Non-Normal Diffusion Models

Henry Li

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 $\Delta x_k := x_k - x_{k+1}$. 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 $\Delta x_k$ 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; Dhariwal 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 $\Delta x_k := x_k - x_{k+1}$ 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 $\Delta x_k$.

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.

2. Background

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b) take the form $p_\theta(x) = \int p_\theta(x_{0:T}) \, dx_{1:T}$ where data $x_0 := x$ are related to a set of latent variables $x_{1:T} := (x(t_1), \ldots, x(t_T))$ distributed as marginals of a diffusion process governed by an Itô stochastic differential equation (SDE)

$$\frac{dx}{dt} = f(x,t) \, dt + g(t) \, dw$$

(1)

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

$$dx = \left[ f(x,t) - g(t)^2 \nabla_x \log p(x,t) \right] \, dt + g(t) \, dw,$$

(2)

where $w$ 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 $\nabla_x \log p(x,t)$ with a neural network $s_\theta(x,t)$ 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

\[
p_\theta(x_0:T) = p(x_T) \prod_{k=0}^{T-1} p_\theta(x_k|x_{k+1})
\]

(3)
or equivalently

\[
p_\theta(x_0:T) = p(x_T) \prod_{k=0}^{T-1} p_\theta(\Delta x_k|x_{k+1}),
\]

(4)

where \( \nu_\theta, \rho_\theta \) are Markov models and \( \Delta x_k := x_{k+1} - x_k \).

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

\[
\log p_\theta(x) \geq \mathbb{E}_q \left[ \log p(x_0|x_1) \right]_{\mathcal{L}_0} - \sum_{k=1}^T KL(q(\Delta x_k|x_{k+1})||p_\theta(\Delta x_k|x_{k+1})) - KL(q(x_T)||p(x_T)) \right] \quad (5)
\]

that reduces to a simple function of \( s_\theta(x, t) \).

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 \( \Delta x_k \) (Sohl-Dickstein et al., 2015; Song et al., 2020b; Särkkä & Solin, 2019), we shall extend this result to arbitrarily distributed \( \Delta x_k \) in Section 3.

Ultimately, either choice of learning \( s_\theta(x, t) \) allows for unbiased estimates of \( \log p_\theta(x) \) 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)

\[
dx = \left[ f(x, t) - \frac{1}{2} g(t)^2 \nabla_x \log p(x, t) \right] dt,
\]

(6)

and substituting the score with \( s_\theta(x, t) \).

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 \( x_k \). Let \( x(t) \) be the solution to Eq. (1). Intuitively, one might expect a similar convergence of \( x_k \) to \( x(t) \) if we constrain the first and second moments of its increments \( \Delta x_k := x_{k+1} - x_k \) to

\[
\mathbb{E}[\Delta x_k|x_k] = f(x_k, t_k)\Delta t_k,
\]

\[
\text{Var}(\Delta x_k|x_k) = g(t_k)^2 \Delta t_k.
\]

(7)

This type of convergence has been previously explored for normally distributed \( \Delta x_k \) 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 \( \Delta x_k \) 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 \( \Delta x_k \) 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 \( x_k \) 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 \( f \) and \( g \), the time step \( \Delta t_k \), and a sequence of independent variables \( z_k \).

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

\[
\Delta x_k = f(x_k, t_k)\Delta t_k + g(t_k)\sqrt{\Delta t_k}z_k,
\]

(8)

where \( \mathbb{E}[z_k] = 0 \), \( \text{Var}(z_k) = 1 \), \( \Delta t_k := t_{k+1} - t_k \), and \( f, g \) 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 \( f(x, t) \) is linear in \( x \), 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 \( k \in \{0, \ldots, T\} \).
### Table 1. Summary of the diffusion models proposed in Section 4. *A = [μ₁ - b₁, μ₁ + b₁]. **rₖ := ε - εₖ(xₖ, tₖ).

| p | q | KL(p(x)||q(x)) | 𝓉ₖ (note: **) |
|---|---|---|---|
| N(μ₁, σ²) | N(μ₂, σ²) | exp \(\frac{1}{2σ²}||μ₁ - μ₂||² + \frac{1}{2}E[exp(-vk|r_k|)] \) | \(w_k E[|r_k|^2] \) |
| Laplace(μ₁, σ²) | Laplace(μ₂, σ²) | exp \(\frac{1}{2\sigma}||μ₁ - μ₂||² + \frac{1}{2}E[exp(-vk|r_k|)] \) | \(w_k E[|r_k|^2] \) |
| Uniform[μ₁ - √3σ, μ₁ + √3σ] | N(μ₂, σ²) | \(\frac{1}{2}(\frac{σ}{σ^2}(μ₁ - μ₂)^2 + log \frac{σ}{σ^2} + 1) + E[exp(-vk|r_k|)] \) | \(w_k E[|r_k|^2] + \frac{1}{2}(1 + log \frac{σ}{σ^2}) \) |
| Uniform[μ₁ - √3σ, μ₁ + √3σ] | Laplace(μ₂, σ²) | \(\frac{1}{2}|μ₁ - μ₂|^2 \) if \(ε_θ(x, t) \in A \)
| | | \(w_k E[|r_k|^2] + \frac{1}{2} \) if \(ε_θ(x, t) \notin A \) | |
maximizing the likelihood of a standard diffusion model with Gaussian increments minimizes a quadratic error term between the score function \( s_\theta(x_k, t_k) = \frac{1}{\sigma(t_k)} \varphi(x_k, t_k) \).

### 4.2. Laplace \( q \) and \( p_\theta \)

We now consider the case of Laplace distributed \( \Delta x_k \). Invoking Theorem 3.3, we can derive the alternative loss

\[
\mathcal{L}_k = \mathbb{E}_x \left[ \exp \left( -v_k \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_1 \right) - 1 + v_k \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_1 \right],
\]

(10)

where \( v_k := \sqrt{w_k} \).

While the term in the expectation \( d(\varepsilon - \varepsilon_\theta(x_k, t_k)) := \exp \left( -v_k \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_1 \right) - 1 + v_k \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_1 \) appears somewhat opaque, we can see that it converges to a weighted \( L1 \) norm of the error \( r_k := \varepsilon - \varepsilon_\theta(x_k, t_k) \) under two conditions:

\[
\lim_{t_k \to 0} v_k \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_1 = 1, \quad (11)
\]

i.e., when \( t \) is small, and

\[
\lim_{|r_k|_1 \to \infty} v_k \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_1 = 1 \quad (12)
\]

i.e., when \( |r_k|_1 \) is large.

#### 4.3. Uniform \( q \), Gaussian \( p_\theta \)

Next, we note that \( q \) and \( p_\theta \) 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

\[
\mathcal{L}_k = w_k \mathbb{E}_\varepsilon \left[ \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_2^2 + C, \right]
\]

(13)

where \( C = \frac{1}{2} \left( 1 + \log \frac{e}{2} \right) \approx 0.34 \) 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 \( p_\theta \) and \( q \) are not from the same family of distributions.

#### 4.4. Uniform \( q \), Laplace \( p_\theta \)

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 \( p_\theta \) is the Laplace distribution. This configuration of distributions produces the piecewise loss

\[
\mathcal{L}_k = \begin{cases} 
  w_k \mathbb{E}_\varepsilon \left[ \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_2^2 + \frac{1}{2} \right] & \text{if } \varepsilon_\theta(x, t) \in A \\
  v_k \mathbb{E}_\varepsilon \left[ \| \varepsilon - \varepsilon_\theta(x_k, t_k) \|_1 \right] & \text{if } \varepsilon_\theta(x, t) \notin A,
\end{cases}
\]

(14)

where \( A = [\mu_1 - \sqrt{3} \sigma w_k, \mu_2 + \sqrt{3} \sigma w_k] \). 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 \( r_k := \varepsilon - \varepsilon_\theta(x_k, t_k) \) is large. Both imply that \( \varepsilon_\theta(x, t) \notin A \).

### 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 \( \Delta x_k := x_k - x_{k+1} \). Freeing
Non-Normal Diffusion Models

up the distributional assumption on $\Delta x_k$ 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


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($\mu_1$, $b_1$) and Laplace($\mu_2$, $b_2$), i.e.,

\[
p(x) = \frac{1}{2b_1} \exp\left(-\frac{|x - \mu_1|}{b_1}\right)
\]

\[
q(x) = \frac{1}{2b_2} \exp\left(-\frac{|x - \mu_2|}{b_2}\right).
\]

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

\[
KL(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \log p(x) dx - \int_{-\infty}^{\infty} p(x) \log q(x) dx
\]

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

\[
\int_{-\infty}^{\infty} \frac{|x - \mu_2|}{2b_1b_2} \exp\left(-\frac{|x - \mu_1|}{b_1}\right) dx = \int_{-\infty}^{\mu_2} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{\mu_1 - x}{b_1}\right) dx + \int_{\mu_2}^{\infty} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{\mu_1 - x}{b_1}\right) dx
\]

\[
= \int_{-\infty}^{\mu_2} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{x - \mu_1}{b_1}\right) dx + \int_{\mu_2}^{\infty} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{x - \mu_1}{b_1}\right) dx
\]

\[
= \int_{-\infty}^{\mu_2} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{x - \mu_1}{b_1}\right) dx + \int_{\mu_2}^{\infty} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{x - \mu_1}{b_1}\right) dx
\]

\[
= \frac{\mu_1 - \mu_2}{b_2} + \frac{b_1}{b_2} \exp\left(-\frac{\mu_1 - \mu_2}{b_1}\right),
\]

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

\[
\int_{-\infty}^{\infty} \frac{|x - \mu_2|}{2b_1b_2} \exp\left(-\frac{|x - \mu_1|}{b_1}\right) dx
\]

\[
= \int_{-\infty}^{\mu_2} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{x - \mu_1}{b_1}\right) dx + \int_{\mu_2}^{\infty} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{x - \mu_1}{b_1}\right) dx
\]

\[
+ \int_{\mu_2}^{\infty} \frac{x - \mu_1}{2b_1b_2} \exp\left(-\frac{x - \mu_1}{b_1}\right) dx
\]

\[
= \left[ \frac{\mu_2 - \mu_1 + b_1}{2b_2} \right] + \left[ \frac{b_1}{2b_2} \exp\left(-\frac{\mu_2 - \mu_1}{b_1}\right) \right] + \left[ \frac{\mu_2 - \mu_1}{2b_2} \right]
\]

\[
= \frac{\mu_2 - \mu_1}{b_2} + \frac{b_1}{b_2} \exp\left(-\frac{\mu_2 - \mu_1}{b_1}\right),
\]

Non-Normal Diffusion Models
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
\[ KL(p(x) || q(x)) = \frac{b_1}{b_2} \exp \left( - \frac{|\mu_2 - \mu_1|}{b_1} \right) + \frac{b_1}{b_2} + \log \frac{b_2}{b_1} - 1. \tag{20} \]

**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
\[ p(x) = \mathbb{1}_{x \in [\mu_1 - b_1, \mu_1 + b_1]} \frac{1}{2b_1}, \tag{21} \]
\[ q(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( - \frac{(x - \mu_2)^2}{2\sigma^2} \right). \tag{22} \]

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 \(*\*\):
\[ - \int_{-\infty}^{\infty} p(x) \log q(x) \, dx = \int_{\mu_1 - b_1}^{\mu_1 + b_1} \frac{1}{2b_1} \left( \log(\sigma \sqrt{2\pi}) + \frac{1}{2\sigma^2} (x - \mu_2)^2 \right) \, dx 
= \log(\sigma \sqrt{2\pi}) + \int_{\mu_1 - b_1}^{\mu_1 + b_1} \frac{1}{4b_1 \sigma^2} (x - \mu_2)^2 \, dx 
= \log(\sigma \sqrt{2\pi}) + \frac{1}{4b_1 \sigma^2} \left( \frac{1}{3} x^3 - \mu_2 x^2 + \mu_2^2 x \right) \bigg|_{\mu_1 - b_1}^{\mu_1 + b_1} 
= \log(\sigma \sqrt{2\pi}) + \frac{1}{4b_1 \sigma^2} \left( 2b_1 \left( (\mu_1 - \mu_2)^2 + \frac{1}{3} \sigma^2 \right) \right) \]

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\(\mu_2, b_2\) 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 *:

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

$$
= \frac{\log(2b_1)}{2b_1} (\mu_1 - b_1) - \frac{\log(2b_1)}{2b_1} (\mu_1 + b_1)
$$

$$
= - \log(2b_1).
$$

Turning to the cross-entropy term **:

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

$$
= \log(2b_2) + \int_{\mu_1-b_1}^{\mu_1+b_1} \frac{1}{2b_1} \frac{|x - \mu_2|}{b_2} dx.
$$

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

$$
\int_{\mu_1-b_1}^{\mu_1+b_1} \frac{1}{2b_1} \frac{|x - \mu_2|}{b_2} dx = \int_{\mu_1-b_1}^{\mu_1+b_1} \frac{1}{2b_1} b_2 \frac{x - \mu_2}{b_2} dx
$$

$$
= \frac{1}{2b_1 b_2} \left( \frac{1}{2} x^2 - \mu_2 x \right) \bigg|_{\mu_1-b_1}^{\mu_1+b_1}
$$

$$
= \frac{1}{2b_1 b_2} \left( 2b_1 (\mu_1 - \mu_2) \right),
$$

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

$$
\int_{\mu_1-b_1}^{\mu_1+b_1} \frac{1}{2b_1} \frac{|x - \mu_2|}{b_2} dx = \int_{\mu_1-b_1}^{\mu_1+b_1} \frac{1}{2b_1} b_2 \frac{x - \mu_2}{b_2} dx
$$

$$
= \frac{1}{2b_1 b_2} \left( \mu_2 x - \frac{1}{2} x^2 \right) \bigg|_{\mu_1-b_1}^{\mu_1+b_1}
$$

$$
= \frac{1}{2b_1 b_2} \left( 2b_1 (\mu_2 - \mu_1) \right).
$$

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

$$
\int_{\mu_1-b_1}^{\mu_1+b_1} \frac{1}{2b_1 b_2} |x - \mu_2| dx = \frac{1}{2b_1 b_2} \left[ \int_{\mu_1-b_1}^{\mu_2} (\mu_2 - x) dx + \int_{\mu_2}^{\mu_1+b_1} \frac{1}{b_2} (x - \mu_2) dx \right]
$$

$$
= \frac{1}{2b_1 b_2} \left[ \left( \frac{1}{2} (\mu_1 - \mu_2)^2 + \frac{1}{b_2^2} \right) + \left( \frac{1}{2} (\mu_1 - b_1)^2 + \frac{1}{b_2^2} + b_1 (\mu_2 - \mu_1) \right) \right]
$$

$$
= \frac{1}{2b_1 b_2} \left[ (\mu_1 - \mu_2)^2 + b_1^2 \right].
$$

Combining the cases, we obtain the KL divergence

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

(26)
Non-Normal Diffusion Models

A.4. Deriving $L_k$

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

$$f_\theta = f(x, t) - \frac{1}{2}g(t)\nabla_x \log p(x, t),$$

$$\hat{f}_\theta = f(x, t) - \frac{1}{2}g(t)\sigma_\theta(x, t),$$

where $f$ 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 $\Delta_{x_k}$ be normally distributed, i.e.,

$$p_\theta(\Delta_{x_k} | x_k) = \mathcal{N} \left( \Delta_{x_k}; \hat{f}_\theta(x_k, t_k) \Delta_{t_k}, g(t_k^2) \Delta_{t_k} \right),$$

$$q(\Delta_{x_k} | x_k) = \mathcal{N}(\Delta_{x_k}; f_\theta(x_k, t_k) \Delta_{t_k}, g(t_k^2) \Delta_{t_k}).$$

Then

$$L_k = w_k \mathbb{E}_{\epsilon \sim q}[\epsilon - \epsilon_\theta(x_k, t_k)]^2,$$

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,

$$L_k = KL(p_\theta(\Delta_{x_k} | x_k) || q(\Delta_{x_k} | x_k))$$

$$= \mathbb{E} \left[ \frac{||\mu_{p_k}(x_k) - \mu_{q_k}(x_k)||^2}{2\sigma^2} \right].$$

Since $\mu_{q_k} = (f(x_k, t_k) - g(t_k^2)\nabla \log p(x_k))\Delta_{t_k}$, $\mu_{p_k} = (f(x_k, t_k) - g(t_k^2)\nabla \log p_\theta(x_k))\Delta_{t_k}$, and $\sigma_{p_k} = \sigma_{q_k} = g(t_k)\sqrt{\Delta_{t_k}}$, we have

$$L_k = \frac{1}{2} \mathbb{E} \left[ \frac{||g(t_k^2)\nabla \log p(x_k) - g(t_k^2)\nabla \log p_\theta(x_k)||^2 \Delta_{t_k}^2}{g(t_k^2)\Delta_{t_k}} \right]$$

$$= \frac{1}{2} \mathbb{E} \left[ g(t_k^2)||\nabla \log p(x_k) - \nabla \log p_\theta(x_k)||^2 \Delta_{t_k} \right].$$

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

$$L_k = w_k \mathbb{E} [||\epsilon - \epsilon_\theta(x_k, t_k)||^2],$$

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 $\Delta_{x_k}$ be Laplace distributed, i.e.,

$$p_\theta(\Delta_{x_k} | x_k) = \text{Laplace} \left( \Delta_{x_k}; \hat{f}_\theta(x_k, t_k) \Delta_{t_k}, g(t_k^2) \Delta_{t_k} \right),$$

$$q(\Delta_{x_k} | x_k) = \text{Laplace} \left( \Delta_{x_k}; f_\theta(x_k, t_k) \Delta_{t_k}, g(t_k^2) \Delta_{t_k} \right).$$

Then, letting $r_k := \epsilon - \epsilon_\theta(x_k, t_k)$,

$$L_k = \exp (-w_k||r_k||_1) - 1 + w_k||r_k||_1,$$

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

\hfill \Box
Lemma A.3. Let \( p(\Delta_{x_k}|x_{k+1}) \) be normally distributed and \( q(\Delta_{x_k}|x_{k+1}) \) be the uniform distribution on the interval \([\mu_1-\sqrt{3}\sigma, \mu_1+\sqrt{3}\sigma]\), i.e.,
\[
p_\theta(\Delta_{x_k}|x_{k}) = \mathcal{N}\left(\Delta_{x_k}; \hat{f}_\theta(x_k, t_k) \Delta_{t_k}, g(t_k)^2 \Delta_{t_k}\right) \
q(\Delta_{x_k}|x_{k}) = \text{Uniform}\left(\Delta_{x_k}; f_\theta(x_k, t_k) \Delta_{t_k}, g(t_k)^2 \Delta_{t_k}\right).
\]
Then
\[
L_k = w_k \mathbb{E}_x ||\epsilon - \epsilon_\theta(x_k, t_k)||^2 + C,
\]
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
\[
L_k = KL(p(\Delta_{x_k}|x_{k})||q(\Delta_{x_k}|x_{k}))
= \mathbb{E} \left[ ||\mu_{p, k}(x_{k}) - \mu_{q, k}(x_{k})||^2 \right] + C,
\]
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 \( p(\Delta_{x_k}|x_{k+1}) \) be Laplace distributed and \( q(\Delta_{x_k}|x_{k+1}) \) be the uniform distribution on the interval \([\mu_1-\sqrt{3}\sigma, \mu_1+\sqrt{3}\sigma]\), i.e.,
\[
p_\theta(\Delta_{x_k}|x_{k}) = \text{Laplace}\left(\Delta_{x_k}; \hat{f}_\theta(x_k, t_k) \Delta_{t_k}, g(t_k)^2 \Delta_{t_k}\right) \
q(\Delta_{x_k}|x_{k}) = \text{Uniform}\left(\Delta_{x_k}; f_\theta(x_k, t_k) \Delta_{t_k}, g(t_k)^2 \Delta_{t_k}\right).
\]

Proof. Plugging in the closed form solution to the KL divergence between two Laplace distributions into the likelihood lower bound (Appendix A.1),
\[
L_k = KL(p(\Delta_{x_k}|x_{k})||q(\Delta_{x_k}|x_{k}))
= \exp \left( \frac{-||\mu_{p, k} - \mu_{q, k}||_1}{\sigma_p} \right) - 1 + \frac{||\mu_{p, k} - \mu_{q, k}||_1}{\sigma_p}.
\]
Observe that \( d_k \) can be simplified as
\[
d_k = \frac{||\mu_{p, k} - \mu_{q, k}||_1}{\sigma_p}
= g(t_k)^2 \sigma \|\nabla_x p(x_k) - \nabla_x q(x_k)||_1
= \frac{\sqrt{\Delta t_k} g(t_k) \|\epsilon - \epsilon_\theta(x_k, t_k)||_1}{\sigma(t_k)}
= v_k \|\epsilon - \epsilon_\theta(x_k, t_k)||_1, \tag{36}
\]
where \( v_k := \frac{g(t_k)}{\sigma(t_k)} \sqrt{\Delta t_k} \). Therefore,
\[
L_k = (-v_k \|\epsilon - \epsilon_\theta(x_k, t_k)||_1) - 1 + v_k \|\epsilon - \epsilon_\theta(x_k, t_k)||_1.
\]
Then

\[
\mathcal{L}_k = \begin{cases} 
 w_k \mathbb{E}[e - e_\theta(x_k, t_k)]^2 + \frac{1}{2} & \mu_2 \in [\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma] \\
 v_k \mathbb{E}[e - e_\theta(x_k, t_k)]^1 & \mu_2 \not\in [\mu_1 - \sqrt{3}\sigma w_k, \mu_1 + \sqrt{3}\sigma w_k]
\end{cases},
\]

(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

\[
\mathcal{L}_k = \begin{cases} 
 \mathbb{E} \left[ \frac{||\mu_{x_k} - \mu_{x_{k+1}}||^2}{\sigma^2} \right] + \frac{1}{2} & \mu_2 \in [\mu_1 - \sqrt{3}\sigma, \mu_1 + \sqrt{3}\sigma] \\
 \mathbb{E} \left[ ||\mu_{x_k} - \mu_{x_{k+1}}||^1 / \sigma \right] & \mu_2 \not\in [\mu_1 - \sqrt{3}\sigma w_k, \mu_1 + \sqrt{3}\sigma w_k]
\end{cases}.
\]

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

**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 \(x_k\) be a structured random walk and \(f(x, t_k) = \beta(t_k)x\) be linear. Then

\[
\mathbb{E}[x_k] = \alpha_k x_0 \quad \text{and} \quad \sigma(t_k)^2 := \text{Var}(x_k) = \gamma_k,
\]

where \(\alpha_k = \prod_{i=1}^k (1 + \beta_i)\) and \(\gamma_k = \sum_{i=1}^k \left(\frac{\alpha_k}{\alpha_{i+1}} g_i\right)^2\). 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 \(\mathbb{E}[x_k]\). Observe that

\[
\mathbb{E}[x_k] = \mathbb{E}[x_{k-1} + \Delta x_k] = \mathbb{E} \left[ x_{k-1} + f(x_{k-1}, t_k) \Delta t_k + g(t_k) \sqrt{\Delta t_k} \right] = \mathbb{E}[x_{k-1}] (1 + \beta(t_k) \Delta t_k).
\]

Applying this operation \(k - 1\) more times, we obtain

\[
\mathbb{E}[x_k] = \mathbb{E}[x_0] \prod_{i=1}^k (1 + \beta(t_i) \Delta t_k).
\]

Turning to \(\text{Var}(x_k)\), we first note that

\[
\mathbb{E}[x_k^2] = \mathbb{E}[\{(x_{k-1} + \Delta x_k)^2\}] = \mathbb{E}[x_{k-1}^2] + \mathbb{E}[\Delta x_k^2] + 2\mathbb{E}[x_{k-1} \Delta x_k]
\]

where

\[
\mathbb{E}[x_{k-1} \Delta x_k] = \mathbb{E} \left[ x_{k-1} \left( f(x_{k-1}, t_k) \Delta t_k + g(t_k) \sqrt{\Delta t_k} \right) \right] = \beta(t_k) \Delta t_k \mathbb{E}[x_{k-1}^2] + g(t_k) \sqrt{\Delta t_k} \mathbb{E}[x_{k-1} \Delta x_k] = \beta(t_k) \Delta t_k \mathbb{E}[x_{k-1}^2] + g(t_k) \sqrt{\Delta t_k} \mathbb{E}[x_{k-1} \Delta x_k]
\]

Then

\[
\mathcal{L}_k = \left\{ \begin{array}{l}
 w_k \mathbb{E}[e - e_\theta(x_k, t_k)]^2 + \frac{1}{2} \\
 v_k \mathbb{E}[e - e_\theta(x_k, t_k)]^1
\end{array} \right\},
\]

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

\[ E[\Delta^2_{x_k}] = E \left[ (f(x_{k-1}, t_k) \Delta_{t_k})^2 + (g(t_k) z_k \sqrt{\Delta_{t_k}})^2 + 2(f(x_{k-1}, t_k) \Delta_{t_k}) (g(t_k) z_k \sqrt{\Delta_{t_k}}) \right] \]

\[ = \beta (t_k)^2 \Delta^2_{t_k} E[x^2_{k-1}] + g(t_k)^2 \Delta^2_{t_k} E[z^2_k] + \beta (t_k) g(t_k) \Delta^2_{t_k} E[x_{k-1} z_k] \]

Putting things together, we have

\[ E[x^2_k] = E[(x_{k-1} + \Delta_{x_k})^2] \]

\[ = E[x^2_{k-1}] + E[\Delta^2_{x_k}] + 2E[x_{k-1} \Delta_{x_k}] \]

\[ = E[x^2_{k-1}] \left( 1 + \beta (t_k)^2 \Delta^2_{t_k} + 2 \beta (t_k) \Delta_{t_k} \right) + g(t_k)^2 \Delta^2_{t_k} \]

This gives, by induction,

\[ E[x^2_k] = E[x^2_0] \prod_{i=1}^{k} (1 + \beta (t_i) \Delta_{t_i})^2 + \sum_{j=1}^{k} \prod_{i=j+1}^{k} (1 + \beta (t_i) \Delta_{t_i})^2 g(t_j)^2 \Delta_{t_j} \]

Finally, we can write

\[ \text{Var}(x_k) = E[x^2_k] - E[x_k]^2 \]

\[ = \text{Var}(x_0) \prod_{i=1}^{k} (1 + \beta (t_i) \Delta_{t_i})^2 + \sum_{j=1}^{k} \prod_{i=j+1}^{k} (1 + \beta (t_i) \Delta_{t_i})^2 g(t_j)^2 \Delta_{t_j} \]

Assuming that \( \text{Var}(x_0) = 0 \), we now have

\[ \sigma^2_k := \text{Var}(x_k) = \sum_{j=1}^{k} \prod_{i=j+1}^{k} (1 + \beta (t_i) \Delta_{t_i})^2 g(t_j)^2 \Delta_{t_j}. \] (43)

\[ \square \]

B.2. Deriving Previous Methods in Our Framework

**Corollary 3.1.** Let \( x_k \) be a structured random walk, \( z_k \) be normally distributed, and \( f(x, t) = \beta(t)x \) 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

\[ p_0(x_{k+1}|x_k) = \mathcal{N}(x_{k+1}; \sqrt{1-\beta_k} x_k, \beta_k I), \] (44)

and thus that \( x_{k+1} \) may be written in terms of \( x_k \) as

\[ x_{k+1} = \sqrt{1-\beta_k} x_k + \sqrt{\beta_k} \epsilon, \] (45)

where \( \epsilon \sim \mathcal{N}(0, I) \). Subtracting \( x_k \) from both sides and leveraging the fact that \( x_k \) and \( x_{k-1} \) are both normally distributed, we obtain

\[ \Delta x_k = (\sqrt{1-\beta_k} - 1)x_k + \sqrt{\beta_k} \epsilon. \] (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

\[ \tilde{\alpha}_k = \prod_{i=1}^{k} (\sqrt{1-\beta_i}), \quad \tilde{\gamma}_k = \sum_{i=1}^{k} \left( \frac{\tilde{\alpha}_k}{\tilde{\alpha}_{i+1}} \right)^2 \beta_i. \] (47)

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

\[ p(x_k|x_0) = \mathcal{N}(x_k; \tilde{\alpha}_k x_0, (1 - \tilde{\alpha}^2) I). \] (48)
Non-Normal Diffusion Models

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

\[ p(x_t) = \mathcal{N} \left( \frac{\alpha_{k+1}}{\alpha_k} x_k, \sigma^2_{k+1} - \left( \frac{\alpha_{k+1}}{\alpha_k} \right)^2 \sigma^2_k \right), \]  

(49)

where \( \alpha_k \) and \( \sigma_k \) are related to each other by a monotonic function \( \gamma(t) \)

\[ \alpha_k^2 = \text{sigmoid}(\gamma(t)), \]  

(50)

\[ \sigma_k^2 = \text{sigmoid}(\gamma(t)). \]  

(51)

According to Eq. 49, \( x_{k+1} \) can be written in terms of \( x_k \) as

\[ x_{k+1} = \frac{\alpha_{k+1}}{\alpha_k} x_k + \left( \sigma^2_{k+1} - \left( \frac{\alpha_{k+1}}{\alpha_k} \right)^2 \sigma^2_k \right) \epsilon, \]  

(52)

where \( \epsilon \sim \mathcal{N}(0, I) \). Subtracting off \( x_k \) on both sides, we obtain

\[ \Delta x_{k+1} = \left( \frac{\alpha_{k+1}}{\alpha_k} - 1 \right) x_k + \left( \sigma^2_{k+1} - \left( \frac{\alpha_{k+1}}{\alpha_k} \right)^2 \sigma^2_k \right) \epsilon. \]  

(53)

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

\[ \bar{\alpha}_k = \prod_{i=1}^{k} \frac{\alpha_{i+1}}{\alpha_i} = \alpha_k \]  

(54)

and

\[ \bar{\gamma}_k = \sum_{i=1}^{k} \frac{\alpha_k}{\alpha_i} \left( \frac{\sigma^2_{i+1} - \left( \frac{\alpha_i}{\alpha_{i-1}} \right)^2 \sigma^2_i}{\sigma^2_{i+1}} \right) \]  

\[ = \sum_{i=1}^{k} \frac{\alpha_k}{\alpha_i} \sigma^2_{i+1} - \alpha_k \sigma^2_{i+1} \]  

(55)

\[ = \sigma^2_k, \]  

(56)

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

\[ ||f(x) - f(y)|| \leq K ||x - y||, \quad \text{and} \quad |g(t) - g(s)| \leq K |t - s|. \]  

(58)

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

\[ ||f(x)|| \leq K (1 + ||x||), \quad \text{and} \quad |g(t)| \leq K (1 + |t|). \]  

(59)

**Assumption 3** (Integrability of \( z_k \)). There exists \( K \in \mathbb{R} \) such that

\[ \mathbb{E}[||z_k||^4] < K. \]  

(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 $\Delta x_k$, which is heavily utilized in the original proof. By exploiting the structural properties of Definition 1, we can decompose $x_k$ into a set of auxiliary processes with the same limit, which we can show to converge to $X$ 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 $\{x_k\}_{k=1}^n$ is a structured random walk on $\mathbb{R}^d$. Let $X_T(t) = x_0 + \sum_{k=1}^{n} \Delta x_k$ be the continuous-time càdàg extension of $x_k$, where $n_t = \lfloor t \ast T \rfloor$. Then $X_T$ converges in distribution to $x(t)$, as $\Delta t_k \to 0$.

**Proof.** Using Eq. 8 we may define the continuous-time extension of $x_k$ as the process

$$x_T(t) = x_0 + \sum_{i=1}^{[t*T]} \Delta^{(T)}_{x_i} + (t*T - [t*T]) \Delta^{(T)}_{x_{[t*T]}}, \tag{61}$$

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

$$\Delta^{(T)}_{x_i} := x_{k+1} - x_k = f(x_k, t_k) \Delta^{(T)}_{x_k} + g(t_k) \sqrt{\Delta^{(T)}_{x_k} Z^{(T)}_{k}}, \tag{62}$$

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 $\{X_T\}, x$ be processes (with associated probability measures $\{P_T\}, \mathcal{P}$) such that $X_T$ converges to $x$ in finite dimensional distributions (f.d.d.), i.e., for any $k$ time steps $t_1, t_2, \ldots, t_k$,

$$(X_T(t_1), X_T(t_2), \ldots, X_T(t_k)) \overset{P}{\to} (x(t_1), x(t_2), \ldots, x(t_k)). \tag{63}$$

If $\{P_T\}$ are also tight, then $X_T \Rightarrow_T x$.

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 $\{P_{x_k}\}_{k=1}^n$ corresponding to the structured random walk $\{x_k\}_{k=1}^n$ is tight.

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

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:

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

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

$$s \in \left[ \frac{k-1}{n}, \frac{k}{n} \right] \text{ and } t \in \left[ \frac{\ell-1}{n}, \frac{\ell}{n} \right]. \tag{65}$$
First, observe that, applying Definition 1, Assumption 2, and the fact that $z_k \in \mathcal{L}^4 \implies E[||z_k||^4] \leq M$ for some $M \in \mathbb{R}$ and all $k \in \{0, \ldots, n\}$,

$$E[||x_n(s - \Delta t) - x_n(s)||_4] = E[||\Delta x_k||_4] = E \left[ ||f(x_k) \Delta t + g \left( \frac{k}{n} \right) z_k \sqrt{\Delta t} ||_4 \right] \leq \Delta t E[||f(x_k)||_4 + \sqrt{\Delta t} \left| g \left( \frac{k}{n} \right) \right| E[||z_k||_4]] \leq \Delta t E[K (1 + ||x_k||_4)] + \sqrt{\Delta t K M} \left( 1 + \frac{k}{n} \right) \leq C \sqrt{\Delta t}, \quad (66)$$

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

We will bound $E \left[ ||x_n(s) - x_n(t)||^4 \right]$ in three regimes:

**Case 1: $k = \ell$**

$$E \left[ ||x_n(s) - x_n(t)||_4 \right] = E \left[ \left| \left| x_0 + \sum_{i=1}^{k-1} \Delta x_i + (ns - k) \Delta x_k - x_0 - \sum_{i=1}^{\ell-1} \Delta x_i - (nt - \ell) \Delta \ell \right| \right|_4 \right] = E \left[ ||(ns - k) \Delta x_k - (nt - k) \Delta x_k ||_4 \right] \leq (n|t - s|) E \left[ ||\Delta x_k||_4 \right] \leq C_1 \sqrt{n}|t - s|, \quad (67)$$

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

$$E \left[ ||x_n(s) - x_n(t)||^4 \right] \leq C_1 |t - s|^2. \quad (68)$$

**Case 2: $k = \ell + 1$**

$$E \left[ ||x_n(s) - x_n(t)||_4 \right] = E \left[ \left| \left| x_0 + \sum_{i=1}^{k-1} \Delta x_i + (ns - k) \Delta x_k - x_0 - \sum_{i=1}^{\ell-1} \Delta x_i - (nt - \ell) \Delta \ell \right| \right|_4 \right] = E \left[ ||\Delta_\ell + (ns - k) \Delta x_k - (nt - \ell) \Delta \ell ||_4 \right] = E \left[ ||(ns - k) \Delta x_k - (nt - k) \Delta \ell ||_4 \right] \leq ||(ns - nt)|| E \left[ ||\Delta x_k||_4 \right] \leq C_1 \sqrt{n}|t - s|,$$

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,

$$E \left[ ||x_n(s) - x_n(t)||^4 \right] = 4C_1 |t - s|^2. \quad (69)$$
Case 3: $k > \ell + 2$

\[
\mathbb{E}[||x_n(s) - x_n(t)||_4^4] 
\leq \mathbb{E} \left[ ||x_n(s) - x_n \left( \frac{k-1}{n} \right) ||_4^4 + ||x_n \left( \frac{k-1}{n} \right) - x_n \left( \frac{\ell-1}{n} \right) ||_4^4 \right] 
+ C_1 \sqrt{s - \frac{k-1}{n}} + \mathbb{E} \left[ \left| \left| x_n \left( \frac{k-1}{n} \right) - x_n \left( \frac{\ell-1}{n} \right) \right| \right|_4^4 \right] 
+ C_1 \sqrt{t - \frac{\ell-1}{n}} 
\leq C_1 \left( \sqrt{s - \frac{k-1}{n}} - \sqrt{t - \frac{\ell-1}{n}} \right) + \mathbb{E} \left[ \left| \left| x_n \left( \frac{k-1}{n} \right) - x_n \left( \frac{\ell-1}{n} \right) \right| \right|_4^4 \right].
\]

Inspecting (*), we can see that

\[
\mathbb{E} \left[ \left| \left| x_n \left( \frac{k-1}{n} \right) - x_n \left( \frac{\ell-1}{n} \right) \right| \right|_4^4 \right] = \mathbb{E} \left[ \left| \left| x_0 + \sum_{i=1}^{k-1} \Delta x_i - x_0 - \sum_{i=1}^{\ell-1} \Delta x_i \right| \right|_4^4 \right] 
= \mathbb{E} \left[ \left| \left| \sum_{i=1}^{k-1} \Delta x_i \right| \right|_4^4 \right] 
\leq \sum_{i=1}^{k-1} \mathbb{E}[||f(x_i)\Delta t||] + \mathbb{E} \left[ \left| \left| \sum_{i=\ell-1}^{k-1} g \left( \frac{i}{n} \right) z_i \sqrt{\Delta t} \right| \right|_4^4 \right] 
\leq C_1 \sqrt{\frac{k-1}{n} - \frac{\ell}{n}} + \mathbb{E} \left[ \left| \left| \sum_{i=\ell-1}^{k-1} g \left( \frac{i}{n} \right) z_i \sqrt{\Delta t} \right| \right|_4^4 \right].
\]

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

**Lemma B.3.**

\[
\mathbb{E} \left[ \left| \left| \sum_{i=\ell-1}^{k-1} g \left( \frac{i}{n} \right) z_i \sqrt{\Delta t} \right| \right|_4^4 \right] \leq C_2 \left( \frac{k-1}{n} - \frac{\ell}{n} \right)^2
\]

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

\[
\mathbb{E} \left[ \left| \left| \sum_{i=\ell-1}^{k-1} g \left( \frac{i}{n} \right) z_i \sqrt{\Delta t} \right| \right|_4^4 \right] = \mathbb{E} \left[ \left| \left| \sum_{i=\ell-1}^{k-1} W_i \right| \right|_4^4 \right],
\]

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

\[
\mathbb{E} \left[ \left| \left| \sum_{i=\ell-1}^{k-2} W_i + W_{k-1} \right| \right|_4^4 \right] = \mathbb{E} \left[ \left| \left| \sum_{i=\ell-1}^{k-2} W_i \right| \right|_4^4 + \left| \left| W_{k-1} \right| \right|_4^4 + \left| \left| \sum_{i=\ell-1}^{k-2} W_i \right| \right|_2^2 \left| \left| W_{k-1} \right| \right|_2^2 \right],
\]
where the odd terms containing first moments of \( W_i \) go to zero. Leveraging Assumption 2 and the fact that \( z_k \in L^4 \) we can further simplify left hand side to

\[
\mathbb{E} \left[ \left\| \sum_{i=\ell-1}^{k-2} W_i + W_{k-1} \right\|^4 \right] \leq \mathbb{E} \left[ \left\| \sum_{i=\ell-1}^{k-2} W_i \right\|^4 \right] + M^4 \left( g^4 \left( \frac{k-1}{n} \right) (\Delta t)^2 \right)
\]

\[
\quad + \left( \sum_{i=\ell-1}^{k-2} M^2 g^2 \left( \frac{i}{n} \right) \Delta t \right) \left( g^2 \left( \frac{k-1}{n} \right) \Delta t \right)
\]

\[
\quad \leq \mathbb{E} \left[ \left\| \sum_{i=\ell-1}^{k-2} W_i \right\|^4 \right] + C_2 (k-\ell) (\Delta t)^2.
\]

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

\[
\mathbb{E} \left[ \left\| \sum_{i=\ell-1}^{k-1} W_i \right\|^4 \right] \leq C_2 (k-\ell)^2 (\Delta t)^2
\]

\[
= C_2 \left( \frac{k}{n} - \frac{\ell}{n} \right)^2.
\]

Assembling the parts, we obtain

\[
\mathbb{E}[||x_n(s) - x_n(t)||^4] \leq C_1 \left( \left( s - \frac{k-1}{n} \right)^2 \left( t - \frac{\ell-1}{n} \right)^2 + \left( \frac{k}{n} - \frac{\ell}{n} \right)^2 \right) + C_2 \left( \frac{k}{n} - \frac{\ell}{n} \right)^2
\]

\[
\leq C_3 |t - s|^2.
\]

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

\[
\mathbb{E}[||x_n(s) - x_n(t)||^4] \leq \max(4C_1, C_3)|t - s|^2.
\]

**Proof.** (of Lemma B.2)

Let us define the auxiliary processes

\[
\bar{x}_T(t) = x_0 + \sum_{i=1}^{[t/T]} \Delta^{(T)}_{x_i} + (t * T - [t * T]) \Delta^{(T)}_{x_{[t/T]}}
\]

\[
\bar{x}_{T,S}(t) = x_0 + \sum_{i=1}^{[t/T]} \Delta^{(T,S)}_{x_i} + (t * T - [t * T]) \Delta^{(T,S)}_{x_{[t/T]}}.
\]

where

\[
\Delta^{(T)}_{x_k} = f(x_k, t_k) \Delta t_k + g(t_k) \mathcal{W}(\Delta t_k)
\]

\[
\Delta^{(T,S)}_{x_k} = f(x_k, t_k) \Delta^{(T)}_{t_k} + \sum_{i=0}^{S-1} g(t_k) \sqrt{\Delta^{(T)}_{t_k}} z_k^{(T,S)}
\]

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

\[
\mathcal{A}(\{t_i\}_{i=1}^k) := (\mathcal{A}(t_1), \ldots, \mathcal{A}(t_k))
\]

(78)
for a diffusion process $A(t)$ evaluated at times $(t_1, t_2, \ldots, t_k)$, we may obtain the desired result by observing that
\[
\lim_{T \to \infty} \text{CDF}[x_T^{k}]
\]
\[
= \lim_{T \to \infty} \lim_{S \to \infty} \text{CDF}[x_T^{k} + (x_{T,S}^{k} - x_T^{k})]
\]
\[
= \lim_{T \to \infty} \lim_{S \to \infty} \text{CDF}[x_{T,S}^{k}]
\]
\[
= \text{CDF}[x^{k}].
\]

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

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

Finally, the result $\bar{x}_T \overset{\text{f.d.d.}}{\longrightarrow} x$ 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
\[
\bar{x}(t) = \sum_{k=1}^{n} x_k \mathbb{1}_{t \in [t_k, t_{k+1})}(t)
\]
be the continuous-time càdlàg extensions of the random walk $x_k$. Now, $\bar{x}_T$ can also be written as the Itô integral
\[
\bar{x}_T = x(0) + \int_0^t f(x(s))ds + \int_0^t g \left( \frac{s + T}{n} \right) dW_s.
\]
Of course, the solution $x$ to Eq. 1 can also be expressed in the similar form
\[
x(t) = x(0) + \int_0^t f(x(s))ds + \int_0^t g(s)dW_s.
\]
Now we may state the following lemma.

Lemma B.5. Let $\bar{x}_T$ be defined as above and Assumption 1 hold. Then $\bar{x}_T$ converges to $x$ 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
\[
\bar{x}_{k+1}^{(T)} = \bar{x}_k^{(T)} + f(\bar{x}_k^{(T)}, t_k)\Delta t_k + g(t_k)\sqrt{\Delta t_k} W(\Delta t_k),
\]
and Eq. (75) of the random walk
\[
x_{k+1}^{(T,S)} = x_k^{(T,S)} + f(x_k^{(T,S)}, t_k)\Delta t_k + \sum_{i=1}^{S} g(t_k)\sqrt{\Delta t_k} Z_{i+T_k}.\]

Applying the Central Limit Theorem, we may see that
\[
A_k := \sum_{i=1}^{S} g(t_k)\sqrt{\Delta t_k} Z_{i+T_k} \overset{D}{\longrightarrow} g(t)\sqrt{\Delta t_k} W(\Delta t_k)
\]
for each $0 \leq k < T$. We now show our result by recursion. In the base case we have that $x_0^{(T,S)} := x_0^{(T)} := x_0$, so clearly $x_0^{(T,S)} \overset{D}{\longrightarrow} \bar{x}_0^{(T)}$. For any subsequent $k + 1 > 0$, we may invoke Slutsky’s Theorem on the independent sequences $x_k^{(T,S)} \overset{D}{\longrightarrow} \bar{x}_k^{(T)}$ and $A_k \overset{D}{\longrightarrow} g(t)\sqrt{\Delta t_k} W(\Delta t_k)$ to obtain
\[
x_{k+1}^{(T,S)} := x_k^{(T,S)} + A_k \overset{D}{\longrightarrow} \bar{x}_k^{(T)} + g(t)\sqrt{\Delta t_k} W(\Delta t_k) =: \bar{x}_{k+1}^{(T)}.
\]
Therefore, we have that \( x_{k+1}^{(T,S)} \xrightarrow{D} x_k^{(T)} \) 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

\[
\epsilon(t) = \sup_{0 \leq s \leq t} \mathbb{E} \left[ \left| \| Y_n(s) - x(s) \| \right|^2 \right].
\]  

(90)

Recalling the definitions \( \mathbf{x}(t) := x_{[nt]} \) and \( \mathbf{z}(t) = z_{[nt]} \), we have

\[
\epsilon(t) = \sup_{0 \leq s \leq t} \mathbb{E} \left[ \left| \int_0^s [f(\mathbf{x}(u)) - f(\mathbf{z}(u))] du + \int_0^s \left[ g \left( \frac{n \cdot u}{n} \right) - g(u) \right] dW_u \right|^2 \right]
\]

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

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

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

We now leverage Assumption 1 to obtain

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

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

\[
\epsilon(t) \leq C \left( \int_0^s \epsilon(u) du + \Delta t \right),
\]

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

\[
\sup_{0 \leq s \leq T} \mathbb{E} \left[ \| Y_n(s) - x(s) \|^2 \right] = \epsilon(T) \leq C \Delta t.
\]  

(91)

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

\[
\mathbb{E}[\| (Y_n(t_1), \ldots, Y_n(t_k)) - (Y(t_1), \ldots, Y(t_k)) \|^2] \leq \sum_{i=1}^k \mathbb{E}[\| Y_n(t_i) - Y(t_i) \|^2]
\]

\[
\leq k \sup_{0 \leq s \leq T} \mathbb{E} \left[ \| Y_n(s) - x(s) \|^2 \right]
\]

\[
\leq kC \Delta t \to 0
\]

as \( \Delta t \to 0 \). This shows \( (Y_n(t_1), \ldots, Y_n(t_k)) \xrightarrow{L_2} (Y(t_1), \ldots, Y(t_k)) \), which implies the desired result.

\[ \square \]

**C. Implementation**

We use directly with no changes the models and training protocols in (Kingma et al., 2021) to parameterize our score network \( \epsilon(x, t) \) 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.