Laplace(μ_2, b_2) distributions, respectively. Then

$$p(x) = \mathbb{1}_{x\in[\mu_1-b_1,\mu_1+b_1]}\frac{1}{2b_1} \tag{24}$$

$$q(x) = \frac{1}{2b_2}\exp\left(-\frac{|x-\mu_2|}{b_2}\right). \tag{25}$$

We once again write the KL decomposition between p and q as Eq. 17, and begin with the entropy term $*$:

$$\begin{aligned}\int_{-\infty}^{\infty} p(x)\log p(x)dx &= -\int_{\mu_1-b_1}^{\mu_1+b_1}\frac{\log(2b_1)}{2b_1}dx \\ &= \frac{\log(2b_1)}{2b_1}(\mu_1-b_1)-\frac{\log(2b_1)}{2b_1}(\mu_1+b_1) \\ &= -\log(2b_1).\end{aligned}$$

Turning to the cross-entropy term $**$:

$$\begin{aligned}-\int_{-\infty}^{\infty} p(x)\log q(x)dx &= \int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1}\left(\log(2b_2)+\frac{|x-\mu_2|}{b_2}\right)dx \\ &= \log(2b_2)+\int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1}\frac{|x-\mu_2|}{b_2}dx.\end{aligned}$$

Considering the case where $\mu_2 < \mu_1 - b_1$, the above integral reduces to

$$\begin{aligned}\int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1b_2}|x-\mu_2|dx &= \int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1b_2}x-\mu_2 dx \\ &= \frac{1}{2b_1b_2}\left(\frac{1}{2}x^2-\mu_2 x\right)\Big|_{\mu_1-b_1}^{\mu_1+b_1} \\ &= \frac{1}{2b_1b_2}\left(2b_1(\mu_1-\mu_2)\right),\end{aligned}$$

whereas the case $\mu_2 < \mu_1 - b_1$ gives

$$\begin{aligned}\int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1b_2}|x-\mu_2|dx &= \int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1b_2}x-\mu_2 dx \\ &= \frac{1}{2b_1b_2}\left(\mu_2 x-\frac{1}{2}x^2\right)\Big|_{\mu_1-b_1}^{\mu_1+b_1} \\ &= \frac{1}{2b_1b_2}\left(2b_1(\mu_2-\mu_1)\right).\end{aligned}$$

Finally, when $\mu_2 \in [\mu_1 - b_1, \mu_1 + b_1]$, we have

$$\begin{aligned}\int_{\mu_1-b_1}^{\mu_1+b_1}\frac{1}{2b_1b_2}|x-\mu_2|dx &= \frac{1}{2b_1b_2}\left[\int_{\mu_1-b_1}^{\mu_2}(\mu_2-x)dx+\int_{\mu_2}^{\mu_1+b_1}\frac{1}{2b_1b_2}(x-\mu_2)dx\right] \\ &= \frac{1}{2b_1b_2}\left[\left(\mu_2 x-\frac{1}{2}x^2\right)\Big|_{\mu_1-b_1}^{\mu_2}+\left(\frac{1}{2}x^2-\mu_2 x\right)\Big|_{\mu_2}^{\mu_1+b_1}\right] \\ &= \frac{1}{2b_1b_2}\left[\left(\frac{1}{2}(\mu_1-\mu_2)^2+\frac{1}{2}b_1^2+b_1(\mu_2-\mu_1)\right)+\left(\frac{1}{2}(\mu_1-\mu_2)^2+\frac{1}{2}b_1^2+b_1(\mu_1-\mu_2)\right)\right] \\ &= \frac{1}{2b_1b_2}\left[(\mu_1-\mu_2)^2+b_1^2\right].\end{aligned}$$

Combining the cases, we obtain the KL divergence

$$KL(p(x)||q(x)) = \begin{cases}\log\frac{b_2}{b_1}+\frac{1}{2b_1b_2}((\mu_1-\mu_2)^2+b_1^2) & \mu_2\in[\mu_1-b_1,\mu_1+b_1] \\ \log\frac{b_2}{b_1}+\frac{1}{b_2}|\mu_1-\mu_2| & \mu_2\notin[\mu_1-b_1,\mu_1+b_1]\end{cases}. \tag{26}$$

A.4. Deriving $\mathcal{L}_k$

We use the following lemmas to obtain Eqs. (9) and (10) in Sections 4.1 and 4.2. Throughout this section, we will use

$$\mathbf{f}_\theta = \mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p(\mathbf{x}, t), \tag{27}$$

$$\hat{\mathbf{f}}_\theta = \mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \mathbf{s}_\theta(\mathbf{x}, t), \tag{28}$$

where $\mathbf{f}$ and g are defined as in Eq. (1), to denote the true and learned reverse drift terms described in Eq (2).

Lemma A.1. Let $\boldsymbol{\Delta}_{\mathbf{x}_k}$ be normally distributed, i.e.,

$$p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \mathcal{N}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \hat{\mathbf{f}}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2 \Delta_{t_k}\right), \tag{29}$$

$$q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \mathcal{N}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \mathbf{f}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2 \Delta_{t_k}\right). \tag{30}$$

Then

$$\mathcal{L}_k = w_k \mathbb{E}_{\boldsymbol{\epsilon}\sim q}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||^2, \tag{31}$$

where $w_k := \frac{g(t_k)^2}{2\sigma(t_k)^2}\Delta_{t_k}$.

Proof. Plugging in the closed form solution to the KL divergence between two Gaussian distributions into the likelihood lower bound,

$$\begin{aligned}\mathcal{L}_k &= KL(p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k)||q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k)) \\ &= \mathbb{E}\left[\frac{||\boldsymbol{\mu}_{p_\theta,k}(\mathbf{x}_k) - \boldsymbol{\mu}_{q,k}(\mathbf{x}_k)||^2}{2\sigma^2}\right].\end{aligned}$$

Since $\boldsymbol{\mu}_{q,k} = (\mathbf{f}(\mathbf{x}_k, t_k) - g(t_k)^2\nabla \log p(\mathbf{x}_k))\Delta_{t_k}$, $\boldsymbol{\mu}_{p,k} = (\mathbf{f}(\mathbf{x}_k, t_k) - g(t_k)^2\nabla \log p_\theta(\mathbf{x}_k))\Delta_{t_k}$, and $\sigma_{p_\theta,k} = \sigma_{q,k} = g(t_k)\sqrt{\Delta_{t_k}}$, we have

$$\begin{aligned}\mathcal{L}_k &= \frac{1}{2}\mathbb{E}\left[\frac{||g(t_k)^2\nabla \log p(\mathbf{x}_k) - g(t_k)^2\nabla \log p_\theta(\mathbf{x}_k)||^2\Delta_{t_k}^2}{g(t_k)^2\Delta_{t_k}}\right] \\ &= \frac{1}{2}\mathbb{E}\left[g(t_k)^2||\nabla \log p(\mathbf{x}_k) - \nabla \log p_\theta(\mathbf{x}_k)||^2\Delta_{t_k}\right].\end{aligned}$$

Finally, following the parameterization of the score model ($i.e.$, $\boldsymbol{\epsilon}_\theta(\mathbf{x}, t) = \sigma(t_k)\nabla p_\theta(\mathbf{x}, t)$) in (Ho et al., 2020), we may write

$$\mathcal{L}_k = w_k \mathbb{E}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||^2\right], \tag{32}$$

where $w_k := \frac{1}{2}g(t_k)^2\sigma(t_k)^2\Delta_{t_k}$, and $\sigma(t_k)^2$ is as defined in Lemma 3.1. □

Lemma A.2. Let $\boldsymbol{\Delta}_{\mathbf{x}_k}$ be Laplace distributed, i.e.,

$$p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \text{Laplace}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \hat{\mathbf{f}}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2 \Delta_{t_k}\right), \tag{33}$$

$$q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \text{Laplace}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \mathbf{f}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2 \Delta_{t_k}\right). \tag{34}$$

Then, letting $\mathbf{r}_k := \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)$,

$$\mathcal{L}_k = \exp\left(-w_k||\mathbf{r}_k||_1\right) - 1 + w_k||\mathbf{r}_k||_1, \tag{35}$$

where $w_k := \frac{g(t_k)}{\sigma(t_k)}\sqrt{\Delta_{t_k}}$..

Proof. Plugging in the closed form solution to the KL divergence between two Laplace distributions into the likelihood lower bound (Appendix A.1),

$$\begin{aligned}\mathcal{L}_k &= KL(p(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k)||q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k)) \\ &= \exp\underbrace{\left(\frac{-||\mu_{p_\theta,k}-\mu_{q,k}||_1}{\sigma_p}\right)}_{\mathbf{d}_k} - 1 + \underbrace{\frac{||\mu_{p_\theta,k}-\mu_{q,k}||_1}{\sigma_p}}_{\mathbf{d}_k}\end{aligned}$$

Observe that $\mathbf{d}_k$ can be simplified as

$$\begin{aligned}\mathbf{d}_k &= \frac{||\mu_{p_\theta,k}-\mu_{q,k}||_1}{\sigma_p} \\ &= \frac{g(t_k)^2||\nabla_{\mathbf{x}}\log p(\mathbf{x}_k) - \nabla_{\mathbf{x}}\log p_\theta(\mathbf{x}_k)||_1}{g(t_k)\sqrt{\Delta_{t_k}}} \\ &= \frac{\sqrt{\Delta_{t_k}}g(t_k)||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_k,t_k)||_1}{\sigma(t_k)} \\ &= v_k||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_k,t_k)||_1, \end{aligned} \tag{36}$$

where $v_k := \frac{g(t_k)}{\sigma(t_k)}\sqrt{\Delta_{t_k}}$. Therefore,

$$\mathcal{L}_k = (-v_k||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_k,t_k)||_1) - 1 + v_k||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_k,t_k)||_1.$$

□

Lemma A.3. Let $p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1})$ be normally distributed and $q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1})$ be the uniform distribution on the interval $[\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma]$, i.e.,

$$p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \mathcal{N}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \hat{\mathbf{f}}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2\Delta_{t_k}\right) \tag{37}$$

$$q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \text{Uniform}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \mathbf{f}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2\Delta_{t_k}\right). \tag{38}$$

Then

$$\mathcal{L}_k = w_k\mathbb{E}_{\boldsymbol{\epsilon}}||\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_k,t_k)||^2 + C, \tag{39}$$

where $w_k := \frac{g(t_k)}{\sigma(t_k)}\sqrt{\Delta_{t_k}}$ and $C = \frac{1}{2}\left(1 + \log\frac{\pi}{6}\right)$.

Proof. Plugging in the closed form solution to the KL divergence between a Gaussian distribution and a Uniform distribution (Appendix A.2), we have

$$\begin{aligned}\mathcal{L}_k &= KL(p(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k)||q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k)) \\ &= \mathbb{E}\left[\frac{||\boldsymbol{\mu}_{p_\theta,k}(\mathbf{x}_k) - \boldsymbol{\mu}_{q,k}(\mathbf{x}_k)||^2}{2\sigma^2}\right] + C,\end{aligned}$$

where $C = \frac{1}{2}\left(1 + \log\frac{\pi}{6}\right)$. Since the expectation is the same as Eq. 32 in Theorem A.1, we are done. □

Lemma A.4. Let $p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1})$ be Laplace distributed and $q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_{k+1})$ be the uniform distribution on the interval $[\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma]$, i.e.,

$$p_\theta(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \text{Laplace}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \hat{\mathbf{f}}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2\Delta_{t_k}\right) \tag{40}$$

$$q(\boldsymbol{\Delta}_{\mathbf{x}_k}|\mathbf{x}_k) = \text{Uniform}\left(\boldsymbol{\Delta}_{\mathbf{x}_k}; \mathbf{f}_\theta\left(\mathbf{x}_k, t_k\right)\Delta_{t_k}, g\left(t_k\right)^2\Delta_{t_k}\right). \tag{41}$$

Then

$$\mathcal{L}_k = \begin{cases} w_k \mathbb{E}_{\boldsymbol{\epsilon}} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||^2 + \frac{1}{2} & \mu_2 \in [\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma] \\ v_k \mathbb{E}_{\boldsymbol{\epsilon}} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, t_k)||^1 & \mu_2 \notin [\mu_1 - \sqrt{3}\sigma w_k, \mu_1 + \sqrt{3}\sigma w_k] \end{cases}, \tag{42}$$

where $w_k := \frac{g(t_k)}{\sigma(t_k)}\sqrt{\Delta_{t_k}}$ and $v_k := \sqrt{w_k}$.

Proof. Plugging in the closed form solution to the KL divergence between a Laplace distribution and a Uniform distribution (Appendix A.3), we have

$$\mathcal{L}_k = \begin{cases} \mathbb{E}\left[\frac{||\boldsymbol{\mu}_{p_\theta,k}(\mathbf{x}_k) - \boldsymbol{\mu}_{q,k}(\mathbf{x}_k)||^2}{\sigma^2}\right] + \frac{1}{2} & \mu_2 \in [\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma] \\ \mathbb{E}\left[\frac{||\boldsymbol{\mu}_{p_\theta,k}(\mathbf{x}_k) - \boldsymbol{\mu}_{q,k}(\mathbf{x}_k)||^1}{\sigma}\right] & \mu_2 \notin [\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma] \end{cases}.$$

We observe that the event $\mu_2 \in [\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma]$ is equivalent to the event $\boldsymbol{\epsilon}(\mathbf{x}, t) \in [\boldsymbol{\epsilon} - \sqrt{3}\sigma w_k, \mu_1 + \sqrt{3}\sigma w_k]$. Using Eqs. 32 and 36 from Theorems A.1 and A.2 respectively, we are done. □

B. Proofs

B.1. Simple Properties of Structured Random Walks

We show several immediate properties of structured random walks discussed in Section 3.1.

Theorem 3.1 (Moments of Structured Random Walks). Let $\mathbf{x}_k$ be a structured random walk and $\mathbf{f}(\mathbf{x}, t_k) = \beta(t_k)\mathbf{x}$ be linear. Then

$$\boldsymbol{\mu}(t_k) := \mathbb{E}[\mathbf{x}_k] = \bar{\alpha}_k \mathbf{x}_0 \quad \text{and} \quad \boldsymbol{\sigma}(t_k)^2 := \mathrm{Var}(\mathbf{x}_k) = \bar{\gamma}_k,$$

where $\bar{\alpha}_k = \prod_{i=1}^k (1 + \beta_i)$ and $\bar{\gamma}_k = \sum_{i=1}^k \left(\frac{\bar{\alpha}_k}{\bar{\alpha}_{i+1}} g_i\right)^2$. For notational convenience, we let $\beta_i := \beta(t_i)\Delta_{t_k}$ and $g_i := g(t_i)\sqrt{\Delta_{t_k}}$.

Proof. We first show the derivation for $\mathbb{E}[\mathbf{x}_k]$. Observe that

$$\begin{aligned} \mathbb{E}[\mathbf{x}_k] &= \mathbb{E}[\mathbf{x}_{k-1} + \boldsymbol{\Delta}_{\mathbf{x}_k}] \\ &= \mathbb{E}\left[\mathbf{x}_{k-1} + \mathbf{f}\left(\mathbf{x}_{k-1}, t_k\right)\Delta_{t_k} + g\left(t_k\right)\mathbf{z}_k\sqrt{\Delta_{t_k}}\right] \\ &= \mathbb{E}[\mathbf{x}_{k-1}]\left(1 + \beta\left(t_k\right)\Delta_{t_k}\right). \end{aligned}$$

Applying this operation $k-1$ more times, we obtain

$$\mathbb{E}[\mathbf{x}_k] = \mathbb{E}[\mathbf{x}_0]\prod_{i=1}^k \left(1 + \beta\left(t_i\right)\Delta_{t_k}\right).$$

Turning to $\mathrm{Var}(\mathbf{x}_k)$, we first note that

$$\begin{aligned} \mathbb{E}[\mathbf{x}_k^2] &= \mathbb{E}[(\mathbf{x}_{k-1} + \boldsymbol{\Delta}_{\mathbf{x}_k})^2] \\ &= \mathbb{E}[\mathbf{x}_{k-1}^2] + \mathbb{E}[\boldsymbol{\Delta}_{\mathbf{x}_k}^2] + 2\mathbb{E}[\mathbf{x}_{k-1}\boldsymbol{\Delta}_{\mathbf{x}_k}] \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}[\mathbf{x}_{k-1}\boldsymbol{\Delta}_{\mathbf{x}_k}] &= \mathbb{E}\left[\mathbf{x}_{k-1}\left(\mathbf{f}\left(\mathbf{x}_{k-1}, t_k\right)\Delta_{t_k} + g\left(t_k\right)\mathbf{z}_k\sqrt{\Delta_{t_k}}\right)\right] \\ &= \beta\left(t_k\right)\Delta_{t_k}\mathbb{E}\left[\mathbf{x}_{k-1}^2\right] + g\left(t_k\right)\sqrt{\Delta_{t_k}}\mathbb{E}\left[\mathbf{x}_{k-1}\mathbf{z}_k\right] \\ &= \beta\left(t_k\right)\Delta_{t_k}\mathbb{E}\left[\mathbf{x}_{k-1}^2\right] \end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Delta}_{\mathbf{x}_k}^2] &= \mathbb{E}\left[\left(\mathbf{f}\left(\mathbf{x}_{k-1}, t_k\right)\Delta_{t_k}\right)^2 + \left(g\left(t_k\right)\mathbf{z}_k\sqrt{\Delta_{t_k}}\right)^2 + 2\left(\mathbf{f}\left(\mathbf{x}_{k-1}, t_k\right)\Delta_{t_k}\right)\left(g\left(t_k\right)\mathbf{z}_k\sqrt{\Delta_{t_k}}\right)\right] \\ &= \beta\left(t_k\right)^2\Delta_{t_k}^2\mathbb{E}\left[\mathbf{x}_{k-1}^2\right] + g\left(t_k\right)^2\Delta_{t_k}\mathbb{E}\left[\mathbf{z}_k^2\right] + \beta\left(t_k\right)g\left(t_k\right)\Delta_{t_k}^{\frac{3}{2}}\mathbb{E}[\mathbf{x}_{k-1}\mathbf{z}_k] \\ &= \beta\left(t_k\right)^2\Delta_{t_k}^2\mathbb{E}\left[\mathbf{x}_{k-1}^2\right] + g\left(t_k\right)^2\Delta_{t_k}\end{aligned}$$

Putting things together, we have

$$\begin{aligned}\mathbb{E}[\mathbf{x}_k^2] &= \mathbb{E}[(\mathbf{x}_{k-1} + \boldsymbol{\Delta}_{\mathbf{x}_k})^2] \\ &= \mathbb{E}[\mathbf{x}_{k-1}^2] + \mathbb{E}[\boldsymbol{\Delta}_{\mathbf{x}_k}^2] + 2\mathbb{E}[\mathbf{x}_{k-1}\boldsymbol{\Delta}_{\mathbf{x}_k}] \\ &= \mathbb{E}[\mathbf{x}_{k-1}^2]\left(1 + \beta\left(t_k\right)^2\Delta_{t_k}^2 + 2\beta\left(t_k\right)\Delta_{t_k}\right) + g\left(t_k\right)^2\Delta_{t_k} \\ &= \mathbb{E}[\mathbf{x}_{k-1}^2]\left(1 + \beta\left(t_k\right)\Delta_{t_k}\right)^2 + g\left(t_k\right)^2\Delta_{t_k}.\end{aligned}$$

This gives, by induction,

$$\mathbb{E}[\mathbf{x}_k^2] = \mathbb{E}[\mathbf{x}_0^2]\prod_{i=1}^{k}\left(1 + \beta\left(t_i\right)\Delta_{t_k}\right)^2 + \sum_{j=1}^{k}\prod_{i=j+1}^{k}\left(1 + \beta\left(t_i\right)\Delta_{t_i}\right)^2 g\left(t_j\right)^2\Delta_{t_j}.$$

Finally, we can write

$$\begin{aligned}\mathrm{Var}(\mathbf{x}_k) &= \mathbb{E}[\mathbf{x}_k^2] - \mathbb{E}[\mathbf{x}_k]^2 \\ &= \mathrm{Var}(\mathbf{x}_0)\prod_{i=1}^{k}\left(1 + \beta\left(t_i\right)\Delta_{t_k}\right)^2 + \sum_{j=1}^{k}\prod_{i=j+1}^{k}\left(1 + \beta\left(t_i\right)\Delta_{t_i}\right)^2 g\left(t_j\right)^2\Delta_{t_j}.\end{aligned}$$

Assuming that $\mathrm{Var}(\mathbf{x}_0) = 0$, we now have

$$\boldsymbol{\sigma}_k^2 := \mathrm{Var}(\mathbf{x}_k) = \sum_{j=1}^{k}\prod_{i=j+1}^{k}\left(1 + \beta\left(t_i\right)\Delta_{t_i}\right)^2 g\left(t_j\right)^2\Delta_{t_j}. \tag{43}$$

□

B.2. Deriving Previous Methods in Our Framework

Corollary 3.1. Let $\mathbf{x}_k$ be a structured random walk, $\mathbf{z}_k$ be normally distributed, and $\mathbf{f}(\mathbf{x}, t) = \beta(t)\mathbf{x}$ where $\beta(t)$ is one of the noise schedules defined in (Sohl-Dickstein et al., 2015; Ho et al., 2020; Kingma et al., 2021). Then we recover their respective forward processes.

Denoising Diffusion Probabilistic Models We first examine the forward processes in (Ho et al., 2020) and (Sohl-Dickstein et al., 2015), which have the forward Markov chain

$$p_\theta(\mathbf{x}_{k+1}|\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_{k+1}; \sqrt{1-\beta_k}\mathbf{x}_k, \beta_k\mathbf{I}), \tag{44}$$

and thus that $\mathbf{x}_{k+1}$ may be written in terms of $\mathbf{x}_k$ as

$$\mathbf{x}_{k+1} = \sqrt{1-\beta_k}\mathbf{x}_k + \sqrt{\beta_k}\boldsymbol{\epsilon}, \tag{45}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Subtracting $\mathbf{x}_k$ from both sides and leveraging the fact that $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ are both normally distributed, we obtain

$$\boldsymbol{\Delta}_{\mathbf{x}_k} = (\sqrt{1-\beta_k} - 1)\mathbf{x}_k + \sqrt{\beta_k}\boldsymbol{\epsilon}. \tag{46}$$

Now, we see that we can clearly write Eq. (46) as a structured random walk (Eq. 8). Applying Theorem 3.1, we have that

$$\bar{\alpha}_k = \prod_{i=1}^{k}(\sqrt{1-\beta_i}) \qquad \bar{\gamma}_k = \sum_{i=1}^{k}\left(\frac{\bar{\alpha}_k}{\bar{\alpha}_{i+1}}\right)^2\beta_i. \tag{47}$$

This converges numerically to the form given in (Ho et al., 2020)

$$p(\mathbf{x}_k|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_k; \bar{\alpha}_k\mathbf{x}_0, (1-\bar{\alpha}^2)\mathbf{I}). \tag{48}$$

Variational Diffusion Models We can obtain a similar closed form solution for the forward process in (Kingma et al., 2021). The sampling chain of the process can be written as

$$p(\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{k+1}; \frac{\alpha_{k+1}}{\alpha_k}\mathbf{x}_k, \sigma_{k+1}^2 - \left(\frac{\alpha_{k+1}}{\alpha_k}\right)\sigma_k^2\right), \tag{49}$$

where α_k and σ_k are related to each other by a monotonic function $\gamma(t)$

$$\alpha_k^2 = \text{sigmoid}(-\gamma(t)), \tag{50}$$
$$\sigma_k^2 = \text{sigmoid}(\gamma(t)). \tag{51}$$

According to Eq. 49, $\mathbf{x}_{k+1}$ can be written in terms of $\mathbf{x}_k$ as

$$\mathbf{x}_{k+1} = \frac{\alpha_{k+1}}{\alpha_k}\mathbf{x}_k + \left(\sigma_{k+1}^2 - \left(\frac{\alpha_{k+1}}{\alpha_k}\right)\sigma_k^2\right)\boldsymbol{\epsilon}, \tag{52}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Subtracting off $\mathbf{x}_k$ on both sides, we obtain

$$\boldsymbol{\Delta}_{\mathbf{x}_{k+1}} = \left(\frac{\alpha_{k+1}}{\alpha_k} - 1\right)\mathbf{x}_k + \left(\sigma_{k+1}^2 - \left(\frac{\alpha_{k+1}}{\alpha_k}\right)\sigma_k^2\right)\boldsymbol{\epsilon}. \tag{53}$$

Now we can once again apply Theorem 3.1, and see that

$$\bar{\alpha}_k = \prod_{i=1}^{k}\frac{\alpha_{i+1}}{\alpha_i} = \alpha_k \tag{54}$$

and

$$\bar{\gamma}_k = \sum_{i=1}^{k}\frac{\alpha_k}{\alpha_i}\left(\sigma_{i+1}^2 - \left(\frac{\alpha_i}{\alpha_{i-1}}\right)\sigma_i^2\right) \tag{55}$$
$$= \sum_{i=1}^{k}\frac{\alpha_k}{\alpha_i}\sigma_{i+1}^2 - \frac{\alpha_k}{\alpha_{i-1}}\sigma_i^2 \tag{56}$$
$$= \sigma_k^2, \tag{57}$$

which agrees with Eq. **??**.

B.3. Regularity Conditions

To show our main result, we state the following regularity conditions. Assumptions 1 and 2 are standard for finite-step discretizations of SDEs (Särkkä & Solin, 2019). Assumption 3 simplifies the subsequent proof for tightness.

Assumption 1 ($\mathbf{f}$ and g are Lipschitz). There exists $K > 0$ such that, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $t, s \in [0, 1]$

$$||\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|| \leq K||\mathbf{x} - \mathbf{y}||, \qquad \text{and} \qquad |g(t) - g(s)| \leq K|t - s|. \tag{58}$$

Assumption 2 (Linear growth of $\mathbf{f}$ and g). There exists $K > 0$ such that, for any $\mathbf{x} \in \mathbb{R}^d$ and $t \in [0, 1]$

$$||\mathbf{f}(\mathbf{x})|| \leq K(1 + ||\mathbf{x}||), \qquad \text{and} \qquad |g(t)| \leq K(1 + |t|). \tag{59}$$

Assumption 3 (Integrability of $\mathbf{z}_k$.). There exists $K \in \mathbb{R}$ such that

$$\mathbb{E}[||\mathbf{z}_k||^4]) < K. \tag{60}$$

B.4. Main Result

Our theorem below can be seen as a generalization of Donsker's Invariance Principle, and certain parts of the proof resembles that of the original theorem. Differences appear where we can no longer rely on the independence of the increments $\mathbf{\Delta}_{\mathbf{x}_k}$, which is heavily utilized in the original proof. By exploiting the structural properties of Definition 1, we can decompose $\mathbf{x}_k$ into a set of auxiliary processes with the same limit, which we can show to converge to $\mathbf{X}$ with techniques borrowed from the strong convergence of SDE solvers and central limit theorems.

Theorem 3.3 (Structured Invariance Principle). Suppose regularity conditions hold and $\{\mathbf{x}_k\}_{k=1}^n$ is a structured random walk on $\mathbb{R}^d$. Let $\overline{\mathbf{x}}_T(t) = \mathbf{x}_0 + \sum_{k=1}^{n_t} \mathbf{\Delta}_{\mathbf{x}_k}$ be the continuous-time càdlàg extension of $\mathbf{x}_k$, where $n_t = \lfloor t * T \rfloor$. Then $\overline{\mathbf{x}}_T$ converges in distribution to $\mathbf{x}(t)$, as $\Delta_{t_k} \to 0$.

Proof. Using Eq. 8 we may define the continuous-time extension of $\mathbf{x}_k$ as the process

$$\mathbf{x}_T(t) = \mathbf{x}_0 + \sum_{i=1}^{\lfloor t*T \rfloor} \mathbf{\Delta}_{\mathbf{x}_i}^{(T)} + (t * T - \lfloor t * T \rfloor)\mathbf{\Delta}_{\mathbf{x}_{\lceil t*T \rceil}}^{(T)}, \tag{61}$$

which is produced by linearly interpolating between the iterates of the random walk. We write the increments

$$\mathbf{\Delta}_{\mathbf{x}_i}^{(T)} := \mathbf{x}_{k+1} - \mathbf{x}_k = \mathbf{f}(\mathbf{x}_k, t_k)\Delta_{t_k}^{(T)} + g(t_k)\sqrt{\Delta_{t_k}^{(T)}} Z_k^{(T)} \tag{62}$$

with the superscript (T) to emphasize its dependence on T. We show convergence by invoking the following theorem.

Theorem B.1. (Theorem 13.1 from (Billingsley, 2013).) Let $\{\mathbf{x}_T\}, \mathbf{x}$ be processes (with associated probability measures $\{\mathcal{P}_T\}, \mathcal{P}$) such that $\mathbf{x}_T$ converges to $\mathbf{x}$ in finite dimensional distributions (f.d.d.), i.e., for any k time steps $t_1, t_2, \ldots, t_k$,

$$(\mathbf{x}_T(t_1), \mathbf{x}_T(t_2), \ldots, \mathbf{x}_T(t_k)) \xrightarrow{\mathcal{D}} (\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_k)). \tag{63}$$

If $\{\mathcal{P}_T\}$ are also tight, then $\mathbf{x}_T \Rightarrow_T \mathbf{x}$.

Theorem B.1 relates the pointwise weak convergence (of a sequence of marginals of a process) on a finite set of points to weak convergence of the path measures. This is made possible by Prohorov's theorem, which connects tightness to relative compactness. Thus, to show convergence, we must show two conditions are satisfied: 1) convergence in f.d.d., and 2) tightness of the associated sequence of measures. These are given by the following two lemmas.

Lemma B.1. The sequence of measures $\{P_{\mathbf{x}_k}\}_{k=1}^T$ corresponding to the **structured random walk** $\{\mathbf{x}_k\}_{k=1}^T$ is tight.

Lemma B.2. The continuous-time random walk interpolation $\mathbf{x}_T$ converges in finite dimensional distributions (f.d.d.) to the diffusion process (i.e., solution to Eq. (1) $\mathbf{x}$.

Combining Theorem B.1 with Lemmas B.1 and B.2, we obtain our result. □

Proof. (of Lemma B.1)

The result can be obtained via Kolmogorov's tightness criterion, which provides the following sufficient condition for tightness:

$$\sup_n \mathbb{E}\left[||\mathbf{x}_n(s) - \mathbf{x}_n(t)||^p\right] \leq C|s-t|^{1+\epsilon}, \quad \text{for some } \epsilon > 0, p \geq 1+\epsilon. \tag{64}$$

We shall demonstrate Eq. 64 for $\epsilon = 1, p = 4$. For any $s, t \in [0, T]$, choose k, ℓ such that

$$s \in \left[\frac{k-1}{n}, \frac{k}{n}\right) \quad \text{and} \quad t \in \left[\frac{\ell-1}{n}, \frac{\ell}{n}\right). \tag{65}$$

First, observe that, applying Definition 1, Assumption 2, and the fact that $\mathbf{z}_k \in \mathcal{L}^4 \implies \mathbf{E}[||\mathbf{z}_k||^4] \leq M$ for some $M \in \mathbb{R}$ and all $k \in \{0, \dots, n\}$,

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}_n(s-\Delta t) - \mathbf{x}_n(s)||_4] &= \mathbb{E}[||\mathbf{\Delta}_{\mathbf{x}_k}||_4] \\
&= \mathbb{E}\left[\left|\left|\mathbf{f}(\mathbf{x}_k)\Delta t + g\left(\frac{k}{n}\right)\mathbf{z}_k\sqrt{\Delta t}\right|\right|_4\right] \\
&\leq \Delta t\mathbb{E}[||\mathbf{f}(\mathbf{x}_k)||_4] + \sqrt{\Delta t}\left|g\left(\frac{k}{n}\right)\right|\mathbb{E}[||\mathbf{z}_k||_4]] \\
&\leq \Delta t\mathbb{E}[K(1+||\mathbf{x}_k||_4)] + \sqrt{\Delta t}KM\left(1+\frac{k}{n}\right) \\
&\leq C\sqrt{\Delta t}, \qquad (66)
\end{aligned}$$

where $C_1 \leq \mathcal{O}(\sqrt{\Delta t})$.

We will bound $\mathbb{E}\left[||\mathbf{x}_n(s) - \mathbf{x}_n(t)||^4\right]$ in three regimes:

Case 1: $k = \ell$

$$\begin{aligned}
\mathbb{E}\left[||\mathbf{x}_n(s) - \mathbf{x}_n(t)||_4\right] &= \mathbb{E}\left[\left|\left|\mathbf{x}_0 + \sum_{i=1}^{k-1}\mathbf{\Delta}_{\mathbf{x}_i} + (ns-k)\mathbf{\Delta}_{\mathbf{x}_k} - \mathbf{x}_0 - \sum_{i=1}^{\ell-1}\mathbf{\Delta}_{\mathbf{x}_i} - (nt-\ell)\mathbf{\Delta}_\ell\right|\right|_4\right] \\
&= \mathbb{E}\left[||(ns-k)\mathbf{\Delta}_{\mathbf{x}_k} - (nt-k)\mathbf{\Delta}_{\mathbf{x}_k}||_4\right] \\
&\leq (n|t-s|)\mathbb{E}\left[||\mathbf{\Delta}_{\mathbf{x}_k}||_4\right] \\
&\leq C_1\sqrt{n}|t-s|,
\end{aligned}$$

where we used Eq. 66 the fact that $\Delta t = \frac{1}{n}$. Finally, since $k = \ell \implies |t-s| \leq n^{-1}$, we take the fourth power of both sides of the inequality to obtain

$$\mathbb{E}\left[||\mathbf{x}_n(s) - \mathbf{x}_n(t)||^4\right] \leq C_1|t-s|^2. \qquad (67)$$

Case 2: $k = \ell + 1$

$$\begin{aligned}
\mathbb{E}\left[||\mathbf{x}_n(s) - \mathbf{x}_n(t)||_4\right] &= \mathbb{E}\left[\left|\left|\mathbf{x}_0 + \sum_{i=1}^{k-1}\mathbf{\Delta}_{\mathbf{x}_i} + (ns-k)\mathbf{\Delta}_{\mathbf{x}_k} - \mathbf{x}_0 - \sum_{i=1}^{\ell-1}\mathbf{\Delta}_{\mathbf{x}_i} - (nt-\ell)\mathbf{\Delta}_\ell\right|\right|_4\right] \\
&= \mathbb{E}\left[||\mathbf{\Delta}_\ell + (ns-k)\mathbf{\Delta}_{\mathbf{x}_k} - (nt-\ell)\mathbf{\Delta}_\ell||_4\right] \\
&= \mathbb{E}\left[||(ns-k)\mathbf{\Delta}_{\mathbf{x}_k} - (nt-k)\mathbf{\Delta}_\ell||_4\right] \\
&\leq |(ns-nt)|\mathbb{E}\left[||\mathbf{\Delta}_{\mathbf{x}_k}||_4\right] \\
&\leq C_1\sqrt{n}|t-s|,
\end{aligned}$$

where we again use Eq. 66 the fact that $\Delta t = \frac{1}{n}$. This time, we have that $|t-s| \leq 2n^{-1}$. Therefore,

$$\mathbb{E}\left[||\mathbf{x}_n(s) - \mathbf{x}_n(t)||^4\right] = 4C_1|t-s|^2. \qquad (68)$$

Case 3: $k > \ell + 2$

$$\begin{aligned}
&\mathbb{E}[||\mathbf{x}_n(s) - \mathbf{x}_n(t)||_4] \\
&\quad\leq \mathbb{E}\left[\left|\left|\mathbf{x}_n(s) - \mathbf{x}_n\left(\frac{k-1}{n}\right)\right|\right|_4 + \left|\left|\mathbf{x}_n\left(\frac{k-1}{n}\right) - \mathbf{x}_n\left(\frac{\ell-1}{n}\right)\right|\right|_4\right. \\
&\qquad\left. + \left|\left|\mathbf{x}_n(t) - \mathbf{x}_n\left(\frac{\ell-1}{n}\right)\right|\right|_4\right] \\
&\quad\leq C_1\sqrt{s - \frac{k-1}{n}} + \mathbb{E}\left[\left|\left|\mathbf{x}_n\left(\frac{k-1}{n}\right) - \mathbf{x}_n\left(\frac{\ell-1}{n}\right)\right|\right|_4\right] \\
&\qquad + C_1\sqrt{t - \frac{\ell-1}{n}} \\
&\quad\leq C_1\left(\sqrt{s - \frac{k-1}{n}} - \sqrt{t - \frac{\ell-1}{n}}\right) + \underbrace{\mathbb{E}\left[\left|\left|\mathbf{x}_n\left(\frac{k-1}{n}\right) - \mathbf{x}_n\left(\frac{\ell-1}{n}\right)\right|\right|_4\right]}_{(*)}
\end{aligned}$$

Inspecting $(*)$, we can see that

$$\begin{aligned}
\mathbb{E}\left[\left|\left|\mathbf{x}_n\left(\frac{k-1}{n}\right) - \mathbf{x}_n\left(\frac{\ell-1}{n}\right)\right|\right|_4\right] &= \mathbb{E}\left[\left|\left|\mathbf{x}_0 + \sum_{i=1}^{k-1}\boldsymbol{\Delta}_{\mathbf{x}_i} - \mathbf{x}_0 - \sum_{i=1}^{\ell-1}\boldsymbol{\Delta}_{\mathbf{x}_i}\right|\right|_4\right] \\
&= \mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-1}\boldsymbol{\Delta}_{\mathbf{x}_i}\right|\right|_4\right] \\
&\leq \sum_{i=\ell-1}^{k-1}\mathbb{E}[||\mathbf{f}(\mathbf{x}_i)\Delta t||] + \mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-1} g\left(\frac{i}{n}\right)\mathbf{z}_i\sqrt{\Delta t}\right|\right|_4\right] \\
&\leq C_1\sqrt{\frac{k}{n} - \frac{\ell}{n}} + \mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-1} g\left(\frac{i}{n}\right)\mathbf{z}_i\sqrt{\Delta t}\right|\right|_4\right]. \qquad (69)
\end{aligned}$$

We make the following observation about the second term in Eq. 69.

Lemma B.3.

$$\mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-1} g\left(\frac{i}{n}\right)\mathbf{z}_i\sqrt{\Delta t}\right|\right|^4\right] \leq C_2\left(\frac{k}{n} - \frac{\ell}{n}\right)^2 \qquad (70)$$

Proof. Letting $\mathbf{W}_i := g\left(\frac{i}{n}\right)\mathbf{z}_i\sqrt{\Delta t}$, the second term in Eq. 69, taken to the fourth power, can be written as

$$\mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-1} g\left(\frac{i}{n}\right)\mathbf{z}_i\sqrt{\Delta t}\right|\right|^4\right] = \mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-1}\mathbf{W}_i\right|\right|^4\right], \qquad (71)$$

where $\mathbb{E}[\mathbf{W}_i] = 0$ and $\mathbb{E}[\mathbf{W}_i^2] = g^2\left(\frac{i}{n}\right)\Delta t$. The result will be shown by induction. Separating an element of the sum and then expanding the norm, we can write this term as

$$\mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-2}\mathbf{W}_i + W_{k-1}\right|\right|^4\right] = \mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-2}\mathbf{W}_i\right|\right|^4 + \left|\left|W_{k-1}\right|\right|^4 + \left|\left|\sum_{i=\ell-1}^{k-2}\mathbf{W}_i\right|\right|^2\left|\left|W_{k-1}\right|\right|^2\right],$$

where the odd terms containing first moments of $\mathbf{W}_i$ go to zero. Leveraging Assumption 2 and the fact that $\mathbf{z}_k \in \mathcal{L}^4$ we can further simplify left hand side to

$$\begin{aligned}\mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-2}\mathbf{W}_i + W_{k-1}\right|\right|^4\right] &\leq \mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-2}\mathbf{W}_i\right|\right|^4\right] + M^4\left(g^4\left(\frac{k-1}{n}\right)(\Delta t)^2\right) \\ &\quad + \left(\sum_{i=\ell-1}^{k-2} M^2 g^2\left(\frac{i}{n}\right)\Delta t\right)\left(g^2\left(\frac{k-1}{n}\right)\Delta t\right) \\ &\leq \mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-2}\mathbf{W}_i\right|\right|^4\right] + C_2(k-\ell)(\Delta t)^2.\end{aligned}$$

Applying this operation $k-\ell-1$ more times, we obtain our desired result

$$\begin{aligned}\mathbb{E}\left[\left|\left|\sum_{i=\ell-1}^{k-1}\mathbf{W}_i\right|\right|^4\right] &\leq C_2(k-\ell)^2(\Delta t)^2 \\ &= C_2\left(\frac{k}{n}-\frac{\ell}{n}\right)^2.\end{aligned}$$

□

Assembling the parts, we obtain

$$\begin{aligned}\mathbb{E}[||\mathbf{x}_n(s)-\mathbf{x}_n(t)||^4] &\leq C_1\left[\left(s-\frac{k-1}{n}\right)^2 - \left(t-\frac{\ell-1}{n}\right)^2 + \left(\frac{k}{n}-\frac{\ell}{n}\right)^2\right] + C_2\left(\frac{k}{n}-\frac{\ell}{n}\right)^2 \\ &\leq C_3|t-s|^2. \end{aligned} \tag{72}$$

Finally, we combine Eqs. 67, 68, and 72, which provides a bound that satisfies Kolmogorov's tightness criterion:

$$\mathbb{E}[||\mathbf{x}_n(s)-\mathbf{x}_n(t)||^4] \leq \max(4C_1, C_3)|t-s|^2. \tag{73}$$

□

Proof. (of Lemma B.2)

Let us define the auxiliary processes

$$\widetilde{\mathbf{x}}_T(t) = \mathbf{x}_0 + \sum_{i=1}^{\lfloor t*T\rfloor}\widetilde{\boldsymbol{\Delta}}^{(T)}_{\mathbf{x}_i} + (t*T - \lfloor t*T\rfloor)\widetilde{\boldsymbol{\Delta}}^{(T)}_{\mathbf{x}_{\lceil t*T\rceil}} \tag{74}$$

$$\mathbf{x}_{T,S}(t) = \mathbf{x}_0 + \sum_{i=1}^{\lfloor t*T\rfloor}\boldsymbol{\Delta}^{(T,S)}_{\mathbf{x}_i} + (t*T - \lfloor t*T\rfloor)\boldsymbol{\Delta}^{(T,S)}_{\mathbf{x}_{\lceil t*T\rceil}}, \tag{75}$$

where

$$\widetilde{\boldsymbol{\Delta}}^{(T)}_{\mathbf{x}_k} = \mathbf{f}(\mathbf{x}_k, t_k)\Delta_{t_k} + g(t_k)\mathbf{W}(\Delta_{t_k}) \tag{76}$$

$$\boldsymbol{\Delta}^{(T,S)}_{\mathbf{x}_k} = \mathbf{f}(\mathbf{x}_k, t_k)\Delta^{(T)}_{t_k} + \sum_{i=0}^{S-1} g(t_k)\sqrt{\Delta^{(T)}_{t_k}}\mathbf{z}^{(S*T)}_{k+Ti}, \tag{77}$$

and $\mathbf{W}(\Delta_{t_k}) := \mathcal{N}(0, I*\Delta_{t_k})$. Slightly overloading our notation and letting

$$\mathbf{A}(\{t_i\}_{i=1}^k) := (\mathbf{A}(t_1), \ldots, \mathbf{A}(t_k)) \tag{78}$$

for a diffusion process $\mathbf{A}(t)$ evaluated at times $(t_1, t_2, \ldots, t_k)$, we may obtain the desired result by observing that

$$\lim_{T\to\infty} \text{CDF}[\mathbf{x}_T(\{t_i\}_{i=1}^k)] \tag{79}$$

$$= \lim_{T\to\infty} \lim_{S\to\infty} \text{CDF}[\mathbf{x}_T(\{t_i\}_{i=1}^k) + (\mathbf{x}_{T,S}(\{t_i\}_{i=1}^k) - \mathbf{x}_T(\{t_i\}_{i=1}^k))] \tag{80}$$

$$= \lim_{T\to\infty} \lim_{S\to\infty} \text{CDF}[\mathbf{x}_{T,S}(\{t_i\}_{i=1}^k)] \tag{81}$$

$$= \lim_{T\to\infty} \text{CDF}[\widetilde{\mathbf{x}}_T(\{t_i\}_{i=1}^k)] \qquad \text{Lemma B.4} \tag{82}$$

$$= \text{CDF}[\mathbf{x}(\{t_i\}_{i=1}^k)]. \qquad \text{Lemma B.5} \tag{83}$$

Next, we may interpret $\widetilde{\mathbf{x}}_T(t)$ (Eq. 74) as a variant of $\mathbf{x}_T(t)$ (Eq. (61) with "normalized" increments, which can be formally shown to be the limit of $\mathbf{x}_{T,S}(t)$ (Eq. 75) as $S \to \infty$ by the central limit theorem.

Lemma B.4. Let $\mathbf{x}_{T,S}(t)$ and $\widetilde{\mathbf{x}}_T(t)$ be defined as above. Then $\mathbf{x}_{T,S}(t)$ converges in f.d.d. to $\widetilde{\mathbf{x}}_T(t)$.

Finally, the result $\widetilde{\mathbf{x}}_T \xrightarrow{\text{f.d.d.}} \mathbf{x}$ can be shown via techniques that follow closely to the proof for the strong convergence of SDE solvers. For $i \in \{0, \ldots, n\}$ and $t \in [0, T]$ we let

$$\overline{\mathbf{x}}(t) = \sum_{k=1}^{n} \mathbf{x}_k \mathbb{1}_{t\in[t_k, t_{k+1})}(t) \tag{84}$$

be the continuous-time càdlàg extensions of the random walk $\mathbf{x}_k$. Now, $\widetilde{\mathbf{x}}_T$ can also be written as the Itô integral

$$\widetilde{\mathbf{x}}_T = \mathbf{x}(0) + \int_0^t f(\overline{\mathbf{x}}(s))ds + \int_0^t g\left(\frac{\lfloor s * T \rfloor}{n}\right) dW_s. \tag{85}$$

Of course, the solution $\mathbf{x}$ to Eq. 1 can also be expressed in the similar form

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t f(\mathbf{x}(s))ds + \int_0^t g(s)dW_s. \tag{86}$$

Now we may state the following lemma.

Lemma B.5. Let $\widetilde{\mathbf{x}}_T$ be defined as above and Assumption 1 hold. Then $\widetilde{\mathbf{x}}_T$ converges to $\mathbf{x}$ in finite dimensional distributions (f.d.d.).

□

Proof. (of Lemma B.4)

Observe that Eq. (74) can be seen as the continuous-time interpolation of the random walk

$$\widetilde{\mathbf{x}}_{k+1}^{(T)} = \widetilde{\mathbf{x}}_k^{(T)} + \mathbf{f}(\widetilde{\mathbf{x}}_k^{(T)}, t_k)\Delta_{t_k} + g(t_k)\sqrt{\Delta_{t_k}}\mathbf{W}(\Delta_{t_k}), \tag{87}$$

and Eq. (75) of the random walk

$$\mathbf{x}_{k+1}^{(T,S)} = \mathbf{x}_k^{(T,S)} + \mathbf{f}(\mathbf{x}_k^{(T,S)}, t_k)\Delta_{t_k} + \sum_{i=1}^{S} g(t_k)\sqrt{\Delta_{t_k}^{(T)}}\mathbf{z}_{k+Ti}^{(S*T)}. \tag{88}$$

Applying the Central Limit Theorem, we may see that

$$\mathbf{A}_k := \sum_{i=1}^{S} g(t_k)\sqrt{\Delta_{t_k}^{(T)}}\mathbf{z}_{k+Ti}^{(S*T)} \xrightarrow{\mathcal{D}} g(t)\sqrt{\Delta_{t_k}}\mathbf{W}(\Delta_{t_k})$$

for each $0 \le k < T$. We now show our result by recursion. In the base case we have that $\mathbf{x}_0^{(T,S)} := \widetilde{\mathbf{x}}_0^{(T)} := \mathbf{x}_0$, so clearly $\mathbf{x}_0^{(T,S)} \xrightarrow{\mathcal{D}} \widetilde{\mathbf{x}}_0^{(T)}$. For any subsequent $k + 1 > 0$, we may invoke Slutsky's Theorem on the independent sequences $\mathbf{x}_k^{(T,S)} \xrightarrow{\mathcal{D}} \widetilde{\mathbf{x}}_k^{(T)}$ and $\mathbf{A}_k \xrightarrow{\mathcal{D}} g(t)\sqrt{\Delta_{t_k}}\mathbf{W}(\Delta_{t_k})$ to obtain

$$\mathbf{x}_{k+1}^{(T,S)} := \mathbf{x}_k^{(T,S)} + \mathbf{A}_k \xrightarrow{\mathcal{D}} \widetilde{\mathbf{x}}_k^{(T)} + g(t)\sqrt{\Delta_{t_k}}\mathbf{W}(\Delta_{t_k}) =: \widetilde{\mathbf{x}}_{k+1}^{(T)}. \tag{89}$$

Therefore, we have that $\mathbf{x}_{k+1}^{(T,S)} \xrightarrow{\mathcal{D}} \mathbf{x}_{k+1}^{(T)}$ for all k. Since Eqs. 74 and 75 are purely functions of t and their respective random walks (Eqs. 87 and 88), we have our result. □

Proof. (of Lemma B.5) Let us define

$$\boldsymbol{\epsilon}(t) = \sup_{0\leq s\leq t} \mathbb{E}\left[\left|\left|\mathbf{Y}_n(s) - \mathbf{x}(s)\right|\right|^2\right]. \tag{90}$$

Recalling the definitions $\overline{\mathbf{x}}(t) := \mathbf{x}_{\lfloor nt \rfloor}$ and $\bar{\mathbf{z}}(t) = \mathbf{z}_{\lfloor nt \rfloor}$, we have

$$\begin{aligned}
\boldsymbol{\epsilon}(t) &= \sup_{0\leq s\leq t} \mathbb{E}\left[\left|\left|\int_0^s [f(\mathbf{x}(u)) - f(\overline{\mathbf{x}}(u))]du + \int_0^s \left[g\left(\frac{\lfloor n\cdot u\rfloor}{n}\right) - g(u)\right] dW_u\right|\right|^2\right] \\
&\leq 4 \sup_{0\leq s\leq t} \mathbb{E}\left[\left|\left|\int_0^s [f(\mathbf{x}(u)) - f(\overline{\mathbf{x}}(u))]du\right|\right|^2 + \left|\left|\int_0^s \left[g\left(\frac{\lfloor n\cdot u\rfloor}{n}\right) - g(u)\right] dW_u\right|\right|^2\right].
\end{aligned}$$

Invoking the Itô isometry, Cauchy-Schwarz inequality, and linearity of expectations,

$$\boldsymbol{\epsilon}(t) \leq 4 \sup_{0\leq s\leq t} \left(t\mathbb{E}\left[\int_0^s \left|\left|f(\mathbf{x}(u)) - f(\overline{\mathbf{x}}(u))\right|\right|^2 du\right] + \mathbb{E}\left[\int_0^s \left|\left|g\left(\frac{\lfloor n\cdot u\rfloor}{n}\right) - g(u)\right|\right|^2 du\right]\right).$$

We now leverage Assumption 1 to obtain

$$\boldsymbol{\epsilon}(t) \leq 4K \sup_{0\leq s\leq t} \left(t\mathbb{E}\left[\int_0^s \left|\left|\mathbf{x}(u)) - \overline{\mathbf{x}}(u)\right|\right|^2 du\right] + \mathbb{E}\left[\int_0^s \frac{1}{n^2} du\right]\right).$$

Applying Theorem 4.5.4 in (Kloeden & Platen, 1992) and folding all constants that depend on $T, \mathbb{E}[X_0], K$ into C, we have

$$\boldsymbol{\epsilon}(t) \leq C\left(\int_0^s \boldsymbol{\epsilon}(u)du + \Delta t\right),$$

which, by Gronwall's inequality, results in the bound

$$\sup_{0\leq s\leq T} \mathbb{E}\left[\left|\mathbf{Y}_n(s) - \mathbf{x}(s)\right|^2\right] = \boldsymbol{\epsilon}(T) \leq C\Delta t. \tag{91}$$

Now, fix k and choose times $t_1, \ldots, t_k$. We see that

$$\begin{aligned}
\mathbb{E}[||(\mathbf{Y}_n(t_1), \ldots, \mathbf{Y}_n(t_k)) - (\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_k))||^2] &\leq \sum_{i=1}^k \mathbb{E}[||\mathbf{Y}_n(t_i) - \mathbf{Y}(t_i)||^2] \\
&\leq k \sup_{0\leq s\leq T} \mathbb{E}\left[\left|\mathbf{Y}_n(s) - \mathbf{x}(s)\right|^2\right] \\
&\leq kC\Delta t \to 0
\end{aligned}$$

as $\Delta t \to 0$. This shows $(\mathbf{Y}_n(t_1), \ldots, \mathbf{Y}_n(t_k)) \xrightarrow{\mathcal{L}_2} (\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_k))$, which implies the desired result. □

C. Implementation

We use directly with no changes the models and training protocols in (Kingma et al., 2021) to parameterize our score network $\boldsymbol{\epsilon}(\mathbf{x}, t)$ to evaluate the log-likelihoods of our proposed diffusion models. To evaluate FID, we instead use the architecture and training procedures in (Karras et al., 2022), again with no changes. All training is performed on NVIDIA RTX A6000 GPUs.