How Discrete and Continuous Diffusion Meet: Comprehensive Analysis of Discrete Diffusion Models via a Stochastic Integral Framework

Yinuo Ren
ICME
Stanford University
yinuoren@stanford.edu

Haoxuan Chen
ICME
Stanford University
haoxuanc@stanford.edu

Grant M. Rotskoff
Department of Chemistry and ICME
Stanford University
rotskoff@stanford.edu

Lexing Ying
Department of Mathematics and ICME
Stanford University
lexing@stanford.edu

Abstract

Discrete diffusion models have gained increasing attention for their ability to model complex distributions with tractable sampling and inference. However, the error analysis for discrete diffusion models remains less well-understood. In this work, we propose a comprehensive framework for the error analysis of discrete diffusion models based on Lévy-type stochastic integrals. By generalizing the Poisson random measure to that with a time-independent and state-dependent intensity, we rigorously establish a stochastic integral formulation of discrete diffusion models and provide the corresponding change of measure theorems that are intriguingly analogous to Itô integrals and Girsanov's theorem for their continuous counterparts. Our framework unifies and strengthens the current theoretical results on discrete diffusion models and obtains the first error bound for the τ-leaping scheme in KL divergence. With error sources clearly identified, our analysis gives new insight into the mathematical properties of discrete diffusion models and offers guidance for the design of efficient and accurate algorithms for real-world discrete diffusion model applications.

1 Introduction

Diffusion and flow-based models designed for discrete distributions have gained significant attention in recent years due to their versatility and wide applicability across various domains. These models have been proposed and refined in several key works [1, 2, 3, 4, 5, 6, 7, 8, 9]. Starting from molecule, protein, and DNA sequence design [10, 11, 12, 13, 14, 15], discrete diffusion models have also proven effective in many other applications, including text, image, sound, motion [16, 17, 18, 19], and have synergized with other methodologies, *e.g.* tensor networks [20]. These developments highlight the growing importance of discrete modeling in advancing both theoretical understanding and efficient implementations. We refer to Appendix A for a more detailed review of related works.

Partly due to the absence of a discrete equivalent to Girsanov's theorem, the error analysis for discrete diffusion models remains underdeveloped compared to their continuous counterparts. [21] conducts a Markov chain-based error analysis for τ-leaping in total variation distance, with further advancements for the particular state space $\mathbb{X} = \{0, 1\}^d$ by [22]. In this work, our goal is to establish a comprehensive framework for discrete diffusion models through a stochastic analysis perspective,

M3L: Mathematics of Modern Machine Learning @ NeurIPS 2024

completely different from previous works. Drawing on tools from Lévy processes and methodologies for analyzing chemical reaction simulations [23, 24], we extend Poisson random measures to those with evolving intensities, *i.e.* both time-inhomogeneous and state-dependent intensities [25], introduce Lévy-type stochastic integrals [26], and articulate corresponding change of measure theorems, which are analogous to the Itô integrals and Girsanov's theorem in continuous settings.

We further demonstrate that discrete diffusion models, both the τ-leaping and uniformization schemes, can be formulated as stochastic integrals w.r.t. Poisson random measures with evolving intensity, which allows for a unified error analysis framework. This new framework, marking a first for discrete diffusion models, is especially convenient and straightforward for decomposing inference error, drawing satisfying parallels with state-of-the-art theories for continuous diffusion model theories [27, 28]. Our approach thus provides intuitive explanations for the loss design and unifies the error analysis across both schemes. Notably, we achieve stronger convergence results in KL divergence, relaxing some of the stringent assumptions previously required, thereby paving the way for the analysis of a broader class of discrete diffusion models of interest and providing valuable insights and tools for designing efficient and accurate algorithms tailored to the practical demands of discrete diffusion models in real-world applications.

1.1 Contributions

Our main contributions are summarized as follows:

- We develop a rigorous framework for discrete diffusion models using Lévy-type stochastic integrals based on the Poisson random measure with evolving intensity, including formulating discrete diffusion models into stochastic integrals and establishing change of measure theorems that facilitate explicit log-likelihood ratio calculations;
- Our framework extends to a comprehensive, continuous-time analysis for error decomposition in discrete diffusion models, drawing clear parallels with the methodologies used in continuous models and enabling more effective adaptations of techniques across different model types;
- We unify and fortify existing research on discrete diffusion models by deriving the first error bound for τ-leaping in terms of KL divergence, provide a comparative study of τ-leaping and uniformization implementations, and shed light on establishing convergence guarantees for a broader spectrum of discrete diffusion models.

2 Preliminaries

In this section, we introduce the basic concepts of discrete diffusion models. A brief review of continuous diffusion models and their error decomposition and analysis is provided in Appendix B for comparison.

In discrete diffusion models, instead of an Itô process, one considers a continuous-time Markov chain $(\boldsymbol{x}_t)_{0\le t\le T}$ in a space $\mathbb{X}$ of finite cardinality as the *forward process*. We denote the probability distribution of $\boldsymbol{x}_t$ by a vector $\boldsymbol{p}_t \in \Delta^{|\mathbb{X}|}$, where $\Delta^{|\mathbb{X}|}$ denotes the probability simplex in $\mathbb{R}^{|\mathbb{X}|}$. Given the target distribution $\boldsymbol{p}_0$, the Markov chain satisfies the following master equation:

$$\frac{\mathrm{d}\boldsymbol{p}_t}{\mathrm{d}t} = \boldsymbol{Q}_t \boldsymbol{p}_t, \quad \text{where} \quad \boldsymbol{Q}_t = (Q_t(y,x))_{x,y\in\mathbb{X}} \in \mathbb{R}^{|\mathbb{X}|\times|\mathbb{X}|} \tag{2.1}$$

is the rate matrix at time t. The rate matrix $\boldsymbol{Q}_t$ satisfies the following two conditions: (i) $Q_t(x,x) = -\sum_{y\neq x} Q_t(y,x),\ \forall x \in \mathbb{X}$; (ii) $Q_t(x,y) \ge 0,\ \forall x \neq y \in \mathbb{X}$.

In the following, we will use a shorthand notation $\widetilde{\boldsymbol{Q}}_t$ to denote the matrix $\boldsymbol{Q}_t$ with the diagonal elements set to zero. It can be shown that the corresponding backward process is of the same form but with a different rate matrix [29]:

$$\frac{\mathrm{d}\overleftarrow{\boldsymbol{p}}_s}{\mathrm{d}s} = \overline{\boldsymbol{Q}}_s \overleftarrow{\boldsymbol{p}}_s, \quad \text{where} \quad \overline{Q}_s(y,x) = \begin{cases} \frac{\overleftarrow{p}_s(y)}{\overleftarrow{p}_s(x)} \overleftarrow{Q}_s(x,y), & \forall x \neq y \in \mathbb{X}, \\ -\sum_{y'\neq x} \overline{Q}_s(y',x), & \forall x = y \in \mathbb{X}. \end{cases} \tag{2.2}$$

The rate matrix $\boldsymbol{Q}_t$ is often chosen to possess certain sparse structures such that the forward process converges to a simple distribution that is easy to sample from. Several popular choices include the uniform and absorbing transitions [30].

The common practice is to define the score function (or rather the score vector) as $\boldsymbol{s}_t(x) = (s_t(x,y))_{y\in\mathbb{X}} := \frac{\boldsymbol{p}_t}{p_t(x)}$, $\forall x \in \mathbb{X}$, and estimate it by a neural network $\widehat{\boldsymbol{s}}_t^\theta(x)$, where the neural network θ is trained by minimizing the score entropy [31, 30]:

$$\theta = \arg\min_\theta \int_0^T \psi_t \mathbb{E}_{x_t\sim p_t}\left[\sum_{y\neq x}\left(-\log\frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} - 1 + \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)}\right) s_t(x,y)Q_t(x,y)\right]\mathrm{d}t. \tag{2.3}$$

Similar to the continuous case, the backward process is approximated by the continuous-time Markov chain with the following master equation with $\boldsymbol{q}_0 = \boldsymbol{p}_\infty$ and rate matrix $\widehat{\overleftarrow{\boldsymbol{Q}}}_s^\theta$:

$$\frac{\mathrm{d}\boldsymbol{q}_s}{\mathrm{d}s} = \widehat{\overleftarrow{\boldsymbol{Q}}}_s^\theta \boldsymbol{q}_s, \text{ where } \widehat{\overleftarrow{Q}}_s^\theta(y,x) = \overleftarrow{s}_s^\theta(x,y)\overleftarrow{Q}_s(x,y),\ \forall x\neq y\in\mathbb{X}. \tag{2.4}$$

and sampling is accomplished by first sampling from the distribution $\boldsymbol{p}_\infty$ and then evolving the Markov chain accordingly.

3 Stochastic Integral Formulation of Discrete Diffusion Models

In this section, we introduce the stochastic integral formulation of discrete diffusion models. The goal is to establish a path evolution equation analogous to Itô integral (or equivalently, SDEs), with the master equation (2.1) and (2.2) analogous to the Fokker-Planck equation, in the continuous case.

3.1 Poisson Random Measure with Evolving Intensity

In the following, the Poisson distribution with expectation λ is denoted by $\mathcal{P}(\lambda)$.

Definition 3.1 (Poisson Random Measure with Evolving Intensity)**.** *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and $(\mathbb{X},\mathcal{B},\nu)$ be a measure space and $\lambda_t(y)$ is a non-negative predictable process on $\mathbb{R}^+\times\mathbb{X}\times\Omega$ satisfying for any $T>0$, $\int_0^T\int_{\mathbb{X}} 1\vee|y|\vee|y|^2\lambda_t(y)\nu(\mathrm{d}y)\mathrm{d}t<\infty$, a.s.. The random measure $N[\lambda](\mathrm{d}t,\mathrm{d}y)$ on $\mathbb{R}^+\times\mathbb{X}$ is called a* Poisson random measure with evolving intensity $\lambda_t(y)$ *if*

(i) For any $B\in\mathcal{B}$ and $0\le s<t$, $N[\lambda]((s,t]\times B)\sim\mathcal{P}\left(\int_s^t\int_B\lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau\right)$;

(ii) For any $t\ge 0$ and disjoint sets $\{B_i\}_{i\in[n]}\subset\mathcal{B}$, $\{N_t[\lambda](B_i):=N[\lambda]((0,t]\times B_i)\}_{i\in[n]}$ are independent stochastic processes.

The well-definedness of this definition is non-trivial, with further details provided in Appendix C.1. The Poisson random measure defined admits Lévy-type stochastic integral (*cf.* Figure 1), Itô isometry, Itô's formula (Theorem C.9), and Lévy's characterization theorem (Theorem C.8), for which we refer readers to Appendix C.2 for details.

Now we turn to the setting of discrete diffusion models, where the state space $\mathbb{X}$ is finite endowed with the natural σ-algebra $\mathcal{B}=2^{\mathbb{X}}$ and the counting measure $\nu=\sum_{y\in\mathbb{X}}\delta_y$.

Proposition 3.2 (Stochastic Integral Formulation of Discrete Diffusion Models)**.** *The forward process in discrete diffusion models (2.1) can be represented by the following stochastic integral:*

$$x_t = x_0 + \int_0^t\int_{\mathbb{X}}(y-x_{t^-})N[\lambda](\mathrm{d}t,\mathrm{d}y),\ \textit{with}\ \lambda_t(y)=\widetilde{Q}_t(y,x_{t^-}), \tag{3.1}$$

and the backward process in discrete diffusion models (2.2) can be represented by the following stochastic integral:

$$\overleftarrow{x}_s = \overleftarrow{x}_0 + \int_0^s\int_{\mathbb{X}}(y-\overleftarrow{x}_{s^-})N[\mu](\mathrm{d}s,\mathrm{d}y),\ \textit{with}\ \mu_s(y)=\overleftarrow{s}_s(\overleftarrow{x}_{s^-},y)\widetilde{Q}_s(\overleftarrow{x}_{s^-},y), \tag{3.2}$$

where X_{t^-} denotes the left limit of a càdlàg process X_t at time t.

The proof of Proposition 3.2 is provided in Appendix C.4. We would like to remark that the stochastic integral formulation in Proposition 3.2 is tantalizingly close to the Itô integral in continuous diffusion models in the form of SDEs (*cf.* (B.1) and (B.2)). Recalling that in the continuous case, Girsanov's theorem is applied for deriving the score-matching loss (B.4) and the error analysis, one may wonder if similar techniques can be applied to discrete diffusion models. The following section provides an affirmative answer to it.

3.2 Change of Measure

The following theorem provides a change-of-measure argument for stochastic integrals w.r.t. Poisson random measures with evolving intensity, analogous to Girsanov's theorem for Itô integrals w.r.t. Brownian motions.

Theorem 3.3 (Change of Measure for Poisson Random Measure with Evolving Density)**.** *Let $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ be a Poisson random measure with evolving intensity $\lambda_t(y)$ in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $h_t(y)$ be a positive predictable process on $\mathbb{R}^+ \times \mathbb{X} \times \Omega$. Suppose the following exponential process is a local $\mathcal{F}_t$-martingale:*

$$Z_t[h] := \exp\left(\int_0^t \int_{\mathbb{X}} \log h_t(y) N[\lambda](\mathrm{d}t \times \mathrm{d}y) - \int_0^t \int_{\mathbb{X}} (h_t(y) - 1)\lambda_t(y)\nu(\mathrm{d}y)\right), \tag{3.3}$$

and $\mathbb{Q}$ is another probability measure on $(\Omega, \mathcal{F})$ such that $\mathbb{Q} \ll \mathbb{P}$ with Radon-Nikodym derivative $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}|_{\mathcal{F}_t} = Z_t[h]$. Then the Poisson random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ under the measure $\mathbb{Q}$ is a Poisson random measure with evolving intensity $\lambda_t(y)h_t(y)$.

Then it is straightforward to derive Corollary C.11, which was derived in [31] with a different technique with Feller processes and adopted in [30, 22] in the design of loss functions. Proofs will be provided in Appendix C.3. One should recall that in the continuous case with Itô integrals, the proximity of two paths in KL divergence only requires a small difference between the drift terms by Girsanov's theorem, and therefore, the score function can be trained with the mean squared error loss (B.4) [32], while in the discrete case, it requires the likelihood ratio to be close to one, accounting for a more complicated score entropy design in the loss (2.3) [31, 30].

4 Error Analysis of Discrete Diffusion Models

In this section, we first introduce two different implementations of the discrete diffusion models, namely τ-leaping [33] and uniformization [34], derive their stochastic integral formulations as in Proposition 3.2, and provide our main results for their error analysis.

4.1 Algorithms

A straightforward algorithm for simulating the backward process is to discretize the integral in (3.2) with an Euler-Maruyama scheme; this leads to the *τ-leaping algorithm* summarized in Algorithm 1. As shown in the following proposition, τ-leaping can be formulated as a stochastic integral.

Proposition 4.1 (Stochastic Integral Formulation of τ-Leaping)**.** *The τ-leaping algorithm (Algorithm 1) is equivalent to solving the following stochastic integral equation:*

$$\widehat{y}_s = \widehat{y}_0 + \int_0^s \int_{\mathbb{X}} (y - \widehat{y}_{\lfloor s \rfloor^-}) N[\widehat{\mu}^\theta_{\lfloor \cdot \rfloor}](\mathrm{d}s, \mathrm{d}y), \tag{4.1}$$

where the evolving intensity $\widehat{\mu}^\theta_s(y)$ is given by $\widehat{\mu}^\theta_{\lfloor s \rfloor}(y) = \overleftarrow{s}^\theta_{\lfloor s \rfloor}(\widehat{y}_{\lfloor s \rfloor^-}, y)\widetilde{Q}_{\lfloor s \rfloor}(\widehat{y}_{\lfloor s \rfloor^-}, y) = \widehat{\mu}^\theta_{s_n}(y)$, in which we used the symbol $\lfloor s \rfloor = s_n$ for $s \in [s_n, s_{n+1})$. We will call the process $\widehat{y}_s$ the interpolating process *of the τ-leaping algorithm and denote the distribution of $\widehat{y}_s$ by $\widehat{q}_s$.*

Another algorithm considered for simulating the backward process in discrete diffusion models is *uniformization*. The algorithm is summarized in Algorithm 2, Appendix C.4. The uniformization algorithm also admits a stochastic integral formulation, as shown in the following proposition.

Proposition 4.2 (Stochastic Integral Formulation of Uniformization)**.** *Under the block discretization scheme $(s_b)_{b\in[0,B]}$ with $s_0 = 0$ and $s_B = T - \delta$, and for any $s \in (s_b, s_{b+1}]$, we define $\overline{\lambda}_s = \sup_{s\in(s_b,s_{b+1}]} \int_{\mathbb{X}} \widehat{\mu}^\theta_s(y)\nu(\mathrm{d}y)$. Then the uniformization algorithm (Algorithm 2) is equivalent to solving the following stochastic integral equation in the augmented measure space $(\mathbb{X} \times [0, \overline{\lambda}], \mathcal{B} \otimes \mathcal{B}([0, \overline{\lambda}]), \nu \otimes m)$:*

$$y_s = y_0 + \int_0^s \int_{\mathbb{X}} \int_{\mathbb{R}} (y - y_{s^-}) \mathbf{1}_{0 \le \xi \le \int_{\mathbb{X}} \widehat{\mu}^\theta_s(y)\nu(\mathrm{d}y)} N[\widehat{\mu}^\theta](\mathrm{d}s, \mathrm{d}y, \mathrm{d}\xi), \tag{4.2}$$

where the evolving intensity $\widehat{\mu}^\theta_s(y)$ is given by $\widehat{\mu}^\theta_s(y) = \overleftarrow{s}^\theta_s(\widehat{y}_{s^-}, y)\widetilde{Q}_s(\widehat{y}_{s^-}, y)$.

Based on Proposition 4.2, one can show that the uniformization algorithm simulates the backward process in discrete diffusion models accurately (*cf.* Theorem C.12), and the proofs of the claims above will be provided in Appendix C.4.

4.2 Assumptions

For simplicity, we assume the rate matrix $\boldsymbol{Q}_t$ is time-homogeneous and symmetric, *i.e.* $\boldsymbol{Q}_t = \boldsymbol{Q}$ for any $t \geq 0$.

Assumption 4.3 (Regularity of the Rate Matrix)**.** *The rate matrix $\boldsymbol{Q}$ satisfies the following conditions:*

(i) For any $x, y \in \mathbb{X}$, $Q(x,y) \leq C$ and $\underline{D} \leq -Q(x,x) \leq \overline{D}$ for some constants $C, \underline{D}, \overline{D} > 0$;

(ii) The modified log-Sobolev constant $\rho(\boldsymbol{Q})$ *of the rate matrix $\boldsymbol{Q}$ (*cf. *Definition D.5) is lower bounded by $\rho > 0$.*

Assumption 4.4 (Bounded Score)**.** *The true score function satisfies $s_t(x,y) \lesssim 1 \vee t^{-1}$, while the learned score function satisfies $\widehat{s}_s^\theta(x,y) \in (0, M]$, for any $x, y \in \mathbb{X}$.*

Assumption 4.5 (Continuity of Score Function)**.** *For any $t > 0$ and $y \in \mathbb{X}$ such that $Q(x_{t^-}, y) > 0$, we have $\left|\frac{\mu_{t+}(y)}{\mu_t(y)}\right| := \left|\frac{p_t(x_{t^-})Q(x_t,y)}{p_t(x_t)Q(x_{t^-},y)} - 1\right| \lesssim 1 \vee t^{-\gamma}$, for some exponent $\gamma \in [0,1]$.*

Assumption 4.6 (ϵ-accurate Score Estimation)**.** *The score function $s_t(x_t)$ is estimated by the neural network $\widehat{s}_t^\theta(x_t)$ with ϵ-accuracy,* i.e.

$$\sum_{n=0}^{N-1}(s_{n+1}-s_n)\mathbb{E}\left[\int_{\mathbb{X}} K\left(\frac{\overleftarrow{\widehat{s}}^\theta_{s_n}(\overleftarrow{x}_{s_n^-},y)}{\overleftarrow{s}_{s_n}(\overleftarrow{x}_{s_n^-},y)}\right)\overleftarrow{s}_{s_n}(\overleftarrow{x}_{s_n^-},y)\widetilde{Q}(\overleftarrow{x}_{s_n^-},y)\nu(\mathrm{d}y)\right] \leq \epsilon.$$

We refer to Appendix E.1 for discussions on these assumptions.

4.3 Error Analysis

4.3.1 τ-Leaping

Theorem 4.7 (Error Analysis of τ-Leaping)**.** *Suppose the time discretization scheme $(s_i)_{i\in[0,N]}$ with $s_0 = 0$ and $s_N = T - \delta$ satisfies for $k \in [0 : N-1]$, $s_{k+1} - s_k \leq \kappa\left(1 \vee (T - s_{k+1})^{1+\gamma-\eta}\right)$, where the exponent η satisfies $\gamma < \eta \lesssim 1 - T^{-1}$ when $\gamma < 1$, and $\eta = 1$ when $\gamma = 1$. Under Assumptions E.1, E.2, E.3, and E.4, we have the following error bound*

$$D_{\mathrm{KL}}(p_\delta \| \widehat{q}_{T-\delta}) \lesssim \exp(-\rho T)\log|\mathbb{X}| + \epsilon + \overline{D}^2\kappa T,$$

and under the following choice of the order of parameters:

$$T = \mathcal{O}\left(\frac{\log\left(\epsilon^{-1}\log|\mathbb{X}|\right)}{\rho}\right),\ \kappa = \mathcal{O}\left(\frac{\epsilon\rho}{\overline{D}^2\log(\epsilon^{-1}\log|\mathbb{X}|)}\right),\ \delta = \begin{cases} 0, & \gamma < 1, \\ \Omega(e^{-\sqrt{T}}), & \gamma = 1, \end{cases} \tag{4.3}$$

where the mixing time t_{mix} is defined in Definition D.11, we have $D_{\mathrm{KL}}(p_\delta \| \widehat{q}_{T-\delta}) \lesssim \epsilon$ with $N = \kappa^{-1}T = \mathcal{O}\left(\frac{\overline{D}^2\rho^2\log^2\left(\epsilon^{-1}\log|\mathbb{X}|\right)}{\epsilon}\right)$ total steps.

The derivation (as provided in Appendix E.4) and conclusions are analogous to the error bound for continuous diffusion models (*cf.* Theorem B.1). We would like to point out the main differences between the continuous and discrete diffusion models:

- **Truncation Error**: While the Ornstein-Uhlenbeck process converges exponentially fast in the continuous diffusion models, the exponential convergence of the forward process in discrete diffusion models is non-trivial for general graphs $\mathcal{G}(\boldsymbol{Q})$. In practice, Assumption E.1 should be verified for the specific problem at hand;
- **Discretization Error**: In continuous diffusion models, the analysis of the discretization error is based on the Itô integral and Girsanov's theorem, while in the discrete case, the Poisson random measure with evolving intensity (*cf.* Definition 3.1) and change of measure (*cf.* Theorem 3.3) that we developed above are employed instead. Early stopping schemes are discussed in Remark E.11.

In Theorem 4.7, the coefficient $\overline{D}$ roughly translates to the dimension d when the discrete diffusion model is applied to $\mathbb{X} = [S]^d$, where S is the number of states along each dimension. Plugging $\overline{D} = \log|\mathbb{X}| = \mathcal{O}(d)$ into the results, we obtain that the total number of steps $N = \widetilde{\mathcal{O}}(d^2)$. This recovers the dependency described in [21, Theorem 1] for τ-leaping with a completely different set of techniques, and importantly, our results do not rely on strong assumptions such as a uniform bound on the true score. We also reduce assumption stringency by relating our assumption on the estimation error (Assumption E.4) more closely to the training loss rather than requiring an L^∞-accurate score estimation error. Most notably, we provide the first convergence guarantees for τ-leaping in KL divergence, strengthened from total variation distance, for discrete diffusion models.

4.3.2 Uniformization

The error analysis of the uniformization algorithm requires the following modified assumption on the accuracy of the learned score function $\widehat{s}_t^\theta(x_t)$:

Assumption 4.6' (ϵ-accurate Learned Score)**.** *The score function $s_t(x_t)$ is estimated by the neural network $\widehat{s}_t^\theta(x_t)$ with ϵ-accuracy,* i.e.

$$\mathbb{E}\left[\int_0^{T-\delta}\int_{\mathbb{X}} K\left(\frac{\overleftarrow{s}_s^\theta(\overleftarrow{x}_{s^-},y)}{\overleftarrow{s}_s(\overleftarrow{x}_{s^-},y)}\right)\overleftarrow{s}_s(\overleftarrow{x}_{s^-},y)\widetilde{Q}(\overleftarrow{x}_{s^-},y)\nu(\mathrm{d}y)\right]\mathrm{d}s \le \epsilon.$$

Theorem 4.8 (Error Analysis of Uniformization)**.** *Suppose the block discretization scheme $(s_b)_{b\in[0,N]}$ with $s_0 = 0$ and $s_N = T-\delta$ satisfies for $k \in [0:N-1]$, $s_{k+1} - s_k \le \kappa\left(1 \vee (T - s_{k+1})\right)$. Under Assumptions E.1, E.2, E.3, and 4.6', we have the following error bound*

$$D_{\mathrm{KL}}(p_\delta \| q_{T-\delta}) \lesssim \exp(-\rho T)\log|\mathbb{X}| + \epsilon,$$

where the mixing time t_{mix} is defined in Definition D.11. Then with $T = \mathcal{O}\left(\frac{\log\left(\epsilon^{-1}\log|\mathbb{X}|\right)}{\rho}\right)$ and $\delta = \Omega(e^{-T})$, we have $D_{\mathrm{KL}}(p_\delta\|q_{T-\delta}) \lesssim \epsilon$ with $\mathbb{E}[N] = \mathcal{O}\left(\frac{\overline{D}\log\left(\epsilon^{-1}\log|\mathbb{X}|\right)}{\rho}\right)$ steps.

The proof of Theorem 4.8 is deferred to Appendix E.4. Following a similar argument for Theorem 4.7, the dimensionality dependency of the uniformization scheme is $\widetilde{\mathcal{O}}(d)$, confirming the result for the special case $\mathbb{X} = \{0,1\}^d$ in [22]. Theorem 4.7 and 4.8 offer a direct comparison of the efficiency of the τ-leaping and uniformization implementations for discrete diffusion models. Our proof reveals that the less favorable quadratic dependency in the τ-leaping scheme arises from the truncation error, which is not present in the uniformization scheme, illustrating a possible advantage of the latter in reducing computational complexity.

Recalling the current result for continuous diffusion models (Theorem B.1) is $\widetilde{\mathcal{O}}(d)$, we conjecture that $\widetilde{\mathcal{O}}(d)$ is also the optimal rate in the discrete case. In the continuous case, the linear dependency is shown to be achievable with Euler-Maruyama schemes via an intricate stochastic localization argument [28]. The corresponding argument for the τ-leaping scheme of discrete diffusion models would be a possible refinement on Proposition E.6, which we believe is of independent interest and will be explored in future work.

5 Conclusion

In this paper, we have developed a comprehensive framework for the error analysis of discrete diffusion models. We rigorously introduced the Poisson random measure with evolving intensity and established the Lévy-type stochastic integral alongside change of measure arguments. These advancements not only hold mathematical significance but also facilitate a clear-cut analysis of discrete diffusion models. Moreover, we demonstrated that the inference process can be formulated as a stochastic integral using the Poisson random measure with evolving intensity, allowing the error to be systematically decomposed and optimized by algorithmic design, mirroring the theoretical framework for continuous diffusion models.

Our framework unifies the error analysis of discrete diffusion models and provides the first error bounds for the τ-leaping scheme in KL divergence. Our results lay a theoretical groundwork for the analysis of discrete diffusion models, adaptable to broader contexts, such as time-inhomogeneous

and non-symmetric rate matrices. We also hope our work will inspire further research on the practical aspects of discrete diffusion models and their applications in various fields.

Acknowledgments

Lexing Ying acknowledges the support of the National Science Foundation under Award No. DMS-2208163.

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[3] Griffin Floto, Thorsteinn Jonsson, Mihai Nica, Scott Sanner, and Eric Zhengyu Zhu. Diffusion on the probability simplex. *arXiv preprint arXiv:2309.02530*, 2023.

[4] Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. *arXiv preprint arXiv:2110.02037*, 2021.

[5] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

[6] Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35:34532–34545, 2022.

[7] Pierre H Richemond, Sander Dieleman, and Arnaud Doucet. Categorical sdes with simplex diffusion. *arXiv preprint arXiv:2210.14784*, 2022.

[8] Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. *arXiv preprint arXiv:2211.16750*, 2022.

[9] Javier E Santos, Zachary R Fox, Nicholas Lubbers, and Yen Ting Lin. Blackout diffusion: generative diffusion models in discrete-state spaces. In *International Conference on Machine Learning*, pages 9034–9059. PMLR, 2023.

[10] Ari Seff, Wenda Zhou, Farhan Damani, Abigail Doyle, and Ryan P Adams. Discrete object generation with reversible inductive construction. *Advances in neural information processing systems*, 32, 2019.

[11] Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X Lu, Nicolo Fusi, Ava P Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pages 2023–09, 2023.

[12] Patrick Emami, Aidan Perreault, Jeffrey Law, David Biagioni, and Peter St John. Plug & play directed evolution of proteins with gradient-based discrete mcmc. *Machine Learning: Science and Technology*, 4(2):025014, 2023.

[13] John J Yang, Jason Yim, Regina Barzilay, and Tommi Jaakkola. Fast non-autoregressive inverse folding with discrete diffusion. *arXiv preprint arXiv:2312.02447*, 2023.

[14] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

[15] Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

[16] Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*, 2024.

[17] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. *arXiv preprint arXiv:2206.07771*, 2022.

[18] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

[19] Seunggeun Chi, Hyung-gun Chi, Hengbo Ma, Nakul Agarwal, Faizan Siddiqui, Karthik Ramani, and Kwonjoon Lee. M2d2m: Multi-motion generation from text with discrete diffusion models. *arXiv preprint arXiv:2407.14502*, 2024.

[20] Luke Causer, Grant M Rotskoff, and Juan P Garrahan. Discrete generative diffusion models without stochastic differential equations: a tensor network approach. *arXiv preprint arXiv:2407.11133*, 2024.

[21] Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

[22] Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *arXiv preprint arXiv:2402.08095*, 2024.

[23] Tiejun Li. Analysis of explicit tau-leaping schemes for simulating chemically reacting systems. *Multiscale Modeling & Simulation*, 6(2):417–436, 2007.

[24] David F Anderson, Arnab Ganguly, and Thomas G Kurtz. Error analysis of tau-leap simulation methods. 2011.

[25] Philip Protter. Point process differentials with evolving intensities. In *Nonlinear stochastic problems*, pages 467–472. Springer, 1983.

[26] David Applebaum. *Lévy processes and stochastic calculus*. Cambridge university press, 2009.

[27] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[28] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

[29] Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[30] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024.

[31] Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. From denoising diffusions to denoising markov models. *arXiv preprint arXiv:2211.03595*, 2022.

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[33] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of chemical physics*, 115(4):1716–1733, 2001.

[34] Nico M Van Dijk. Uniformization for nonhomogeneous markov chains. *Operations research letters*, 12(5):283–291, 1992.

[35] Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, pages 1276–1301. PMLR, 2023.

[36] Nathan C Frey, Daniel Berenberg, Karina Zadorozhny, Joseph Kleinhenz, Julien Lafrance-Vanasse, Isidro Hotzel, Yan Wu, Stephen Ra, Richard Bonneau, Kyunghyun Cho, et al. Protein discovery with discrete walk-jump sampling. *arXiv preprint arXiv:2306.12360*, 2023.

[37] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[38] Thomas J Kerby and Kevin R Moon. Training-free guidance for discrete diffusion models for molecular generation. *arXiv preprint arXiv:2409.07359*, 2024.

[39] Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yuguang Wang. Graph denoising diffusion for inverse protein folding. *Advances in Neural Information Processing Systems*, 36, 2024.

[40] Yang Li, Jinpei Guo, Runzhong Wang, and Junchi Yan. From distribution learning in training to gradient search in testing for combinatorial optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] Ilia Igashov, Arne Schneuing, Marwin Segler, Michael Bronstein, and Bruno Correia. Retrobridge: Modeling retrosynthesis with markov bridges. *arXiv preprint arXiv:2308.16212*, 2023.

[42] Jose Lezama, Tim Salimans, Lu Jiang, Huiwen Chang, Jonathan Ho, and Irfan Essa. Discrete predictor-corrector diffusion models for image synthesis. In *The Eleventh International Conference on Learning Representations*, 2022.

[43] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

[44] Do Huu Dat, Do Duc Anh, Anh Tuan Luu, and Wray Buntine. Discrete diffusion language model for long text summarization. *arXiv preprint arXiv:2407.10998*, 2024.

[45] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4484. PMLR, 2020.

[46] Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382*, 2020.

[47] Yiming Qin, Clement Vignac, and Pascal Frossard. Sparse training of discrete diffusion models for graph generation. *arXiv preprint arXiv:2311.02142*, 2023.

[48] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

[49] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10167–10176, 2023.

[50] Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7226–7236, 2023.

[51] Yunhong Lou, Linchao Zhu, Yaxiong Wang, Xiaohan Wang, and Yi Yang. Diversemotion: Towards diverse human motion generation via discrete diffusion. *arXiv preprint arXiv:2309.01372*, 2023.

[52] Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and Ponnuthurai N Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11502–11511, 2022.

[53] Zhichao Wu, Qiulin Li, Sixing Liu, and Qun Yang. Dctts: Discrete diffusion model with contrastive learning for text-to-speech generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11336–11340. IEEE, 2024.

[54] Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022.

[55] Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Jian Jiao, Juntao Li, Jian Guo, Nan Duan, Weizhu Chen, et al. Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36:39957–39974, 2023.

[56] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. *arXiv preprint arXiv:2310.05793*, 2023.

[57] Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023.

[58] Kun Zhou, Yifan Li, Wayne Xin Zhao, and Ji-Rong Wen. Diffusion-nat: Self-prompting discrete diffusion for non-autoregressive text generation. *arXiv preprint arXiv:2305.04044*, 2023.

[59] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*, 2024.

[60] Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36, 2024.

[61] Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. *arXiv preprint arXiv:2406.01572*, 2024.

[62] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024.

[63] Severi Rissanen, Markus Heinonen, and Arno Solin. Improving discrete diffusion models via structured preferential generation. *arXiv preprint arXiv:2405.17889*, 2024.

[64] Oscar Davis, Samuel Kessler, Mircea Petrache, Avishek Joey Bose, et al. Fisher flow matching for generative modeling over discrete data. *arXiv preprint arXiv:2405.14664*, 2024.

[65] Linfeng Zhang, Weinan E, and Lei Wang. Monge-ampère flow for generative modeling. *arXiv preprint arXiv:1809.10188*, 2018.

[66] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[67] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[68] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021.

[69] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[70] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[71] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[72] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[73] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[74] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[75] Stanley H Chan. Tutorial on diffusion models for imaging and vision. *arXiv preprint arXiv:2403.18103*, 2024.

[76] Lingxiao Wang, Gert Aarts, and Kai Zhou. Generative diffusion models for lattice field theory. *arXiv preprint arXiv:2311.03578*, 2023.

[77] Amira Alakhdar, Barnabas Poczos, and Newell Washburn. Diffusion models in de novo drug design. *Journal of Chemical Information and Modeling*, 2024.

[78] Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024.

[79] Jiahao Fan, Ziyao Li, Eric Alcaide, Guolin Ke, Huaqing Huang, and E Weinan. Accurate conformation sampling via protein structural diffusion. *bioRxiv*, pages 2024–05, 2024.

[80] Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2):136–154, 2024.

[81] Eric A Riesel, Tsach Mackey, Hamed Nilforoshan, Minkai Xu, Catherine K Badding, Alison B Altman, Jure Leskovec, and Danna E Freedman. Crystal structure determination from powder diffraction patterns with generative machine learning. *Journal of the American Chemical Society*, 2024.

[82] Yuchen Zhu, Tianrong Chen, Evangelos A Theodorou, Xie Chen, and Molei Tao. Quantum state generation with structure-preserving diffusion model. *arXiv preprint arXiv:2404.06336*, 2024.

[83] Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR, 2019.

[84] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

[85] Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

[86] Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023.

[87] Sokhna Diarra Mbacke and Omar Rivasplata. A note on the convergence of denoising diffusion probabilistic models. *arXiv preprint arXiv:2312.05989*, 2023.

[88] Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*, 2024.

[89] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

[90] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

[91] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

[92] Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*, 2023.

[93] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

[94] Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*, 2024.

[95] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024.

[96] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36, 2024.

[97] Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

[98] Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *arXiv preprint arXiv:2404.09730*, 2024.

[99] Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024.

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

[101] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[102] Jean Jacod and Albert Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013.

[103] David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955.

[104] Günter Last and Mathew Penrose. *Lectures on the Poisson process*, volume 7. Cambridge University Press, 2017.

[105] Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

[106] Andreas Eberle. Stochastic analysis. *http://www.mi.unikoeln.de/stochana/ws1617/Eberle/StochasticAnalysis2015.pdf*, 2015.

[107] Andreas Eberle. Markov processes. *Lecture Notes at University of Bonn*, 2009.

[108] Steven P Lalley. Continuous time markov chains. *Lecture Notes, University of Chicago*, page 34, 2012.

[109] Sergey G Bobkov and Prasad Tetali. Modified logarithmic sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19:289–336, 2006.

[110] Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4): 1061–1083, 1975.

[111] Persi Diaconis and Laurent Saloff-Coste. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

[112] Daniel W Stroock. Logarithmic sobolev inequalities for gibbs states. In *Dirichlet forms*. 1993.

[113] Laurent Saloff-Coste. Lectures on finite markov chains. lectures on probability theory and statistics (saint-flour, 1996), 301–413. *Lecture Notes in Math*, 1665, 1997.

[114] Sergey G Bobkov and Michel Ledoux. On modified logarithmic sobolev inequalities for bernoulli and poisson measures. *Journal of functional analysis*, 156(2):347–365, 1998.

[115] Tzong-Yow Lee and Horng-Tzer Yau. Logarithmic sobolev inequality for some models of random walks. *The Annals of Probability*, 26(4):1855–1873, 1998.

[116] Sharad Goel. Modified logarithmic sobolev inequalities for some models of random walk. *Stochastic processes and their applications*, 114(1):51–79, 2004.

[117] Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 2006.

A Related Works

Discrete Diffusion Models. The appeal of such models stems from their potential to address challenging problems in fields like computational biology, where they have shown promise in tasks such as molecule, protein, and DNA sequence design [10, 11, 35, 12, 36, 37, 13, 14, 15, 38, 39]. Additionally, these approaches have proven effective in combinatorial optimization [40], modeling retrosynthesis [41], image synthesis [42, 43], text summarization [44] along with the generation of graph [45, 46, 47, 48], layout [49, 50], motion [19, 51], sound [18], image [52, 17], speech [53] and text [54, 55, 56, 57, 58, 16, 59]. Discrete diffusion models also synergize with other methodologies, including tensor networks [20], enhanced guidance mechanisms [60, 61, 62], structured preferential generation [63], and alternative metrics, *e.g.* the Fisher information metric [64]. These developments highlight the growing importance of discrete modeling in advancing both theoretical understanding and efficient implementations.

Continuous Diffusion Models. Continuous diffusion models have been one of the most active research areas in generative modeling. Earlier work on continuous diffusion models and probability flow-based models include [1, 65, 66, 67, 32, 68, 69, 70, 71, 72]. It has shown state-of-the-art performance in various fields of science and engineering. For some recent work and comprehensive review articles, one may refer to [73, 74, 75, 76, 77, 78, 79, 80, 81, 82].

Theory of Continuous Diffusion Models. In addition to the huge success achieved by diffusion models in empirical studies, many works have also tried to establish sampling guarantees for diffusion and probability flow-based models, such as [83, 84, 85, 86, 87, 88]. Regarding the theoretical analysis of continuous diffusion models, [89] provided the first sampling guarantee under the smoothness and isoperimetry assumptions. Follow-up work removed such assumptions [90, 27, 91] and obtained better convergence results [28, 92, 93, 94, 95]. For the probability flow-based implementation, sampling guarantee was also established and further refined in many recent work [96, 97, 98, 99]

B Review of Continuous Diffusion Models

In this section, we provide a brief review of continuous diffusion models and their error analysis.

B.1 Continuous Diffusion Models

In diffusion models, the *forward process* is designed as an Itô process $(\boldsymbol{x}_t)_{0\le t\le T}$ in $\mathbb{R}^d$ satisfying the following stochastic differential equation (SDE):

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{b}_t(\boldsymbol{x}_t)\mathrm{d}t + \boldsymbol{g}_t\mathrm{d}\boldsymbol{w}_t, \text{ with } \quad \boldsymbol{x}_0 \sim p_0, \tag{B.1}$$

where $(\boldsymbol{w}_t)_{t\ge 0}$ is a standard Brownian motion. The probability distribution of $\boldsymbol{x}_t$ is denoted by p_t, and the distribution p_0 at time $t = 0$ is the target distribution for sampling. The time-reversal $(\overleftarrow{\boldsymbol{x}}_s)_{0\le s\le T}$ of (B.1) satisfies the *backward process*:

$$\mathrm{d}\overleftarrow{\boldsymbol{x}}_s = \left[-\overleftarrow{\boldsymbol{b}}_s(\overleftarrow{\boldsymbol{x}}_s) + \overleftarrow{\boldsymbol{g}}_s\overleftarrow{\boldsymbol{g}}_s^\top \nabla \log \overleftarrow{p}_s(\overleftarrow{\boldsymbol{x}}_s)\right]\mathrm{d}s + \overleftarrow{\boldsymbol{g}}_s\mathrm{d}\boldsymbol{w}_s, \tag{B.2}$$

where $\overleftarrow{*}_s$ denotes $*_{T-s}$, with $\overleftarrow{p}_0 = p_T$ and $\overleftarrow{p}_T = p_0$.

One of the common choices for the drift $\boldsymbol{b}_t$ and the diffusion coefficient $\boldsymbol{g}$ is $\boldsymbol{b}_t(\boldsymbol{x}) = -\frac{1}{2}\beta_t\boldsymbol{x}_t$ and $\boldsymbol{g} = \sigma\sqrt{\beta_t}\boldsymbol{I}$, under which (B.1) is an Ornstein-Uhlenbeck (OU) process converging exponentially, *i.e.* $p_T \approx p_\infty := \mathcal{N}(0, \sigma^2\boldsymbol{I})$, and the forward process (B.1) and the backward process (B.2) reduce to the following form:

$$\mathrm{d}\boldsymbol{x}_t = -\frac{1}{2}\beta_t\boldsymbol{x}_t\mathrm{d}t + \sigma\sqrt{\beta_t}\mathrm{d}\boldsymbol{w}_t, \text{ and } \mathrm{d}\overleftarrow{\boldsymbol{x}}_s = \overleftarrow{\beta}_s\left[\frac{1}{2}\overleftarrow{\boldsymbol{x}}_s + \sigma^2\nabla\log\overleftarrow{p}_s(\overleftarrow{\boldsymbol{x}}_s)\right]\mathrm{d}s + \sigma\sqrt{\overleftarrow{\beta}_s}\mathrm{d}\boldsymbol{w}_s. \tag{B.3}$$

In pratice, the score function $\boldsymbol{s}_t(\boldsymbol{x}_t) := \nabla\log p_t(\boldsymbol{x}_t)$ is often estimated by a neural network $\widehat{\boldsymbol{s}}_t^\theta(\boldsymbol{x}_t)$, where θ denotes the parameters, and trained via *denoising score-matching* [100]:

$$\theta = \arg\min_\theta \int_0^T \psi_t\mathbb{E}_{\boldsymbol{x}_t\sim p_t}\left[\left\|\nabla\log p_t(\boldsymbol{x}_t) - \widehat{\boldsymbol{s}}_t^\theta(\boldsymbol{x}_t)\right\|^2\right]\mathrm{d}t \tag{B.4}$$

where $p_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)$ is the transition distribution from $\boldsymbol{x}_0$ to $\boldsymbol{x}_t$ under (B.3) with an explicit form as

$$\mathcal{N}(\boldsymbol{\mu}_t, \sigma_t^2 \boldsymbol{I}), \text{ where } \boldsymbol{\mu}_t = \boldsymbol{x}_0 e^{-\frac{1}{2}\int_0^t \beta_t \mathrm{d}t} \text{ and } \sigma_t^2 = \sigma^2 \left(1 - e^{-\int_0^t \beta_t \mathrm{d}t}\right), \tag{B.5}$$

and ψ_t is a weighting function for the loss at time t. After obtaining the NN-based score function $\widehat{\boldsymbol{s}}_t^\theta(\boldsymbol{x}_s)$, the backward process in (B.3) is approximated as:

$$\mathrm{d}\boldsymbol{y}_s = \left[\frac{1}{2}\boldsymbol{y}_s + \widehat{\boldsymbol{s}}_s^\theta(\boldsymbol{y}_s)\right]\mathrm{d}s + \mathrm{d}\boldsymbol{w}_s, \quad \text{with } \boldsymbol{y}_0 \sim q_0 = \mathcal{N}(0, \sigma^2 \boldsymbol{I}). \tag{B.6}$$

B.2 Error Analysis of Continuous Diffusion Models

For continuous diffusion models, the error analysis is often conducted by considering the following three error terms:

- **Truncation Error:** The error caused by approximating p_T by p_∞, which is often of the order $\mathcal{O}(d\exp(-T))$ due to exponential ergodicity;
- **Approximation Error:** The error caused by approximating the score function $\nabla \log p_t(\boldsymbol{x}_t)$ by a neural network $\widehat{\boldsymbol{s}}_t^\theta(\boldsymbol{x}_t)$, which is often assumed to be of order $\mathcal{O}(\epsilon)$, where ϵ is a small threshold, given a thorough training process;
- **Discretization Error:** The error caused by numerically solving the SDE (B.6) with Euler-Maruyama scheme or other schemes, *e.g.* exponential integrator [101].

Then, the total error is obtained from these three error terms with proper choices of the order of the time horizon T and the design of the numerical scheme. We extract the following theorem from the state-to-the-art theoretical result [28] for later comparison:

Theorem B.1 (Error Analysis of Continuous Diffusion Models)**.** *Suppose the time discretization scheme $(s_i)_{i\in[0,N]}$ with $s_0 = 0$ and $s_N = T - \delta$ satisfies $s_{k+1} - s_k \leq \kappa(T - s_{k+1})$ for $k \in [0 : N-1]$. Assume $\mathrm{cov}(p_0) = \boldsymbol{I}$, and the score function $\nabla \log p_t(\boldsymbol{x}_t)$ is estimated by the neural network $\widehat{\boldsymbol{s}}_t^\theta(\boldsymbol{x}_t)$ with ϵ-accuracy,* i.e.

$$\sum_{k=0}^{N-1}(s_{k+1} - s_k)\mathbb{E}_{\overleftarrow{\boldsymbol{x}}_{s_k}\sim\overleftarrow{p}_{s_k}}\left[\left\|\nabla \log \overleftarrow{p}_{s_k}(\overleftarrow{\boldsymbol{x}}_{s_k}) - \overleftarrow{\widehat{\boldsymbol{s}}}_{s_k}^\theta(\boldsymbol{x}_{s_k})\right\|^2\right] \leq \epsilon.$$

Then, under the following choice of the order of parameters

$$T = \mathcal{O}(\log(d\epsilon^{-1})),\ \kappa = \mathcal{O}(d^{-1}\epsilon^2 \log^{-1}(d\epsilon^{-1})),\ N = \mathcal{O}(d\epsilon^{-1}\log(d\epsilon^{-1})),$$

then we have $D_{\mathrm{KL}}(p_\delta \| \widehat{q}_{t_N}) \leq \epsilon$, where $\widehat{q}_{t_N}$ is the distribution of the approximate backward process (B.6) implemented with exponential integrator after N steps.

C Mathematical Framework of Poisson Random Measure

In this section, we provide a mathematical framework for Poisson random measure with evolving intensity, which is crucial for the error analysis of discrete diffusion models in the main text.

C.1 Preliminaries

We first provide the definition of the ordinary Poisson random measure.

Definition C.1 (Poisson Random Measure)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathbb{X}, \mathcal{B}, \nu)$ be a measure space satisfying that*

$$\int_{\mathbb{X}} 1 \vee |y| \vee |y|^2 \nu(\mathrm{d}y) < \infty,$$

The random measure $N(\mathrm{d}t, \mathrm{d}y)$ on $\mathbb{R}^+ \times \mathbb{X}$ is called a Poisson random measure *w.r.t. measure ν if it is a random counting measure satisfying the following properties:*

(i) For any $B \in \mathcal{B}$ and $0 \leq s < t$, $N((s,t] \times B) \sim \mathcal{P}(\nu(B)(t-s))$;

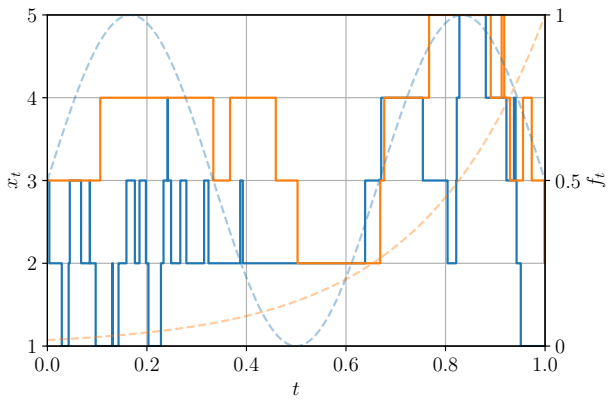

Figure 1: Example trajectories of stochastic integrals (3.1) w.r.t. Poisson random measure with different evolving intensities. The intensity is chosen as $\lambda_t(y) = 50 f_t$ if $|y - x_{t^-}| = 1$ or otherwise 0, as shown in dashed lines. Intuitively, λ_t controls the rate of jumps at time t and location y.

(ii) For any $t \geq 0$ and pairwise disjoint sets $\{B_i\}_{i\in[n]} \subset \mathcal{B}$, $\{N_t(B_i) := N((0,t] \times B_i)\}_{i\in[n]}$ are independent stochastic processes.

The following definition of *predictability* will be frequently used for the well-definedness of stochastic integrals w.r.t. Poisson random measure, and thus the extension from ordinary Poisson random measure to Poisson random measure with evolving intensity.

Definition C.2 (Predictability)**.** *The predictable σ-algebra on $\mathbb{R}^+ \times \mathbb{X}$ is defined as the σ-algebra generated by all sets of the form $(s,t] \times B$ for $0 \leq s < t$ and $B \in \mathcal{B}$. A process X_t is called predictable if and only if X_t is predictable w.r.t. the predictable σ-algebra above.*

In the following, we will define the Poisson random measure with evolving intensity, which is a special case of random measures [102, Definition 1.3].

Definition C.3 (Poisson Random Measure with Evolving Intensity)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathbb{X}, \mathcal{B}, \nu)$ be a measure space. Suppose $\lambda_t(y)$ is a non-negative predictable process on $\mathbb{R}^+ \times \mathbb{X} \times \Omega$ satisfying that for any $0 \leq T < \overline{T}$,*

$$\int_0^T \int_{\mathbb{X}} 1 \vee |y| \vee |y|^2 \lambda_t(y) \nu(\mathrm{d}y)\mathrm{d}t < \infty, \text{ a.s..}$$

The random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ on $\mathbb{R}^+ \times \mathbb{X}$ is called a Poisson random measure *with* evolving intensity *$\lambda_t(y)$ w.r.t. measure ν if it is a random counting measure satisfying the following properties:*

(i) For any $B \in \mathcal{B}$ and $0 \leq s < t$, $N[\lambda]((s,t] \times B) \sim \mathcal{P}\left(\int_s^t \int_B \lambda_\tau(y) \nu(\mathrm{d}y)\mathrm{d}\tau\right)$;

(ii) For any $t \geq 0$ and pairwise disjoint sets $\{B_i\}_{i\in[n]} \subset \mathcal{B}$,

$$\{N_t[\lambda](B_i) := N[\lambda]((0,t] \times B_i)\}_{i\in[n]}$$

are independent stochastic processes.

Theorem C.4 (Well-definedness of Poisson Random Measure with Evolving Intensity)**.** *The Poisson random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ with evolving intensity $\lambda_t(y)$ is well-defined under the conditions in the definition above.*

Proof. We first augment the $(\mathbb{X}, \mathcal{B}, \nu)$ measure space to a product space $(\mathbb{X} \times \mathbb{R}, \mathcal{B} \times \mathcal{B}(\mathbb{R}), \nu \times m)$, where m is the Lebesgue measure on $\mathbb{R}$, and $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$. The Poisson random measure with evolving intensity $\lambda_t(y)$ can be defined in the augmented measure space as

$$N[\lambda]((s,t] \times B) := \int_s^t \int_B \int_{\mathbb{R}} \mathbf{1}_{0 \leq \xi \leq \lambda_\tau(y)} N(\mathrm{d}\tau, \mathrm{d}y, \mathrm{d}\xi), \tag{C.1}$$

where $N(\mathrm{d}\tau, \mathrm{d}y, \mathrm{d}\xi)$ is the Poisson random measure on $\mathbb{R}^+ \times \mathbb{X} \times \mathbb{R}$ w.r.t. measure $\nu(\mathrm{d}y)\mathrm{d}\xi$.

Then it is straightforward to verify the two conditions in the definition of Poisson random measure with evolving intensity by noticing that for pairwise disjoint sets $\{B_i\}_{i\in[n]} \subset \mathcal{B}$, $\{B_i \times \mathbb{R}\}_{i\in[n]} \subset \mathcal{B} \times \mathcal{B}(\mathbb{R})$ are also pairwise disjoint.

The Poisson random process $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ with evolving intensity $\lambda_t(y)$ is well-defined up to an eventual explosion time

$$\overline{T} = \inf_T \left\{ \int_0^T \int_{\mathbb{X}} \lambda_t(y)\nu(\mathrm{d}y)\mathrm{d}t = \infty, \text{ a.s.} \right\}.$$

We refer the readers to [25] for a more rigorous detailed version of the proof. □

Remark C.5 (Relation to the Cox process)**.** *The Poisson random measure with evolving intensity shares multiple similarities with the Cox process [103, 104], including being a point process and with the intensity being a random measure. The main difference is that the Cox process is defined on a general measure space, while the Poisson random measure with evolving intensity is defined on the product space $(\mathbb{X} \times \mathbb{R}, \mathcal{B} \times \mathcal{B}(\mathbb{R}), \nu \times m)$ and the intensity function is required to be predictable to ensure the well-definedness of its stochastic integral.*

C.2 Stochastic Integral w.r.t. Poisson Random Measure

The following theorems provide the properties of stochastic integrals w.r.t. Poisson random measure with evolving intensity. The proofs are based on the observation that with the augmentation of the measure space argument (C.1), the stochastic integral w.r.t. Poisson random measure with evolving intensity in $(\mathbb{X}, \mathcal{B}, \nu)$ can be reduced to the stochastic integral w.r.t. homogeneous Poisson random measure in $(\mathbb{X} \times \mathbb{R}, \mathcal{B} \times \mathcal{B}(\mathbb{R}), \nu \times m)$, and under certain conditions on the measure space $(\mathbb{X}, \mathcal{B}, \nu)$, to the well-known Lévy-type stochastic integral [26]. For simplicity, we will work on the interval $t \in [0, T]$ with $T < \overline{T}$ and the following regularity conditions of the Poisson random measure:

$$0 < \operatorname*{ess\,inf}_{\tau\in[0,T], y\in\mathbb{X}} \lambda_\tau(y) \le \operatorname*{ess\,sup}_{\tau\in[0,T], y\in\mathbb{X}} \lambda_\tau(y) < +\infty.$$

One can easily generalize the following results to their local versions on $[0, \overline{T})$ by considering its compact subsets.

Theorem C.6 (Stochastic Integrals w.r.t. Poisson Random Measure with Evolving Density)**.** *For any predictable process $K_t(y)$ on $\mathbb{R}^+ \times \mathbb{X} \times \Omega$, the stochastic integral w.r.t. Poisson random measure with evolving intensity $\lambda_t(y)$*

$$x_t = x_0 + \int_0^t \int_{\mathbb{X}} K_t(y) N[\lambda](\mathrm{d}t, \mathrm{d}y), \tag{C.2}$$

has a unique solution, for which the following properties hold:

(1) (Expectation) For any $t \ge 0$, we have

$$\mathbb{E}\left[\int_0^t \int_{\mathbb{X}} K_t(y) N[\lambda](\mathrm{d}t, \mathrm{d}y)\right] = \int_0^t \int_{\mathbb{X}} K_t(y)\lambda_t(y)\nu(\mathrm{d}y)\mathrm{d}t;$$

(2) (Martingale) For any $t \ge 0$, we have

$$\int_0^t \int_{\mathbb{X}} K_t(y)\widetilde{N}[\lambda](\mathrm{d}t, \mathrm{d}y) := \int_0^t \int_{\mathbb{X}} K_t(y) N[\lambda](\mathrm{d}t, \mathrm{d}y) - \int_0^t \int_{\mathbb{X}} K_t(y)\lambda_t(y)\nu(\mathrm{d}y)\mathrm{d}t$$

is a local $\mathcal{F}_t$-martingale;

(3) (Itô Isometry) For any $t \ge 0$, we have

$$\mathbb{E}\left[\left(\int_0^t \int_{\mathbb{X}} K_t(y) N[\lambda](\mathrm{d}t, \mathrm{d}y)\right)^2\right] = \int_0^t \int_{\mathbb{X}} K_t(y)^2 \lambda_t(y)\nu(\mathrm{d}y)\mathrm{d}t.$$

Proof. We first write the integral (C.2) in the augmented measure space $(\mathbb{X}\times\mathbb{R}, \mathcal{B}\times\mathcal{B}(\mathbb{R}), \nu\times m)$ as

$$x_t = x_0 + \int_0^t \int_{\mathbb{X}} \int_{\mathbb{R}} K_t(y)\mathbf{1}_{0\le\xi\le\lambda_t(y)} N(\mathrm{d}t, \mathrm{d}y, \mathrm{d}\xi), \tag{C.3}$$

and since $K_t(y)\mathbf{1}_{0\le\xi\le\lambda_t(y)}$ is a predictable process, the desired properties can be derived from the corresponding properties of the stochastic integral w.r.t. Poisson random measure in the augmented measure space.

The subsequent proof will follow a similar argument as the proof of the stochastic integral w.r.t. Brownian motion (*e.g.* in [105]) by starting from proving the properties for elementary processes, which in our case refer to working with the *elementary predictable processes* of the following form:

$$Z_t(y,\xi)(\omega) = \sum_{i=0}^{n-1}\sum_{j=1}^{m}\sum_{k=1}^{l} Z_{i,j,k}(\omega)\mathbf{1}_{t\in(t_i,t_{i+1}]}\mathbf{1}_{y\in B_j}\mathbf{1}_{\xi\in C_k},$$

where $0 = t_0 < \cdots < t_n = T$ is a partition of $[0,T]$, $B_j \in \mathcal{B}$ for $j \in [m]$ are a partition of $\mathbb{X}$ with $\nu(B_k) < \infty$, and $C_k \in \mathcal{B}(\mathbb{R})$ for $k \in [l]$ are a partition of the time interval $[0, \operatorname{ess\,sup}_{\tau\in[0,T], y\in\mathbb{X}} \lambda_\tau(y)]$ with $m(C_k) < \infty$, and $K_{i,j,k}$ is bounded and $\mathcal{F}_{t_i}$-measurable, on which the stochastic integral is defined as

$$\int_0^t \int_{\mathbb{X}} Z_t(y,\xi) N(\mathrm{d}t, \mathrm{d}y, \mathrm{d}\xi) = \sum_{i=0}^{n-1}\sum_{j=1}^{m}\sum_{k=1}^{l} Z_{i,j,k} N_{t_i}((t_i, t_{i+1}]\times B_j \times C_k).$$

Then, it is straightforward to verify the properties of the stochastic integral for the elementary predictable process $Z_t^+(y,\xi)$, using the definition of Poisson random measure (Definition C.1). For general predictable processes $Z_t(y,\xi)$, we write $Z_t(y,\xi) = Z_t^+(y,\xi) - Z_t^-(y,\xi)$, where $Z_t^+(y,\xi)$ and $Z_t^-(y,\xi)$ are positive and negative parts of $Z_t(y,\xi)$, and apply the results to $Z_t^+(y,\xi)$ and $Z_t^-(y,\xi)$ separately.

Finally, we take $Z_t(y,\xi) = K_t(y)\mathbf{1}_{0\le\xi\le\lambda_t(y)}$ to derive the properties of the stochastic integral w.r.t. Poisson random measure with evolving intensity.

We refer readers to [106, Section 2.2] for detailed arguments. For the uniqueness of the solution to the stochastic integral, we also refer to [25, Theorem 3.1]. □

Proposition C.7. *Define the list of jump times $(t_n)_{n\in\mathbb{N}}$ recursively as*

$$t_0 = 0, \quad t_{n+1} = \inf\{t > t_n | \Delta x_t \neq 0\},\ n \ge 0,$$

the Poisson random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ with evolving intensity $\lambda_t(y)$ can be written as

$$N[\lambda](\mathrm{d}t, \mathrm{d}y) = \sum_{n=1}^{\infty} \delta_{t_n}(\mathrm{d}t)\delta_{Y_n}(\mathrm{d}y), \tag{C.4}$$

and the stochastic integral (C.2) is càdlàg and can be rewritten as a sum of jumps:

$$x_t = x_0 + \sum_{n=1}^{N} \Delta x_{t_n} = x_0 + \sum_{n=1}^{N} K_{t_n}(Y_n), \tag{C.5}$$

where N is a random variable satisfying $t_N \le t < t_{N+1}$, and Δx_{t_n} are the jumps $\Delta x_{t_n} = x_{t_n} - x_{t_n^-}$ with $x_{t_n^-} := \lim_{s\to t_n^-} x_s$.

Proof. To see the solution is càdlàg, we notice the following right limit at time t:

$$\lim_{\epsilon\to 0}(x_{t+\epsilon} - x_t) = \int_{(t,t+\epsilon]\times\mathbb{X}} K_t(y) N[\lambda](\mathrm{d}t, \mathrm{d}y) \to 0,$$

and the left limit at time t:

$$\Delta x_t = \lim_{\epsilon\to 0}(x_t - x_{t-\epsilon}) = \int_{(t-\epsilon,t]\times\mathbb{X}} K_t(y) N[\lambda](\mathrm{d}t, \mathrm{d}y) \to \int_{\mathbb{X}} K_t(y) N[\lambda](\{t\}\times \mathrm{d}y), \tag{C.6}$$

where the notation $N[\lambda](\{t\} \times \mathrm{d}y)$ should be understood as $N[\lambda](\{t\} \times \mathrm{d}y) = 0$ if $t \notin \{t_n\}_{n\in\mathbb{N}}$, or otherwise $Y_n = N[\lambda](\{t_n\} \times \mathrm{d}y)$ is a random variable on $\mathbb{X}$.

Since the Poisson random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ with evolving intensity $\lambda_t(y)$ is a random counting measure, it can be represented as a countable sum of Dirac measures as in (C.4), and thus we have

$$\begin{aligned} x_t &= x_0 + \int_0^t \int_{\mathbb{X}} K_t(y) N[\lambda](\mathrm{d}t, \mathrm{d}y) \\ &= x_0 + \int_0^t \int_{\mathbb{X}} K_t(y) \sum_{n=1}^{N} \delta_{t_n}(\mathrm{d}t)\delta_{Y_n}(\mathrm{d}y) = x_0 + \sum_{n=1}^{N} K_{t_n}(Y_n). \end{aligned}$$

By the definition of Poisson random measure, $(t_n)_{n\in[N]}$ are also the jump times of the homogeneous Poisson random measure $N(\mathrm{d}t, \mathrm{d}y, \mathrm{d}\xi)$ in the augmented measure space $(\mathbb{X} \times \mathbb{R}, \mathcal{B} \times \mathcal{B}(\mathbb{R}), \nu \times m)$ w.r.t. measure $\nu(\mathrm{d}y)\mathrm{d}\xi$. Therefore, with a slight abuse of notations, we will assume $(t_n, Y_n, \Xi_n)_{n\in[N]}$ are i.i.d. random variables with probability measure proportional to $\mathrm{d}t\nu(\mathrm{d}y)\mathrm{d}\xi$, for each of which $\Xi_n \le \lambda_{t_n}(Y_n)$ holds because otherwise the jump would not occur.

Then the distribution of Y_n can be derived as a conditional probability of the jump location Y_n given the jump time t_n and $\Xi_n \le \lambda_{t_n}(Y_n)$:

$$\mathbb{P}(Y_n = y)\nu(\mathrm{d}y) = \frac{\int_{\mathbb{R}} \nu(\mathrm{d}y)\mathbf{1}_{\Xi_n \le \lambda_{t_n}(y)}\mathrm{d}\xi}{\int_{\mathbb{R}}\int_{\mathbb{X}} \nu(\mathrm{d}y)\mathbf{1}_{\Xi_n \le \lambda_{t_n}(y)}\mathrm{d}\xi} = \frac{\lambda_{t_n}(y)\nu(\mathrm{d}y)}{\int_{\mathbb{X}} \lambda_{t_n}(y)\nu(\mathrm{d}y)}, \tag{C.7}$$

and the proof is complete. □

The following theorem gives the martingale characterization of Poisson random measure with evolving intensity, which will be crucial for the proof of the change of measure arguments:

Theorem C.8 (Martingale Characterization of Poisson Random Measure with Evolving Density)**.** *Let $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ be a $\mathcal{F}_t$-adapted process in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ is a Poisson random measure with evolving intensity $\lambda_t(y)$ if and only if the complex-valued process*

$$M_t[f] = \exp\left(i\int_0^t \int_{\mathbb{X}} f_\tau(y) N[\lambda](\mathrm{d}\tau, \mathrm{d}y) + \int_0^t \int_{\mathbb{X}} \left(1 - e^{if_\tau(y)}\right) \lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau \right) \tag{C.8}$$

is a local martingale for any predictable process $f_\tau(y)$ satisfying that $f_\tau(y) \in L^1(\mathbb{X}, \nu)$, a.s..

Proof. By Proposition C.7, we rewrite the stochastic integral as a sum of jumps:

$$\int_0^t \int_{\mathbb{X}} f_t(y) N[\lambda](\mathrm{d}\tau, \mathrm{d}y) = \sum_{n=1}^{N} f_{t_n}(Y_n),$$

where $(t_n, Y_n, \Xi_n)_{n\in[N]}$ are i.i.d. random variables with probability measure proportional to $\mathrm{d}t\nu(\mathrm{d}y)\mathrm{d}\xi$, for each of which $\Xi_n \le \lambda_{t_n}(Y_n)$ holds, following a similar argument as in the proof of Proposition C.7.

Then, it is straightforward to derive the following probability of the jump time $t_n = \tau$:

$$\mathbb{P}(t_n = \tau)\mathrm{d}\tau = \frac{\int_{\mathbb{X}} \mathbb{P}(Y_n = y, t_n = \tau)\nu(\mathrm{d}y)}{\int_0^t \int_{\mathbb{X}} \mathbb{P}(Y_n = y, t_n = \tau)\nu(\mathrm{d}y)\mathrm{d}\tau} = \frac{\int_{\mathbb{X}} \lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau}{\int_0^t \int_{\mathbb{X}} \lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau};$$

and by the definition of the Poisson random measure, we have the following probability of the total number of jumps $N = n$:

$$\mathbb{P}(N = n) = \frac{1}{n!} \exp\left(-\int_0^t \int_{\mathbb{X}} \lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau \right) \left(\int_0^t \int_{\mathbb{X}} \lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau \right)^n.$$

Without loss of generality, we only verify $\mathbb{E}[M_t[f]] = 1$ as follows, and general cases are similar by Markov property:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(i\int_0^t\int_{\mathbb{X}} f_{t_n}(y)N[\lambda](\mathrm{d}\tau,\mathrm{d}y)\right)\right]\\
=&\mathbb{E}\left[\exp\left(i\sum_{n=1}^{N} f(Y_n)\right)\Big|\Xi_n \le \lambda_{t_n}(Y_n),\ \forall n\in[N]\right]\\
=&\mathbb{E}\left[\prod_{n=1}^{N}\mathbb{E}\left[e^{if_{t_n}(Y_n)}\big|\Xi_n\le\lambda_{t_n}(Y_n)\right]\right]\\
=&\mathbb{E}\left[\prod_{n=1}^{N}\mathbb{E}\left[\int_{\mathbb{X}}\frac{e^{if_\tau(y)}\lambda_\tau(y)\nu(\mathrm{d}y)}{\int_{\mathbb{X}}\lambda_\tau(y)\nu(\mathrm{d}y)}\Big|t_n=\tau\right]\right]\\
=&\sum_{n=1}^{\infty}\frac{1}{n!}\left(\int_0^t\int_{\mathbb{X}} e^{if_\tau(y)}\lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau\right)^n\exp\left(-\int_0^t\int_{\mathbb{X}}\lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau\right)\\
=&\exp\left(\int_0^t\int_{\mathbb{X}}\left(e^{if_\tau(y)}-1\right)\lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau\right),
\end{aligned}$$

which immediately yields the desired result $\mathbb{E}[M_t[f]] = 1$.

On the other hand, for any $0 \le s < t$ and $B \in \mathcal{B}$, we set

$$Z_t(y) = u\mathbf{1}_{t\in(s,t]}\mathbf{1}_{y\in B},$$

where $u \in \mathbb{R}$, and by assumption, we have

$$\begin{aligned}
&\mathbb{E}\left[M_t[Z]\right] = \mathbb{E}\left[\exp\left(i\int_0^t\int_{\mathbb{X}} Z_\tau(y)N[\lambda](\mathrm{d}\tau,\mathrm{d}y) + \int_0^t\int_{\mathbb{X}}\left(1-e^{iZ_\tau(y)}\right)\lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau\right)\right]\\
=&\mathbb{E}\left[\exp\left(iu\int_s^t\int_B N[\lambda](\mathrm{d}\tau,\mathrm{d}y) + \int_s^t\int_{\mathbb{X}}\left(1-e^{iu}\right)\lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau\right)\right]\\
=&\mathbb{E}\left[\exp\left(iuN[\lambda]((s,t]\times B) + \left(1-e^{iu}\right)(t-s)\int_{\mathbb{X}}\lambda_\tau(y)\nu(\mathrm{d}y)\right)\right] = 1,
\end{aligned}$$

i.e. the following holds for any $u \in \mathbb{R}$:

$$\mathbb{E}\left[\exp\left(iuN[\lambda]((s,t]\times B)\right)\right] = \left(e^{iu}-1\right)(t-s)\int_{\mathbb{X}}\lambda_\tau(y)\nu(\mathrm{d}y),$$

which by Lévy's continuity theorem implies that

$$N[\lambda]((s,t]\times B) \sim \mathcal{P}\left((t-s)\int_{\mathbb{X}}\lambda_\tau(y)\nu(\mathrm{d}y)\right),$$

and thus $N[\lambda](\mathrm{d}t,\mathrm{d}y)$ is a Poisson random measure with evolving intensity $\lambda_t(y)$ by Definition C.1. □

Theorem C.9 (Itô's Formula for Poisson Random Measure with Evolving Density)**.** *Let $N[\lambda](\mathrm{d}t,\mathrm{d}y)$ be a Poisson random measure with evolving intensity $\lambda_t(y)$ in the probability space $(\Omega,\mathcal{F},\mathbb{P})$ and $K_t(y)$ be a predictable process on $\mathbb{R}^+\times\mathbb{X}\times\Omega$. Suppose a process x_t satisfies the stochastic integral*

$$x_t = x_0 + \int_0^t\int_{\mathbb{X}} K_t(y)N[\lambda](\mathrm{d}t,\mathrm{d}y), \tag{C.9}$$

then for any $f_t(y) \in C(\mathbb{R}^+\times\mathbb{X})$ with probability 1, we have

$$f_t(x_t) = f_0(x_0) + \int_0^t \partial_\tau f_\tau(x_\tau)\mathrm{d}\tau + \int_0^t\int_{\mathbb{X}}\left(f_\tau(x_{\tau^-}+K_\tau(y)) - f_\tau(x_{\tau^-})\right)N[\lambda](\mathrm{d}\tau,\mathrm{d}y).$$

Proof. By Proposition C.7, we again rewrite the stochastic integral as a sum of jumps:

$$x_t = x_0 + \sum_{n=1}^{N} K_{t_n}(Y_n),$$

where $(t_n)_{n\in[N]}$ are the jump times with $t_0 = 0$ and $t_N \le t < t_{N+1}$, and $(Y_n)_{n\in[N]}$ are the jump locations. Consequently, it is easy to see that $x_{t_n^-} = x_t = x_{t_{n-1}}$ for $t \in (t_{n-1}, t_n]$ and $n \in [N]$.

Then we have the following decomposition:

$$\begin{aligned}
&f_t(x_t) - f_0(x_0) \\
=&f_t(x_t) - f_{t_N}(x_{t_N}) + \sum_{n=1}^{N} \left(f_{t_n}(x_{t_n}) - f_{t_n}(x_{t_n^-}) + f_{t_n}(x_{t_n^-}) - f_{t_{n-1}}(x_{t_{n-1}}) \right) \\
=&\int_{t_N}^{t} \partial_\tau f_\tau(x_{t_N}) \mathrm{d}\tau + \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \partial_\tau f_\tau(x_{t_{n-1}}) \mathrm{d}\tau + \sum_{n=1}^{N} \left(f_{t_n}(x_{t_{n-1}} + K_{t_n}(Y_n)) - f_{t_n}(x_{t_{n-1}}) \right),
\end{aligned}$$

and for the last term in the above equation, we have

$$\begin{aligned}
&\sum_{n=1}^{N} \left(f_{t_n}(x_{t_{n-1}} + K_{t_n}(Y_n)) - f_{t_n}(x_{t_{n-1}}) \right) = \sum_{n=1}^{N} \left(f_{t_n}(x_{t_n^-} + K_{t_n}(Y_n)) - f_{t_n}(x_{t_n^-}) \right) \\
=&\int_0^t \int_{\mathbb{X}} \left(f_\tau(x_{\tau^-} + K_\tau(y)) - f_\tau(x_{\tau^-}) \right) N[\lambda](\mathrm{d}\tau, \mathrm{d}y).
\end{aligned}$$

Combining the above results, we have the desired result. □

Lemma C.10. *Denote the trajectory obtained by simulating the master equation (2.1) of the forward process of the discrete diffusion model as x_t, then the time interval $\Delta t_n = t_{n+1} - t_n$ is distributed according to the following distribution:*

$$\mathbb{P}(\Delta t_n > \tau) = \exp\left(\int_0^\tau Q_{\tau'}(x_{t_n}, x_{t_n}) \mathrm{d}\tau' \right), \tag{C.10}$$

and the jump location $x_{t_{n+1}}$ is distributed according to the following distribution:

$$\mathbb{P}(x_{t_{n+1}} = y) = -\frac{Q_{t_{n+1}}(y, x_{t_n})}{Q_{t_{n+1}}(x_{t_n}, x_{t_n})}. \tag{C.11}$$

Proof. The results can be found in [107, Section 1.2]. While a fully rigorous proof can be conducted by discretizing the time-inhomogeneous continuous-time Markov chain into 2^n uniform steps and taking the limit as $n \to \infty$, following the approach in [108], we will provide a more intuitive proof here for completeness.

Set $\boldsymbol{p}_{t_n} = \boldsymbol{e}_{x_{t_n}}$, where $\boldsymbol{e}_y$ is the y-th unit vector in $\mathbb{R}^{|\mathbb{X}|}$. then the x_{t_n}-th entry of (2.1) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(x_t = x_{t_n}) = \frac{\mathrm{d}}{\mathrm{d}t} p_t(x_{t_n}) = \sum_{y\in\mathbb{X}} Q_t(x_{t_n}, y) p_t(y),$$

which, by the assumed continuity of the rate matrix $\boldsymbol{Q}_t$, implies

$$\mathbb{P}(\Delta t_n > \tau) = \mathbb{P}(x_{t_n+\tau} = x_{t_n}) = 1 + Q_{t_n}(x_{t_n}, x_{t_n})\tau + o(\tau),$$

and thus

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \log \mathbb{P}(\Delta t_n > \tau) = \lim_{\tau\to 0} \frac{\log \mathbb{P}(\Delta t_n > \tau)}{\tau} = \lim_{\tau\to 0} Q_{t_n}(x_{t_n}, x_{t_n}) + o(1) = Q_{t_n}(x_{t_n}, x_{t_n}),$$

integrating the above equation yields the desired result.

Similarly, by setting $\boldsymbol{p}_{t_{n+1}-\tau} = \boldsymbol{e}_{x_{t_n}}$, we have for all $y \in \mathbb{X} \setminus \{x_{t_n}\}$:

$$\begin{aligned}
&\mathbb{P}(x_{t_{n+1}} = y) = \boldsymbol{p}_{t_{n+1}}(y) \\
=&\lim_{\tau \to 0} \frac{\boldsymbol{p}_{t_{n+1}-\tau}(y) + Q_{t_{n+1}-\tau}(y, x_{t_n})\tau + o(\tau)}{\sum_{y \in \mathbb{X}\setminus\{x_{t_n}\}} \left(\boldsymbol{p}_{t_{n+1}-\tau}(y) + Q_{t_{n+1}-\tau}(y, x_{t_n})\tau + o(\tau)\right)} \\
=&\frac{Q_{t_{n+1}}(y, x_{t_n})}{\sum_{y \in \mathbb{X}\setminus\{x_{t_n}\}} Q_{t_{n+1}}(y, x_{t_n})} = -\frac{Q_{t_{n+1}}(y, x_{t_n})}{Q_{t_{n+1}}(x_{t_n}, x_{t_n})},
\end{aligned}$$

and the result follows. □

C.3 Proofs of Change of Measure Related Arguments

Proof of Theorem 3.3. In the following, we will denote the expectation under the measure $\mathbb{P}$ by $\mathbb{E}_{\mathbb{P}}$ and the expectation under the measure $\mathbb{Q}$ by $\mathbb{E}_{\mathbb{Q}}$.

By Theorem C.8, to verify that the Poisson random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ with evolving intensity $\lambda_t(y)$ is a Poisson random measure with evolving intensity $\lambda_t(y)h_t(y)$ under the measure $\mathbb{Q}$, it suffices to show that for any $f \in L^1(\mathbb{X}, \nu)$, the complex-valued process

$$M_t[f] = \exp\left(i\int_0^t \int_{\mathbb{X}} f(y) N[\lambda](\mathrm{d}\tau, \mathrm{d}y) + \int_0^t \int_{\mathbb{X}} \left(1 - e^{if(y)}\right) \lambda_\tau(y) h_\tau(y) \nu(\mathrm{d}y)\mathrm{d}\tau\right)$$

is a local martingale under the measure $\mathbb{Q}$.

To this end, we perform the following calculation:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\left[M_t[f]\right] = \mathbb{E}_{\mathbb{P}}\left[M_t[f] Z_t[h]\right] \\
=&\mathbb{E}_{\mathbb{P}}\Bigg[\exp\left(i\int_0^t \int_{\mathbb{X}} f(y) N[\lambda](\mathrm{d}\tau, \mathrm{d}y) + \int_0^t \int_{\mathbb{X}} \left(1 - e^{if(y)}\right) \lambda_\tau(y) h_\tau(y) \nu(\mathrm{d}y)\mathrm{d}\tau\right) \\
&\qquad \exp\left(\int_0^t \int_{\mathbb{X}} \log h_t(y) N[\lambda](\mathrm{d}t \times \mathrm{d}y) - \int_0^t \int_{\mathbb{X}} (h_t(y) - 1)\lambda_t(y)\nu(\mathrm{d}y)\right)\Bigg] \\
=&\mathbb{E}_{\mathbb{P}}\left[\exp\left(i\int_0^t \int_{\mathbb{X}} \left(f(y) + \log h_t(y)\right) N[\lambda](\mathrm{d}\tau, \mathrm{d}y) + \int_0^t \int_{\mathbb{X}} \left(1 - e^{if(y)} h_t(y)\right) \lambda_t(y)\nu(\mathrm{d}y)\right)\right] \\
=&\mathbb{E}_{\mathbb{P}}\left[M_t[f + \log h]\right],
\end{aligned}$$

and by assumption, $f + \log h \in L^1(\mathbb{X}, \nu)$, a.s., which implies that $M_t[f + \log h]$ is a local martingale under the measure $\mathbb{P}$ again by Theorem C.8. Consequently, $M_t[f]$ is a local martingale under the measure $\mathbb{Q}$, and the result follows. □

Corollary C.11 (Equivalence between KL Divergence and Score Entropy-based Loss Function)**.** *Let $\overleftarrow{p}_{0:T}$ and $q_{0:T}$ be the path measures of the backward process (2.2) and the approximate backward process (2.4), then it holds that*

$$\begin{aligned}
&D_{\mathrm{KL}}(\overleftarrow{p}_T \| q_T) \le D_{\mathrm{KL}}(\overleftarrow{p}_{0:T} \| q_{0:T}) \\
=&D_{\mathrm{KL}}(\overleftarrow{p}_0 \| q_0) + \mathbb{E}\left[\int_0^T \int_{\mathbb{X}} K\left(\frac{\overleftarrow{s}^\theta_s(\overleftarrow{x}_{s^-}, y)}{\overleftarrow{s}_s(\overleftarrow{x}_{s^-}, y)}\right) \overleftarrow{s}_s(\overleftarrow{x}_{s^-}, y) \widetilde{Q}_s(\overleftarrow{x}_{s^-}, y) \nu(\mathrm{d}y)\mathrm{d}t\right],
\end{aligned} \tag{C.12}$$

where $K(x) = x - 1 - \log x \ge 0$, and the expectation is taken w.r.t. paths generated by the backward process (3.2). Consequently, minimizing the loss (2.3) is equivalent to minimizing the KL divergence between the path measures of the ground truth and the approximate backward process.

Proof of Corollary C.11. With a similar argument as in the proof of Proposition 3.2, we have the following stochastic integral representation of the approximate backward process with the learned neural network score function $\widehat{s}^\theta_t$:

$$y_s = y_0 + \int_0^s \int_{\mathbb{X}} (y - y_{s^-}) N[\widehat{\mu}^\theta](\mathrm{d}s, \mathrm{d}y), \text{ with } \widehat{\mu}^\theta_s(y) = \overleftarrow{\widehat{s}}^\theta_s(y_{s^-}, y) \widetilde{Q}_s(y_{s^-}, y) = \widehat{\mu}^\theta_s(y). \tag{C.13}$$

By the data-processing inequality and the chain rule of KL divergence, we have

$$D_{\mathrm{KL}}(\overleftarrow{p}_T\|q_T) \le D_{\mathrm{KL}}(\overleftarrow{p}_{0:T}\|q_{0:T}) = D_{\mathrm{KL}}(\overleftarrow{p}_0\|q_0) + \mathbb{E}\left[D_{\mathrm{KL}}(\overleftarrow{p}_{0:T}\|q_{0:T}|\overleftarrow{x}_0 = y_0 = y)\right].$$

Then notice that conditioning on the alignment of the initial state $\overleftarrow{x}_0 = y_0 = y$ for any $y \in \mathbb{X}$, the second term in the above equation can be expressed as

$$D_{\mathrm{KL}}(\overleftarrow{p}_{0:T}\|q_{0:T}|\overleftarrow{x}_0 = y_0 = y) = \mathbb{E}\left[\log \frac{\mathrm{d}\overleftarrow{p}_{0:T}}{\mathrm{d}q_{0:T}}\middle|\overleftarrow{x}_0 = y_0 = y\right] = \mathbb{E}\left[\log Z_T^{-1}\left[\frac{\widehat{\mu}^\theta}{\mu}\right]\right],$$

where the last equality is by the change of measure in Theorem 3.3 from the stochastic integral formulation (3.2) of the backward process (2.2) with the true score function $\overleftarrow{s}$ to the stochastic integral formulation (C.13) of the approximate backward process with the learned score function $\widehat{s}^\theta$.

Plug in the expression of Z_T in (3.3) and notice that

$$\frac{\widehat{\mu}_s^\theta}{\mu_s} = \frac{\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y)\widetilde{Q}_s(y_{s^-}, y)}{\overleftarrow{s}_s(y_{s^-}, y)\widetilde{Q}_s(y_{s^-}, y)} = \frac{\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y)}{\overleftarrow{s}_s(y_{s^-}, y)},$$

we have

$$\begin{aligned}
&\mathbb{E}\left[\log Z_T^{-1}\left[\frac{\widehat{\mu}^\theta}{\mu}\right]\right] \\
=&\mathbb{E}\left[-\int_0^T\int_{\mathbb{X}} \log \frac{\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y)}{\overleftarrow{s}_s(y_{s^-}, y)} N[\mu](\mathrm{d}s \times \mathrm{d}y) + \int_0^T\int_{\mathbb{X}} \left(\frac{\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y)}{\overleftarrow{s}_s(y_{s^-}, y)} - 1\right)\mu_s(y)\nu(\mathrm{d}y)\mathrm{d}s\right] \\
=&\mathbb{E}\left[\int_0^T\int_{\mathbb{X}} \left(\frac{\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y)}{\overleftarrow{s}_s(y_{s^-}, y)} - 1 - \log \frac{\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y)}{\overleftarrow{s}_s(y_{s^-}, y)}\right)\overleftarrow{s}_s(y_{s^-}, y)\widetilde{Q}_s(y_{s^-}, y)\nu(\mathrm{d}y)\mathrm{d}s\right] \\
=&\mathbb{E}\left[\int_0^T\int_{\mathbb{X}} \left(\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y) - \overleftarrow{s}_s(y_{s^-}, y) - \overleftarrow{s}_s(y_{s^-}, y)\log \frac{\overleftarrow{\widehat{s}}_s^\theta(y_{s^-}, y)}{\overleftarrow{s}_s(y_{s^-}, y)}\right)\widetilde{Q}_s(y_{s^-}, y)\nu(\mathrm{d}y)\mathrm{d}s\right],
\end{aligned}$$

rearranging the terms in the above equation yields the desired result. □

C.4 Proofs of Stochastic Integral Formulations

Proof of Proposition 3.2. In the following, we will denote the trajectory obtained by simulating the master equation (2.1) of the forward process of the discrete diffusion model as $\widetilde{x}_t$ and the trajectory obtained by the stochastic integral (3.1) as x_t, with $x_0 = \widetilde{x}_0$. We will also use the notation $\widetilde{\cdot}$ to denote the quantities associated with the trajectory $\widetilde{x}_t$. The goal is to show that x_t and $\widetilde{x}_t$ are identically distributed for any $t \in [0, T]$.

We prove this claim by induction. We assume that for any $t \in [0, t_n]$, where $n \in \mathbb{N}$ and t_n is the n-th jump time with $t_0 = 0$, the two trajectories x_t and $\widetilde{x}_t$ are identically distributed. For simplicity, we realign the two processes x_t and $\widetilde{x}_t$ at time t_n by setting $x_{t_n} = \widetilde{x}_{t_n}$.

We first consider the process $\widetilde{x}_t$ generated by the discrete diffusion model (2.1). Recall the definition $\lambda_t(y) = \widetilde{Q}_t(y, \widetilde{x}_{t^-})$, we have that

$$\int_{\mathbb{X}} \lambda_t(y)\nu(\mathrm{d}y) = \sum_{y\in\mathbb{X}} \widetilde{Q}_t(y, \widetilde{x}_{t^-}) = -Q_t(\widetilde{x}_{t^-}, \widetilde{x}_{t^-}) = -Q_t(\widetilde{x}_{t_n}, \widetilde{x}_{t_n}), \text{ for } t \in (t_n, t_{n+1}].$$

By Lemma C.10, the time interval $\Delta\widetilde{t}_n = \widetilde{t}_{n+1} - \widetilde{t}_n$ is distributed according to (C.10), *i.e.*

$$\mathbb{P}(\Delta\widetilde{t}_n > \tau) = \exp\left(\int_0^\tau Q_{\tau'}(\widetilde{x}_{t_n}, \widetilde{x}_{t_n})\mathrm{d}\tau'\right) = \exp\left(-\int_0^\tau\int_{\mathbb{X}} \lambda_{\tau'}(y)\nu(\mathrm{d}y)\mathrm{d}\tau'\right).$$

Similarly, the jump location $\widetilde{x}_{t_{n+1}}$ is distributed according to (C.11), *i.e.*

$$\mathbb{P}(\widetilde{x}_{t_{n+1}} = y) = -\frac{Q_{t_{n+1}}(y, \widetilde{x}_{t_n})}{Q_{t_{n+1}}(\widetilde{x}_{t_n}, \widetilde{x}_{t_n})} = \frac{\lambda_{t_{n+1}}(y)}{\int_{\mathbb{X}} \lambda_{t_{n+1}}(y)\nu(\mathrm{d}y)}.$$

Now we turn to the stochastic integral (3.1). By definition of the Poisson random measure, we have

$$\mathbb{P}(\Delta t_n > \tau) = \mathbb{P}(N[\lambda]((t_n, t_n + \tau] \times \mathbb{X}) = 0) = \exp\left(-\int_{t_n}^{t_n+\tau} \int_{\mathbb{X}} \lambda_{\tau'}(y)\nu(\mathrm{d}y)\mathrm{d}\tau'\right),$$

and the jump location is distributed according to (C.7), *i.e.*

$$\mathbb{P}(x_{t_{n+1}} = y) = \frac{\lambda_{t_{n+1}}(y)}{\int_{\mathbb{X}} \lambda_{t_{n+1}}(y)\nu(\mathrm{d}y)}.$$

Comparing the arguments above, we conclude that the two processes x_t and $\widetilde{x}_t$ are identically distributed for any $t \in [0, t_{n+1}]$, and the induction is complete.

The proof of the equivalence between the backward process of the discrete diffusion model governed by (2.2) and the corresponding stochastic integral (3.2) can be conducted similarly, and the result follows. □

C.4.1 τ-Leaping

The τ-leaping algorithm is summarized in Algorithm 1.

Algorithm 1: τ-Leaping Algorithm for Discrete Diffusion Model Inference

Input: $\widehat{y}_0 \sim q_0$, time discretization scheme $(s_i)_{i\in[0:N]}$ with $s_0 = 0$ and $s_N = T - \delta$, intensity function $\widehat{\mu}_s^\theta$ defined in Proposition 4.1, and neural network-based score function estimation $\widehat{\boldsymbol{s}}_t^\theta$.

Output: A sample $\widehat{y}_{s_N} \sim \widehat{q}_{t_N}$.

for $n = 0$ **to** $N - 1$ **do**

$$\widehat{y}_{s_{n+1}} \leftarrow \sum_{y\in\mathbb{X}} (y - \widehat{y}_{s_n})\mathcal{P}(\widehat{\mu}_s^\theta(\widehat{y}_{s_n})(s_{n+1} - s_n)); \tag{C.14}$$

end

Proof of Proposition 4.1. Without loss of generality, we give the proof for $s = s_N$, and the general case can be proved similarly.

The stochastic integral (4.1) can be partitioned by the time discretization $(s_i)_{i\in[0:N]}$ into N intervals along which the evolving intensity is constant, *i.e.*

$$\begin{aligned}
\widehat{y}_{s_N} =& \widehat{y}_0 + \int_0^s \int_{\mathbb{X}} (y - \widehat{y}_{\lfloor s\rfloor^-}) N[\widehat{\mu}_{\lfloor\cdot\rfloor}](\mathrm{d}s, \mathrm{d}y) \\
=& \widehat{y}_0 + \sum_{i=1}^N \int_{s_{i-1}}^{s_i} \int_{\mathbb{X}} (y - \widehat{y}_{s_{i-1}^-}) N[\widehat{\mu}_{s_{i-1}}](\mathrm{d}s, \mathrm{d}y) \\
=& \widehat{y}_0 + \sum_{i=1}^N \int_{\mathbb{X}} (y - \widehat{y}_{s_{i-1}^-}) N[\widehat{\mu}_{s_{i-1}}]((s_{i-1}, s_i], \mathrm{d}y),
\end{aligned}$$

which given $\mathbb{X}$ is finite, can be further decomposed into the following sum of jumps:

$$\begin{aligned}
\widehat{y}_{s_N} =& \widehat{y}_0 + \sum_{i=1}^N \int_{\mathbb{X}} (y - \widehat{y}_{s_{i-1}^-}) N[\widehat{\mu}_{s_{i-1}}]((s_{i-1}, s_i], \mathrm{d}y) \\
=& \widehat{y}_0 + \sum_{i=1}^N \sum_{y\in\mathbb{X}} (y - \widehat{y}_{s_{i-1}^-}) N[\widehat{\mu}_{s_{i-1}}]((s_{i-1}, s_i], \{y\}) \\
\sim& \widehat{y}_0 + \sum_{i=1}^N \sum_{y\in\mathbb{X}} (y - \widehat{y}_{s_{i-1}^-}) \mathcal{P}((s_i - s_{i-1})\widehat{\mu}_{s_{i-1}}(y)),
\end{aligned}$$

which is exactly (C.14) in the τ-leaping algorithm (Algorithm 1). □

C.4.2 Uniformization

The uniformization algorithm is summarized in Algorithm 2, in which $\sigma_{(m)}$ denotes the m-th order statistic of the M uniform random variables on $[0,1]$, and the randomness in (C.15) should be understood as sampling a categorical distribution and updating the state accordingly. The main idea is to simulate the backward process by a Poisson random measure with a piecewise constant intensity upper bound process and then sample the behavior of each jump according to the intensity $\widehat{\mu}_s^\theta(y)$ at time s.

Algorithm 2: Uniformization Algorithm for Discrete Diffusion Model Inference

Input: $\widehat{y}_0 \sim q_0$, time discretization scheme $(s_b)_{b\in[0,N]}$ with $s_0 = 0$ and $s_B = T - \delta$, intensity upper bound process $\overline{\lambda}_s$, intensity function $\widehat{\mu}_s^\theta$ defined in Proposition 4.1, and neural network-based score function estimation $\widehat{\boldsymbol{s}}_t^\theta$.

Output: A sample $x_{s_B} \sim q_{t_B}$.

for $b = 0$ **to** $B-1$ **do**
  $M \sim \mathcal{P}(\overline{\lambda}_{s_{b+1}}(s_{b+1} - s_b))$, $\sigma_m \sim \mathrm{Unif}([0,1])$ for $m \in [M]$;
  for $m = 1$ **to** M **do**

$$\widehat{y}_{s_b+\sigma_{(m)}} \leftarrow \begin{cases} y, & \text{with prob. } \widehat{\mu}^\theta_{s_b+\sigma_{(m)}}(y)/\overline{\lambda}_{s_{b+1}}, \text{ for } y \in \mathbb{X}, \\ \widehat{y}_{s_b}, & \text{with prob. } 1 - \sum_{y\in\mathbb{X}} \widehat{\mu}^\theta_{s_b+\sigma_{(m)}}(y)/\overline{\lambda}_{s_{b+1}}; \end{cases} \tag{C.15}$$

  end
end

Theorem C.12 (Accurate Simulation by Uniformization)**.** *The uniformization algorithm (Algorithm 2) with its stochastic integral formulation in (4.2) is equivalent to the approximate backward process with its stochastic integral formulation in (C.13).*

Proof of Proposition 4.2 and Theorem C.12. For simplicity, we only consider the stochastic integral (4.2) within the time interval $(s_b, s_{b+1}]$.

We rewrite the stochastic integral (4.2) as a sum of jumps:

$$y_{s_{b+1}} = y_{s_b} + \sum_{i=1}^{N} (Y_n - y_{s_{b,n}^-}) \mathbf{1}_{0\le \Xi_n \le \int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{s_{b,n}}}(y)\nu(\mathrm{d}y)},$$

where by Proposition C.7, $(s_{b,n})_{n\in[N]}$ are the jump times and $(Y_n, \Xi_n)_{n\in[N]}$ are the jump locations that are distributed according to

$$\mathbb{P}(Y_n = y, \Xi_n = \xi) = \frac{\widehat{\mu}^\theta_{s_{b,n}}(y)}{\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{b,n}}(y)\nu(\mathrm{d}y) \int_0^{\overline{\lambda}} \mathrm{d}\xi} = \frac{\widehat{\mu}^\theta_{s_{b,n}}(y)}{\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{b,n}}(y)\nu(\mathrm{d}y)\overline{\lambda}}. \tag{C.16}$$

Therefore, the n-th jump is not performed if $\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{s_{b,n}}}(y)\nu(\mathrm{d}y) < \Xi_n \le \overline{\lambda}$, which is of probability

$$\begin{aligned} \mathbb{P}(\Xi_n > \widehat{\mu}^\theta_{s_{b,n}}(y_{s_{b,n}^-})) &= \frac{\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{b,n}}(y)\nu(\mathrm{d}y)}{\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{b,n}}(y)\nu(\mathrm{d}y)} \frac{\overline{\lambda} - \int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{s_{b,n}}}(y)\nu(\mathrm{d}y)}{\overline{\lambda}} \\ &= 1 - \frac{\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{b,n}}(y)\nu(\mathrm{d}y)}{\overline{\lambda}}, \end{aligned}$$

and is to the state y with probability

$$\begin{aligned} &\mathbb{P}\left(Y_n = y, \Xi_n \le \int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{s_{b,n}}}(y)\nu(\mathrm{d}y)\right) \\ =&\frac{\widehat{\mu}^\theta_{s_{b,n}}(y)}{\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{b,n}}(y)\nu(\mathrm{d}y)} \frac{\int_{\mathbb{X}} \widehat{\mu}^\theta_{s_{b,n}}(y)\nu(\mathrm{d}y)}{\overline{\lambda}} = \frac{\widehat{\mu}^\theta_{s_{b,n}}(y)}{\overline{\lambda}}, \end{aligned}$$

which coincides with (C.15) in the uniformization algorithm (Algorithm 2).

By conditioning on the occurrence of each jump, *i.e.* $0 \le \Xi_n \le \int_{\mathbb{X}} \widehat{\mu}^{\theta}_{s_{s_{b,n}}}(y)\nu(\mathrm{d}y)$, with slight abuse of notation, we have that

$$\mathbb{P}\left(Y_n = y \middle| 0 \le \Xi_n \le \int_{\mathbb{X}} \widehat{\mu}^{\theta}_{s_{s_{b,n}}}(y)\nu(\mathrm{d}y)\right) = \frac{\widehat{\mu}^{\theta}_{s_{b,n}}(y)}{\int_{\mathbb{X}} \widehat{\mu}^{\theta}_{s_{b,n}}(y)\nu(\mathrm{d}y)},$$

which again by Proposition C.7 implies that y_s also satisfies the stochastic integral (C.13) corresponding to the approximate backward process, and vice versa, and the result follows. □

D Results for Continuous-Time Markov Chain

In this section, we will provide some results for the continuous-time Markov chain (CTMC), including the mixing time, the spectral gap, the modified log-Sobolev constant, etc. We will use the notation $\mathcal{G}(\boldsymbol{Q})$ to denote the graph corresponding to the rate matrix $\boldsymbol{Q}$, *i.e.*

$$\mathcal{G}(\boldsymbol{Q}) = (\mathbb{X}, E(\mathcal{G}(\boldsymbol{Q})), Q), \text{ where } E(\mathcal{G}(\boldsymbol{Q})) = \{(x,y) \in \mathbb{X}\times\mathbb{X} | x \neq y, Q(x,y) > 0\},$$

and the weight of the directed edge $(x,y) \in E(\mathcal{G}(\boldsymbol{Q}))$ is $Q(x,y)$. We will assume that the continuous-time Markov chain is irreducible and reversible on the state space $\mathbb{X}$, and the corresponding stationary distribution is $\boldsymbol{\pi}$.

D.1 Spectral Gap

Definition D.1 (Spectral Gap)**.** *Let $\boldsymbol{L} = -\boldsymbol{Q}$ be the* graph Laplacian *matrix with $\boldsymbol{D} = \operatorname{diag}\boldsymbol{L}$, corresponding to the graph $\mathcal{G}(\boldsymbol{Q})$. with*

$$0 = \lambda_1(\boldsymbol{L}) < \lambda_2(\boldsymbol{L}) \le \ldots \le \lambda_{|\mathbb{X}|}(\boldsymbol{L}) \le 2\max_{x\in\mathbb{X}} D(x,x) = 2\overline{D},$$

the spectral gap $\lambda(\boldsymbol{Q})$ of the rate matrix $\boldsymbol{Q}$ is defined as the second smallest eigenvalue of the graph Laplacian $\boldsymbol{L}$, i.e. *$\lambda(\boldsymbol{Q}) = \lambda_2(\boldsymbol{L})$.*

Remark D.2 (Asymptotic Behavior of the score function $\boldsymbol{s}_t$)**.** *Assume $\boldsymbol{Q}$ is symmetric with the following orthogonal eigendecomposition:*

$$\boldsymbol{Q} = -\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top,$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{|\mathbb{X}|})$ is an orthogonal matrix, and the distribution $\boldsymbol{p}_t$ has the following decomposition w.r.t. the eigenvectors of the graph Laplacian $\boldsymbol{L}$:

$$\boldsymbol{p}_t = \sum_{i=1}^{|\mathbb{X}|} \alpha_t(i)\boldsymbol{u}_i = \boldsymbol{U}\boldsymbol{\alpha}(t),$$

then the solution to the master equation (2.1) is given by

$$\boldsymbol{p}_t = \exp(t\boldsymbol{Q})\boldsymbol{p}_0 = \boldsymbol{U}\exp(-t\boldsymbol{\Lambda})\boldsymbol{U}^\top\boldsymbol{p}_0 = \boldsymbol{U}\exp(-t\boldsymbol{\Lambda})\boldsymbol{\alpha}_0 = \sum_{j=1}^{|\mathbb{X}|}\boldsymbol{u}_j\exp(-t\lambda_j)\boldsymbol{\alpha}_0(j),$$

i.e. *$\boldsymbol{\alpha}_t = \exp(-t\boldsymbol{\Lambda})\boldsymbol{\alpha}_0$ and thus for any $i \in [|\mathbb{X}|]$,*

$$p_t(i) - p_0(i) = \sum_{j=1}^{|\mathbb{X}|} u_j(i)(-1+\exp(-t\lambda_j))\alpha_0(j) = -\sum_{j>1} u_j(i)\alpha_0(j)\lambda_j\mathcal{O}(t).$$

Therefore, we have

$$s_t(x,y) = \frac{p_t(y)}{p_t(x)} = \frac{p_0(y) - \sum_{j>1}u_j(y)\alpha_0(j)\lambda_j\mathcal{O}(t)}{p_0(x) - \sum_{j>1}u_j(x)\alpha_0(j)\lambda_j\mathcal{O}(t)} \lesssim 1 \vee (Ft)^{-1},$$

given that the following condition is satisfied

$$F = \min_{x\in\mathbb{X}}\left|\sum_{j>1} u_j(x)\alpha_0(j)\lambda_j\right| > 0,$$

which only depends on the initial distribution $\boldsymbol{p}_0$ and the rate matrix $\boldsymbol{Q}$.

Especially, the bound $s_t(x,y) \lesssim 1$ for any $x \in \mathbb{X}$ s.t. $p_0(x) > 0$ and $s_t(x,y) \lesssim t^{-1}$ for those s.t. $p_0(x) = 0$.

Definition D.3 (Conductance)**.** *The conductance $\phi(\mathcal{G})$ of a graph $\mathcal{G}$ is defined as*

$$\phi(\mathcal{G}) = \min_{S \subset \mathbb{X}} \frac{\sum_{x \in S, y \notin S} Q(x,y)}{\min\left\{\sum_{x \in S} D(x,x), \sum_{y \notin S} D(y,y)\right\}}.$$

Theorem D.4 (Cheeger's Inequality)**.** *Denote the* normalized graph Laplacian *matrix by $\widetilde{\boldsymbol{L}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$ with eigenvalues*

$$0 \leq \lambda_1(\overline{\boldsymbol{L}}) \leq \lambda_2(\overline{\boldsymbol{L}}) \leq \ldots \leq \lambda_{|\mathbb{X}|}(\overline{\boldsymbol{L}}) \leq 2,$$

then the conductance of the graph $\mathcal{G}(\boldsymbol{Q})$ can be bounded by

$$\frac{1}{2}\lambda_2(\overline{\boldsymbol{L}}) \leq \phi(\mathcal{G}(\boldsymbol{Q})) \leq \sqrt{2\lambda_2(\overline{\boldsymbol{L}})}.$$

D.2 Log-Sobolev Inequalities

Definition D.5 (Modified Log-Sobolev Constant [109])**.** *For any function $f, g : \mathbb{X} \to \mathbb{R}$, we denote the* entropy functional $\mathrm{Ent}_\pi(f)$ *of f as*

$$\mathrm{Ent}_\pi(f) := \mathbb{E}_{\boldsymbol{\pi}}[f \log f] - \mathbb{E}_{\boldsymbol{\pi}}[f] \log \mathbb{E}_{\boldsymbol{\pi}}[f],$$

and the Dirichlet form $\mathcal{E}_{\boldsymbol{\pi}}(f, g)$ *as*

$$\mathcal{E}_{\boldsymbol{\pi}}(f,g) = \mathbb{E}_{\boldsymbol{\pi}}[f\boldsymbol{L}^T g] := \sum_{y \in \mathbb{X}} f(y)(\boldsymbol{L}^T g)(y)\pi(y) = \sum_{x,y \in \mathbb{X}} f(y)L(x,y)g(x)\pi(y),$$

where the Laplacian $\boldsymbol{L} = \boldsymbol{Q}$. Then the modified log-Sobolev constant *of the rate matrix $\boldsymbol{Q}$ is defined as*

$$\rho(\boldsymbol{Q}) := \inf\left\{ \frac{\mathcal{E}_{\boldsymbol{\pi}}(f, \log f)}{\mathrm{Ent}_{\boldsymbol{\pi}}(f)} \;\middle|\; f : \mathbb{X} \to \mathbb{R},\ \mathrm{Ent}_{\boldsymbol{\pi}}(f) > 0 \right\}.$$

Theorem D.6 (Theorem 2.4, [109])**.** *For any initial distribution $\boldsymbol{p}_0$, we have for any $t \geq 0$,*

$$D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi}) \leq D_{\mathrm{KL}}(\boldsymbol{p}_0 \| \boldsymbol{\pi}) \exp\left(-\rho(\boldsymbol{Q})t\right),$$

i.e. *the KL divergence converges exponentially fast with rate $\rho(\boldsymbol{Q})$.*

Proof. Noticing that $\mathrm{Ent}_{\boldsymbol{\pi}}(\frac{\boldsymbol{p}_t}{\boldsymbol{\pi}}) = D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi})$, we differentiate $D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi})$ with respect to t to obtain that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi}) =& \frac{\mathrm{d}}{\mathrm{d}t} \sum_{x \in \mathbb{X}} \frac{p_t(x)}{\pi(x)} \log\left(\frac{p_t(x)}{\pi(x)}\right) \pi(x) = \sum_{x \in \mathbb{X}} \left(\log \frac{p_t(x)}{\pi(x)} + 1\right) \pi(x) \frac{\mathrm{d}}{\mathrm{d}t} \frac{p_t(x)}{\pi(x)} \\
=& \sum_{x \in \mathbb{X}} \left(\log \frac{p_t(x)}{\pi(x)} + 1\right) \frac{\mathrm{d}}{\mathrm{d}t} p_t(x) = -\sum_{x,y \in \mathbb{X}} \log \frac{p_t(x)}{\pi(x)} L(x,y) p_t(y) \\
=& -\sum_{y \in \mathbb{X}} \frac{p_t(y)}{\pi(y)} \left(\sum_{x \in \mathbb{X}} L(x,y) \log \frac{p_t(x)}{\pi(x)}\right) \pi(y) \\
=& -\mathcal{E}_{\boldsymbol{\pi}}\left(\frac{p_t}{\pi}, \log \frac{p_t}{\pi}\right) \leq -\rho(\boldsymbol{Q}) D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi}),
\end{aligned} \tag{D.1}$$

and the result follows by applying Grönwall's inequality to both sides above. □

Then, the following proposition connects the modified log-Sobolev constant with the spectral gap.

Proposition D.7 ([109, Proposition 3.5])**.** *The modified log-Sobolev constant $\rho(\boldsymbol{Q})$ of the rate matrix $\boldsymbol{Q}$ is bounded by the spectral gap $\lambda(\boldsymbol{Q})$,* i.e. *$\rho(\boldsymbol{Q}) \leq \lambda(\boldsymbol{Q})$.*

Proof. Below we provide a sketch of the informal proof of the proposition above for the sake of completeness. Let $f : \mathbb{X} \to \mathbb{R}$ be an arbitrary function and $\zeta > 0$ be any positive number. From the definition of the modified log-Sobolev constant, we have

$$\mathcal{E}_{\boldsymbol{\pi}}(e^{\zeta f}, \zeta f) \geq \rho(\boldsymbol{Q}) \mathrm{Ent}_{\boldsymbol{\pi}}(e^{\zeta f}). \tag{D.2}$$

Under the limit $\zeta \to 0^+$, we may apply Taylor expansion to the two terms on the LHS and RHS, which implies

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\pi}}(e^{\zeta f}, \zeta f) &= \mathcal{E}_{\boldsymbol{\pi}}(1 + \zeta f + O(\zeta^2), \zeta f) \\
&= \zeta \mathcal{E}_{\boldsymbol{\pi}}(1, f) + \zeta^2 \mathcal{E}_{\boldsymbol{\pi}}(f, f) + O(\zeta^3) = \zeta^2 \mathcal{E}_{\boldsymbol{\pi}}(f, f) + O(\zeta^3) \\
\mathrm{Ent}_{\boldsymbol{\pi}}(e^{\zeta f}) &= \mathbb{E}_{\boldsymbol{\pi}}[e^{\zeta f} \zeta f] - \mathbb{E}_{\boldsymbol{\pi}}[e^{\zeta f}] \log \mathbb{E}_{\boldsymbol{\pi}}[e^{\zeta f}] \\
&= \mathbb{E}_{\boldsymbol{\pi}}\left[\left(1 + \zeta f + O(\zeta^2)\right) \zeta f\right] - \mathbb{E}_{\boldsymbol{\pi}}[1 + \zeta f + O(\zeta^2)] \log \mathbb{E}_{\boldsymbol{\pi}}[e^{\zeta f}] \\
&= \zeta \mathbb{E}_{\boldsymbol{\pi}}[f] + \zeta^2 \mathbb{E}_{\boldsymbol{\pi}}[f^2] + O(\zeta^3) \\
&\quad - \left(1 + \zeta \mathbb{E}_{\boldsymbol{\pi}}[f] + O(\zeta^2)\right) \log \left(1 + \zeta \mathbb{E}_{\boldsymbol{\pi}}[f] + O(\zeta^2)\right) \\
&= \zeta \mathbb{E}_{\boldsymbol{\pi}}[f] + \zeta^2 \mathbb{E}_{\boldsymbol{\pi}}[f^2] + O(\zeta^3) - \left(1 + \zeta \mathbb{E}_{\boldsymbol{\pi}}[f] + O(\zeta^2)\right) \left(\zeta \mathbb{E}_{\boldsymbol{\pi}}[f] + O(\zeta^2)\right) \\
&= \zeta^2 \left(\mathbb{E}_{\boldsymbol{\pi}}[f^2] - \mathbb{E}_{\boldsymbol{\pi}}[f]^2\right) + O(\zeta^3)
\end{aligned} \tag{D.3}$$

Substituting (D.3) into (D.2) and taking the limit $\zeta \to 0^+$ then yield the following inequality

$$\rho(Q) \leq \frac{\mathcal{E}_{\boldsymbol{\pi}}(f, f)}{\mathbb{E}_{\boldsymbol{\pi}}[f^2] - \mathbb{E}_{\boldsymbol{\pi}}[f]^2}$$

for any non-constant function $f : \mathbb{X} \to \mathbb{R}$. Taking infimum on both sides above with respect to all f then indicates

$$\rho(Q) \leq \inf_f \frac{\mathcal{E}_{\boldsymbol{\pi}}(f, f)}{\mathbb{E}_{\boldsymbol{\pi}}[f^2] - \mathbb{E}_{\boldsymbol{\pi}}[f]^2} \leq \inf_{f : \mathbb{E}_{\boldsymbol{\pi}}[f] = 0} \frac{\mathcal{E}_{\boldsymbol{\pi}}(f, f)}{\mathbb{E}_{\boldsymbol{\pi}}[f^2]} = \lambda_2(\boldsymbol{L}) = \lambda(\boldsymbol{Q}) \tag{D.4}$$

where the last two equalities above follows from the definition of spectral gap, as desired. □

In general, the lower bound of the modified log-Sobolev constant $\rho(\boldsymbol{Q})$ and the spectral gap $\lambda(\boldsymbol{Q})$ depends on the connectivity and other specific structures of the graph $\mathcal{G}(\boldsymbol{Q})$, and the related research is still an active area on a graph-by-graph basis [109].

The properties of the spectral gap $\lambda(\boldsymbol{Q})$ are better known in the literature, as it is closely related to the conductance of the graph $\mathcal{G}(\boldsymbol{Q})$ via Cheeger's inequality (Theorem D.4), and thus when $\boldsymbol{D} = D\boldsymbol{I}$, the spectral gap $\lambda(\boldsymbol{Q})$ satisfies

$$\frac{1}{2D} \lambda(\boldsymbol{Q}) = \frac{1}{2} \lambda_2(\overline{\boldsymbol{L}}) \leq \phi(\mathcal{G}(\boldsymbol{Q})) \leq \sqrt{2\lambda_2(\overline{\boldsymbol{L}})} = \sqrt{\frac{2\lambda(\boldsymbol{Q})}{D}}.$$

However, as shown in Proposition D.7, the lower bound on the modified log-Sobolev constant $\rho(\boldsymbol{Q})$ is hard to obtain, as the KL divergence, the exponential convergence of which is controlled by $\rho(\boldsymbol{Q})$, is stronger than the total variation distance, the exponential convergence of which is controlled by $\lambda(\boldsymbol{Q})$, via Pinsker's inequality. The following theorem gives a rough lower bound on d-regular graphs.

Theorem D.8 ([109, Proposition 5.4])**.** *Suppose $\mathcal{G}$ is a d-regular graph on $\mathbb{X}$ with unit weights and $\boldsymbol{Q}$ is the corresponding rate matrix such that $\mathcal{G}(\boldsymbol{Q}) = \mathcal{G}$, then the modified log-Sobolev constant $\rho(\boldsymbol{Q})$ of the rate matrix $\boldsymbol{Q}$ satisfies*

$$\frac{\lambda(\boldsymbol{Q})}{\log |\mathbb{X}|} \leq \rho(\boldsymbol{Q}) \leq \frac{8d \log |\mathbb{X}|}{\mathrm{diam}(\mathcal{G})^2},$$

where $\mathrm{diam}(\mathcal{G})$ *is the diameter of the graph $\mathcal{G}$.*

For some specific graphs, the modified log-Sobolev constant $\rho(\boldsymbol{Q})$ and the spectral gap $\lambda(\boldsymbol{Q})$ can be explicitly calculated, such as the following examples:

Example D.9 (Hypercube [110])**.** *Let $\mathbb{X} = \{0,1\}^d$ and $\boldsymbol{Q}$ be the rate matrix for which the graph $\mathcal{G}(\boldsymbol{Q})$ is a hypercube, and for any two states $x, y \in \mathbb{X}$, the rate $Q(x,y) = 1$ if x and y differ in exactly one coordinate. Then the modified log-Sobolev constant $\rho(\boldsymbol{Q})$ and the spectral gap $\lambda(\boldsymbol{Q})$ are given by*

$$\rho(\boldsymbol{Q}) = \lambda(\boldsymbol{Q}) = 4,$$

which is dimensionless.

Example D.10 (Asymmetric Hypercube [111])**.** *Let $\mathbb{X} = \{0,1\}^d$ and $\boldsymbol{Q}$ be the rate matrix for which the graph $\mathcal{G}(\boldsymbol{Q})$ is a hypercube, and for any two states $x, y \in \mathbb{X}$, the rate $Q(x,y) = p$ if x and y differ in exactly one coordinate and x is the state with 0 in that coordinate, and $Q(x,y) = q = 1-p$ if with 1 in that coordinate. Then the modified log-Sobolev constant $\rho(\boldsymbol{Q})$ and the spectral gap $\lambda(\boldsymbol{Q})$ are given by*

$$\rho(\boldsymbol{Q}) = \frac{2(p-q)}{pq(\log p - \log q)}, \text{ and } \lambda(\boldsymbol{Q}) = \frac{1}{pq}.$$

Further results on log-Sobolev inequalities related to finite-state Markov chains are beyond the scope of this paper, and we refer the readers to [112, 113, 114, 115, 116, 117] for more detail.

D.3 Mixing Time

Definition D.11 (Mixing Time)**.** *We define the mixing time $t_{\mathrm{mix}}(\epsilon)$ of the continuous-time Markov chain with rate matrix $\boldsymbol{Q}$ as the smallest time t such that starting from any initial distribution $\boldsymbol{p}_0$, the KL divergence $D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi})$ is less than ϵ,* i.e.

$$t_{\mathrm{mix}}(\epsilon) = \inf \left\{ t \in \mathbb{R}_+ \middle| D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi}) = D_{\mathrm{KL}}(e^{-t\boldsymbol{Q}} \boldsymbol{p}_0 \| \boldsymbol{\pi}) \leq \epsilon \right\}.$$

Similarly, we define the mixing time $t_{\mathrm{mix,TV}}(\epsilon)$ as the smallest time t such that starting from any initial distribution $\boldsymbol{p}_0$, the total variation distance $\mathrm{TV}(\boldsymbol{p}_t, \boldsymbol{\pi})$ is less than ϵ, i.e.

$$t_{\mathrm{mix,TV}}(\epsilon) = \inf \left\{ t \in \mathbb{R}_+ \middle| \mathrm{TV}(\boldsymbol{p}_t, \boldsymbol{\pi}) = \mathrm{TV}(e^{-t\boldsymbol{Q}} \boldsymbol{p}_0, \boldsymbol{\pi}) \leq \epsilon \right\}.$$

*With a slight abuse of notation, we will also denote the e^{-1}-*mixing time *as $t_{\mathrm{mix}} = t_{\mathrm{mix,KL}}(e^{-1})$ and $t_{\mathrm{mix,TV}} = t_{\mathrm{mix,TV}}(e^{-1})$.*

Proposition D.12. *The mixing time $t_{\mathrm{mix}}(\epsilon)$ of the continuous-time Markov chain with rate matrix $\boldsymbol{Q}$ is bounded by the modified log-Sobolev constant $\rho(\boldsymbol{Q})$,* i.e.

$$t_{\mathrm{mix}}(\epsilon) \lesssim \rho(\boldsymbol{Q})^{-1} \left(\log \epsilon^{-1} + \log \log \pi_*^{-1} \right),$$

And the mixing time $t_{\mathrm{mix,TV}}(\epsilon)$ is bounded by the spectral gap $\lambda(\boldsymbol{Q})$, i.e.

$$t_{\mathrm{mix,TV}}(\epsilon) \lesssim \lambda(\boldsymbol{Q})^{-1} \left(\log \epsilon^{-1} + \log \log \pi_*^{-1} \right).$$

Proof. Define $\pi_* = \min_{x \in \mathbb{X}} \pi(x)$, we first bound $D_{\mathrm{KL}}(\boldsymbol{p}_0 \| \boldsymbol{\pi})$ as follows:

$$D_{\mathrm{KL}}(\boldsymbol{p}_0 \| \boldsymbol{p}_\infty) \leq \sum_{x \in \mathbb{X}} p_0(x) \log \frac{p_0(x)}{\pi(x)} \leq \log \pi_*^{-1}, \tag{D.5}$$

and thus by Theorem D.6, we have

$$D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi}) \leq D_{\mathrm{KL}}(\boldsymbol{p}_0 \| \boldsymbol{\pi}) \exp\left(-\rho(\boldsymbol{Q}) t\right) \leq \log \pi_*^{-1} \exp\left(-\rho(\boldsymbol{Q}) t\right). \tag{D.6}$$

Therefore, by setting the right-hand side of (D.6) to be ϵ, we have the desired result for the mixing time

$$t_{\mathrm{mix}}(\epsilon) \leq \frac{1}{\rho(\boldsymbol{Q})} \left(\log \epsilon^{-1} + \log \log \pi_*^{-1} \right).$$

For the mixing time $t_{\mathrm{mix,TV}}(\epsilon)$, we use the Pinsker's inequality to obtain:

$$\mathrm{TV}(\boldsymbol{p}_t, \boldsymbol{\pi}) \leq 2\sqrt{D_{\mathrm{KL}}(\boldsymbol{p}_t \| \boldsymbol{\pi})} \leq 2\sqrt{\log \pi_*^{-1}} \exp\left(-\rho(\boldsymbol{Q}) t\right),$$

and therefore,

$$t_{\mathrm{mix,TV}}(\epsilon) \lesssim \frac{1}{\rho(\boldsymbol{Q})} \log\left(\epsilon^{-1}\sqrt{\log \pi_*^{-1}}\right) \lesssim \frac{1}{\rho(\boldsymbol{Q})}\left(\log \epsilon^{-1} + \log\log \pi_*^{-1}\right).$$

□

Corollary D.13. *The e^{-1}-mixing time t_{mix} and $t_{\mathrm{mix,TV}}$ of the continuous-time Markov chain with rate matrix $\boldsymbol{Q}$ satisfy*

$$t_{\mathrm{mix}} \lesssim \rho(\boldsymbol{Q})^{-1}\log\log \pi_*^{-1}, \quad \textit{and} \quad t_{\mathrm{mix,TV}} \lesssim \lambda(\boldsymbol{Q})^{-1}\log\log \pi_*^{-1},$$

and thus $t_{\mathrm{mix}} \gtrsim t_{\mathrm{mix,TV}}$.

E Proofs of Error Analysis in Section 4.3

In this section, we provide the proofs of the main results in Section 4.3. We first introduce the assumptions in Section 4.2 and then provide the proofs of the main results in Section **??**.

E.1 Discussions on Assumptions in Section 4.2

We need the following assumptions to ensure the well-definedness of discrete diffusion models. For simplicity, we assume the rate matrix $\boldsymbol{Q}_t$ is time-homogeneous and symmetric, *i.e.* $\boldsymbol{Q}_t = \boldsymbol{Q}$ for any $t \geq 0$. In fact, the results can be easily extended to the time-inhomogeneous case of the family $\boldsymbol{Q}_t = \beta_t \boldsymbol{Q}$ with a rescaling factor β_t, and asymmetric cases will be left for future works.

Assumption E.1 (Regularity of the Rate Matrix)**.** *The rate matrix $\boldsymbol{Q}$ satisfies the following conditions:*

(i) For any $x, y \in \mathbb{X}$, $Q(x,y) \leq C$ and $\underline{D} \leq -Q(x,x) \leq \overline{D}$ for some constants $C, \underline{D}, \overline{D} > 0$;

(ii) The modified log-Sobolev constant *$\rho(\boldsymbol{Q})$ of the rate matrix $\boldsymbol{Q}$ (*cf. *Definition D.5) is lower bounded by $\rho > 0$.*

Statement (i) assumes the regularity of the rate matrix, which is often trivially satisfied in many applications, while Statement (ii) ensures the exponential convergence of the forward process in discrete diffusion models. In general, $\rho(\boldsymbol{Q})$ may depend on the connectivity and other structures of the corresponding graph $\mathcal{G}(\boldsymbol{Q})$ (*cf.* Definition D.1). Such lower bound has been obtained for specific graphs (*e.g.* Example D.9 and D.10), and general results are in active research [113, 109]. We refer readers to Appendix D for further discussions on the literature of the modified log-Sobolev constant, as well as its relation to the spectral gap, the mixing time, *etc.*.

Assumption E.2 (Bounded Score)**.** *The true score function satisfies $s_t(x,y) \lesssim 1 \vee t^{-1}$, while the learned score function satisfies $\widehat{s}_s^{\theta}(x,y) \in (0, M]$, for any $x, y \in \mathbb{X}$.*

The first part on the asymptotic behavior of the true score corresponds to the estimation $\mathbb{E}[\|\boldsymbol{s}_t\|^2] \sim \mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{\mu}_t\|^2/\sigma_t^2] \sim 1 \vee t^{-1}$ in the continuous case [27, Assumption 1] and further justification is provided in Remark D.2. The bound on the estimated score can be easily satisfied by adding truncation in post-processing in the implementation of the NN-based score estimator.

Assumption E.3 (Continuity of Score Function)**.** *For any $t > 0$ and $y \in \mathbb{X}$ such that $Q(x_{t^-}, y) > 0$, we have $\left|\frac{\mu_{t^+}(y)}{\mu_t(y)}\right| := \left|\frac{p_t(x_{t^-})Q(x_t,y)}{p_t(x_t)Q(x_{t^-},y)} - 1\right| \lesssim 1 \vee t^{-\gamma}$, for some exponent $\gamma \in [0,1]$.*

Assumption E.3 corresponds to the Lipschitz continuity of the score function (*cf.* [90, Assumption 1], [27, Assumption 3]) for continuous diffusion models, and is in light of the postulation that adjacent vertices should have close score function and intensity values. In the worse case, assume $Q(x,y) = \Theta(1)$, then a naïve bound would be $\left|\frac{\mu_{t^+}(y)}{\mu_t(y)}\right| \lesssim |s_t(x_t, x_{t^-})| \lesssim 1 \vee t^{-1}$ with $\gamma = 1$. However, when the initial distribution is both upper and lower bounded, γ may be as small as 0, and we plan to investigate how this (local) continuity of the score function affects the overall performance of discrete diffusion models.

Assumption E.4 (ϵ-accurate Score Estimation)**.** *The score function $s_t(x_t)$ is estimated by the neural network $\widehat{s}_t^\theta(x_t)$ with ϵ-accuracy,* i.e.

$$\sum_{n=0}^{N-1}(s_{n+1}-s_n)\mathbb{E}\left[\int_{\mathbb{X}} K\left(\frac{\overleftarrow{\widehat{s}}^\theta_{s_n}(\overleftarrow{x}_{s_n^-},y)}{\overleftarrow{s}_{s_n}(\overleftarrow{x}_{s_n^-},y)}\right)\overleftarrow{s}_{s_n}(\overleftarrow{x}_{s_n^-},y)\widetilde{Q}(\overleftarrow{x}_{s_n^-},y)\nu(\mathrm{d}y)\right]\le\epsilon.$$

This assumption assumes the expressive power and sufficient training of the NN-based score estimator and is standard in diffusion model-related theories [90, 27, 28, 96].

E.2 Truncation Error

Theorem E.5 (Truncation Error)**.** *The forward process (2.1) converges to the uniform distribution $\boldsymbol{p}_\infty=\mathbf{1}/|\mathbb{X}|$ exponentially fast in terms of the KL divergence,* i.e.

$$D_{\mathrm{KL}}(\boldsymbol{p}_t\|\boldsymbol{p}_\infty)=D_{\mathrm{KL}}\left(\boldsymbol{p}_t\middle\|\frac{\mathbf{1}}{|\mathbb{X}|}\right)\lesssim e^{-\rho t}\log|\mathbb{X}|,$$

where $|\mathbb{X}|$ is the size of the state space, and t_{mix} is the mixing time of the continuous-time Markov chain corresponding to the rate matrix $\boldsymbol{Q}$ defined in Definition D.11.

Proof. Since $\boldsymbol{Q}$ is symmetric, we have the stationary distribution $\boldsymbol{\pi}=\mathbf{1}/|\mathbb{X}|$ and thus $D_{\mathrm{KL}}(\boldsymbol{p}_t\|\boldsymbol{p}_\infty)=D_{\mathrm{KL}}(\boldsymbol{p}_t\|\boldsymbol{\pi})$, and $\pi_*=1/|\mathbb{X}|$.

By Assumption E.1 and Corollary D.13, we have

$$D_{\mathrm{KL}}(\boldsymbol{p}_t\|\boldsymbol{\pi})\le e^{-\rho(\boldsymbol{Q})t}D_{\mathrm{KL}}(\boldsymbol{p}_0\|\boldsymbol{\pi})\le e^{-\rho t}\log\pi_*^{-1}\le e^{-\rho t}\log|\mathbb{X}|,$$

where the last inequality is by (D.5). □

E.3 Discretization Error

Denote the shorthand notation $G(x;y)=x(\log x-\log y)-x$; it is easy to check that $G'(x;y)=\log x-\log y$.

Proposition E.6. *For any $y\in\mathbb{X}$, we have*

$$|\partial_\sigma G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))|\lesssim\left(\log C+\log\left(1\vee(T-\sigma)\right)+\log M\right)\mu_\sigma(y)\overline{D}\left(1\vee(T-\sigma)^{-2}\right).$$

Proof. By the chain rule, we have

$$\partial_\sigma G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))=G'(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))\partial_\sigma\mu_\sigma(y)=\left(\log\mu_\sigma(y)-\log\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y)\right)\partial_\sigma\mu_\sigma(y).$$

We first compute $\partial_\sigma\mu_\sigma(y)$ as

$$\begin{aligned}
&\partial_\sigma\mu_\sigma(y)=\widetilde{Q}(\overleftarrow{x}_{\sigma^-},y)\partial_\sigma\overleftarrow{s}_\sigma(\overleftarrow{x}_{\sigma^-},y)=\widetilde{Q}(\overleftarrow{x}_{\sigma^-},y)\partial_\sigma\left(\frac{\overleftarrow{p}_\sigma(y)}{\overleftarrow{p}_\sigma(x_{\sigma^-})}\right)\\
=&\widetilde{Q}(\overleftarrow{x}_{\sigma^-},y)\left(\frac{1}{\overleftarrow{p}_\sigma(x_{\sigma^-})}\partial_\sigma\overleftarrow{p}_\sigma(y)-\frac{\overleftarrow{p}_\sigma(y)}{\overleftarrow{p}_\sigma(x_{\sigma^-})^2}\partial_\sigma\overleftarrow{p}_\sigma(x_{\sigma^-})\right)\\
=&\widetilde{Q}(\overleftarrow{x}_{\sigma^-},y)\left(-\frac{\overleftarrow{p}_\sigma(y)}{\overleftarrow{p}_\sigma(x_{\sigma^-})}\sum_{y'\in\mathbb{X}}\frac{\overleftarrow{p}_\sigma(y')}{\overleftarrow{p}_\sigma(y)}Q(y,y')+\frac{\overleftarrow{p}_\sigma(y)}{\overleftarrow{p}_\sigma(x_{\sigma^-})}\sum_{y'\in\mathbb{X}}\frac{\overleftarrow{p}_\sigma(y')}{\overleftarrow{p}_\sigma(x_{\sigma^-})}Q(x_{\sigma^-},y')\right)\\
=&\mu_\sigma(y)\left(-\sum_{y'\in\mathbb{X}}\overleftarrow{s}_\sigma(y,y')Q(y,y')+\sum_{y'\in\mathbb{X}}\overleftarrow{s}_\sigma(x_{\sigma^-},y')Q(x_{\sigma^-},y')\right),
\end{aligned}$$

by which we have

$$|\partial_\sigma\mu_\sigma(y)|\lesssim\mu_\sigma(y)\left(\sum_{y'\in\mathbb{X}}\left(1\vee(T-\sigma)^{-2}\right)|Q(x_{\sigma^-},y')|\right)\lesssim\mu_\sigma(y)\overline{D}\left(1\vee(T-\sigma)^{-2}\right),$$

and thus

$$\begin{aligned}
&\left|\partial_\sigma G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))\right| \le \left|\log\mu_\sigma(y)-\log\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y)\right|\left|\partial_\sigma\mu_\sigma(y)\right| \\
\le&\left(|\log\widetilde{Q}(\bar{x}_{\sigma^-},y)|+|\log\bar{s}_\sigma(\bar{x}_{\sigma^-},y)|+|\log\widetilde{Q}(\bar{x}_{s_n^-},y)|+|\log\overleftarrow{\widehat{s}}^\theta_{s_n}(\bar{x}_{s_n^-},y)|\right)|\partial_\sigma\mu_\sigma(y)| \\
\lesssim&\mu_\sigma(y)\left(\log C+\log\left(1\vee(T-\sigma)^{-1}\right)+\log M\right)\overline{D}\left(1\vee(T-\sigma)^{-2}\right).
\end{aligned}$$

□

Proposition E.7. *For any $0<s<t\le T$, we have*

$$\int_s^t\int_{\mathbb{X}}\mu_\sigma(y')\nu(\mathrm{d}y')\mathrm{d}\sigma\lesssim\left(1\vee(T-t)^{-1}\right)\overline{D}(t-s).$$

Proof.

$$\begin{aligned}
&\int_s^t\int_{\mathbb{X}}\mu_\sigma(y')\nu(\mathrm{d}y')\mathrm{d}\sigma=\int_s^t\int_{\mathbb{X}}\bar{s}_\sigma(\bar{x}_{\sigma^-},y')\widetilde{Q}(\bar{x}_{\sigma^-},y')\nu(\mathrm{d}y')\mathrm{d}\sigma \\
\lesssim&\left(1\vee(T-t)^{-1}\right)\int_s^t\int_{\mathbb{X}}\widetilde{Q}(y,\bar{x}_{\sigma^-})\nu(\mathrm{d}y')\mathrm{d}\sigma \qquad \text{(E.1)}\\
\lesssim&\left(1\vee(T-t)^{-1}\right)\int_s^t|Q(\bar{x}_{\sigma^-},\bar{x}_{\sigma^-})|\,\mathrm{d}\sigma\lesssim\left(1\vee(T-t)^{-1}\right)\overline{D}(t-s).
\end{aligned}$$

□

Proposition E.8. *For any $y\in\mathbb{X}$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left|G(\mu_s;\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n};\widehat{\mu}^\theta_{\lfloor s\rfloor}(y))\right|\right] \\
\lesssim&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\mu_\sigma(y)\overline{D}(s-s_n)\left(1\vee(T-s_{n+1})^{-1-\gamma}\right).
\end{aligned}$$

Proof. Applying Theorem C.9 to the backward process (3.2), we have

$$\begin{aligned}
G(\mu_s(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))=&G(\mu_{s_n}(y);\widehat{\mu}^\theta_{\lfloor s\rfloor}(y))+\int_{s_n}^s\partial_\sigma G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))\mathrm{d}\sigma \\
&+\int_{s_n}^s\int_{\mathbb{X}}\left(G(\mu_{\sigma^+}(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))-G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))\right)N[\mu](\mathrm{d}\sigma,\mathrm{d}y'),
\end{aligned}$$

where we adopt the notation $\mu_{\sigma^+}(y)$ as the right limit of the càglàd process $\mu_\sigma(y)$, *i.e.* $\mu_{\sigma^+}(y)=\bar{s}_\sigma(\bar{x}_\sigma,y)\widetilde{Q}(\bar{x}_\sigma,y)$.

For the first term, we have by Proposition E.6 that

$$\begin{aligned}
&\left|\int_{s_n}^s\partial_\sigma G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))\mathrm{d}\sigma\right|\lesssim\int_{s_n}^s\left|\partial_\sigma G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))\right|\mathrm{d}\sigma \\
\lesssim&\int_{s_n}^s\left(\log C+\log\left(1\vee(T-\sigma)^{-1}\right)+\log M\right)\mu_\sigma(y)\overline{D}\left(1\vee(T-\sigma)^{-2}\right)\mathrm{d}\sigma \\
\lesssim&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\mu_\sigma(y)\overline{D}(s-s_n)\left(1\vee(T-s_{n+1})^{-1}\right).
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\left|G(\mu_{\sigma^+}(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))-G(\mu_\sigma(y);\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))\right| \\
=&\left|G'(\xi;\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))(\mu_{\sigma^+}(y)-\mu_\sigma(y))\right|=\left|(\log\xi-\log\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y))(\mu_{\sigma^+}(y)-\mu_\sigma(y))\right| \\
\le&\left(|\log\mu_{\sigma^+}(y)|+|\log\mu_\sigma(y)|+\left|\log\widehat{\mu}^\theta_{\lfloor s_n\rfloor}(y)\right|\right)\left|\frac{\mu_{\sigma^+}(y)}{\mu_\sigma(y)}-1\right|\mu_\sigma(y) \\
\lesssim&\left(\log C+\log\left(1\vee(T-\sigma)^{-1}\right)+\log M\right)\left(1\vee(T-\sigma)^{-\gamma}\right)\mu_\sigma(y),
\end{aligned}$$

where the first equality follows from the mean value theorem and the last inequality is by Assumption E.3.

Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\left|G(\mu_s;\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n};\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y))\right|\right]\\
\leq&\mathbb{E}\left[\left|\int_{s_n}^{s}\partial_\sigma G(\mu_\sigma(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))\mathrm{d}\sigma\right|\right]\\
+&\mathbb{E}\left[\int_{s_n}^{s}\int_{\mathbb{X}}\left|G(\mu_{\sigma^+}(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_\sigma(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))\right|N[\mu](\mathrm{d}\sigma,\mathrm{d}y')\right]\\
\lesssim&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\mu_\sigma(y)\overline{D}(s-s_n)\left(1\vee(T-s_{n+1})^{-1}\right)\\
+&\int_{s_n}^{s}\int_{\mathbb{X}}\left(\log C+\log\left(1\vee(T-\sigma)^{-1}\right)+\log M\right)\mu_\sigma(y)\left(1\vee(T-\sigma)^{-\gamma}\right)\mu_\sigma(y')\nu(\mathrm{d}y')\mathrm{d}\sigma\\
\lesssim&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\mu_\sigma(y)\overline{D}(s-s_n)\left(1\vee(T-s_{n+1})^{-1}\right)\\
+&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\mu_\sigma(y)\left(1\vee(T-s_{n+1})^{-1-\gamma}\right)(s-s_n)\overline{D}\\
\lesssim&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\mu_\sigma(y)\overline{D}(s-s_n)\left(1\vee(T-s_{n+1})^{-1-\gamma}\right),
\end{aligned}
$$

where the second to last inequality is by Proposition E.3. □

Proposition E.9 (Discretization Error)**.** *The following bound holds*

$$
\begin{aligned}
&\int_0^{T-\delta}\int_{\mathbb{X}}\left|G(\mu_s(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n}(y);\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y))\right|\nu(\mathrm{d}y)\mathrm{d}s\\
\lesssim&\begin{cases}\overline{D}^2\kappa T, & \gamma<1,\\ \overline{D}^2\kappa\left(T+\log^2\delta^{-1}\right), & \gamma=1,\end{cases}
\end{aligned}
$$

with $N=\begin{cases}\kappa^{-1}T, & \gamma<1\\ \kappa^{-1}(T+\log\delta^{-1}), & \gamma=1\end{cases}$ *steps, by taking* $\gamma<\eta\lesssim 1-T^{-1}$ *when* $\gamma<1$*, and* $\eta=1$ *when* $\gamma=1$*. In particular, in the former case, early stopping at time* $T-\delta$ *is not necessary,* i.e. $\delta=0$*.*

Proof. We have by Proposition E.8 that

$$
\begin{aligned}
&\int_{s_n}^{s_{n+1}}\int_{\mathbb{X}}\left|G(\mu_s(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n}(y);\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y))\right|\nu(\mathrm{d}y)\mathrm{d}s\\
\lesssim&\int_{s_n}^{s_{n+1}}\int_{\mathbb{X}}\mu_\sigma(y)\nu(\mathrm{d}y)\\
&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\overline{D}(s-s_n)\left(1\vee(T-s_{n+1})^{-1-\gamma}\right)\mathrm{d}s\\
\lesssim&\left(\log C+\log\left(1\vee(T-s_{n+1})^{-1}\right)+\log M\right)\overline{D}^2(s_{n+1}-s_n)^2\left(1\vee(T-s_{n+1})^{-1-\gamma}\right).
\end{aligned}
$$

- *Case 1:* $\gamma<\eta\lesssim 1-T^{-1}$

  The following bound holds

$$
\begin{aligned}
&\int_{s_n}^{s_{n+1}}\int_{\mathbb{X}}\left|G(\mu_s(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n}(y);\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y))\right|\nu(\mathrm{d}y)\mathrm{d}s\\
\lesssim&\left(1+\log\left(1\vee(T-s_{n+1})^{-1}\right)\right)\overline{D}^2\kappa(s_{n+1}-s_n)\left(1\vee(T-s_{n+1})^{-1-\gamma+\eta}\right),
\end{aligned}
$$

and thus, the following error

$$\begin{aligned}
&\sum_{n=0}^{N-1}\int_{s_n}^{s_{n+1}}\int_{\mathbb{X}}\left|G(\mu_s(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n}(y);\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y))\right|\nu(\mathrm{d}y)\mathrm{d}s\\
\lesssim&\sum_{n=0}^{N-1}\left(1+\log\left(1\vee(T-s_{n+1})^{-1}\right)\right)\overline{D}^2\kappa(s_{n+1}-s_n)\left(1\vee(T-s_{n+1})^{-1-\gamma+\eta}\right)\\
\lesssim&\overline{D}^2\kappa\left(T+\int_{\delta}^{1}t^{-1-\gamma+\eta}\log t^{-1}\mathrm{d}t\right)\lesssim\overline{D}^2\kappa\left(T+\int_{1}^{\delta^{-1}}t^{-1-(\eta-\gamma)}\log t\mathrm{d}t\right)\\
\lesssim&\overline{D}^2\kappa\left(T+\delta^{\eta-\gamma}\log\delta^{-1}\right)\to\overline{D}^2\kappa T,\quad\text{as }\delta\to 0,
\end{aligned}$$

is achievable with finite number of steps N, *i.e.*

$$N\lesssim\int_{\delta}^{T}\frac{1}{\kappa\left(1\wedge t^{\eta}\right)}\mathrm{d}t\lesssim\kappa^{-1}T+\kappa^{-1}\int_{\delta}^{1}t^{-\eta}\mathrm{d}t\lesssim\kappa^{-1}\left(T+\frac{1}{1-\eta}\right)\lesssim\kappa^{-1}T,$$

where we take $\eta\lesssim 1-T^{-1}$.

- *Case 2:* $\gamma=\eta=1$

  We have the following bound

$$\begin{aligned}
&\int_{s_n}^{s_{n+1}}\int_{\mathbb{X}}\left|G(\mu_s(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n}(y);\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y))\right|\nu(\mathrm{d}y)\mathrm{d}s\\
\lesssim&\left(1+\log\left(1\vee(T-s_{n+1})^{-1}\right)\right)\overline{D}^2\kappa(s_{n+1}-s_n)\left(1\vee(T-s_{n+1})^{-1}\right),
\end{aligned}$$

  and similarly

$$\begin{aligned}
&\sum_{n=0}^{N-1}\int_{s_n}^{s_{n+1}}\int_{\mathbb{X}}\left|G(\mu_s(y);\widehat{\mu}^{\theta}_{\lfloor s_n\rfloor}(y))-G(\mu_{s_n}(y);\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y))\right|\nu(\mathrm{d}y)\mathrm{d}s\\
\lesssim&\sum_{n=0}^{N-1}\left(1+\log\left(1\vee(T-s_{n+1})^{-1}\right)\right)\overline{D}^2\kappa(s_{n+1}-s_n)\left(1\vee(T-s_{n+1})^{-1}\right)\\
\lesssim&\overline{D}^2\kappa\left(T+\int_{1}^{\delta^{-1}}t^{-1}\log t\mathrm{d}t\right)\lesssim\overline{D}^2\kappa\left(T+\log^2\delta^{-1}\right).
\end{aligned}$$

  However, the number of steps N is now bounded by

$$N\lesssim\int_{\delta}^{T}\frac{1}{\kappa(1\wedge t)}\mathrm{d}t\lesssim\kappa^{-1}(T+\log\delta^{-1}).$$

□

E.4 Overall Error Bound

Theorem E.10. *Let $\overleftarrow{p}_{0:T}$ and $\widehat{q}_{0:T}$ be the path measures of the backward process with the stochastic integral formulation (3.2) and the interpolating process (4.1) of τ-leaping algorithm (Algorithm (1)), then it holds that*

$$D_{\mathrm{KL}}(\overleftarrow{p}_{0:T}\|\widehat{q}_{0:T})=D_{\mathrm{KL}}(\overleftarrow{p}_0\|\widehat{q}_0)+\mathbb{E}\left[\int_0^T\int_{\mathbb{X}}\left(\mu_s(y)\log\frac{\mu_s(y)}{\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y)}-\mu_s(y)+\widehat{\mu}^{\theta}_{\lfloor s\rfloor}(y)\right)\nu(\mathrm{d}y)\mathrm{d}t\right],\tag{E.2}$$

where the expectation is taken w.r.t. paths generated by the backward process (3.2).

Now, we are ready to present the proof of Theorem 4.7.

Proof of Theorem 4.7. We first rewrite the integral in (E.2) as

$$\begin{aligned}
&\int_{s_n}^{s_{n+1}} \int_{\mathbb{X}} \left(\mu_s(y) \log \frac{\mu_s(y)}{\widehat{\mu}^\theta_{\lfloor s \rfloor}(y)} - \mu_s(y) + \widehat{\mu}^\theta_{\lfloor s \rfloor}(y) \right) \nu(\mathrm{d}y)\mathrm{d}s \\
=&\int_{s_n}^{s_{n+1}} \int_{\mathbb{X}} \Bigg(\mu_{s_n}(y) \log \frac{\mu_{s_n}(y)}{\widehat{\mu}^\theta_{\lfloor s_n \rfloor}(y)} - \mu_{s_n}(y) + \widehat{\mu}^\theta_{\lfloor s \rfloor}(y) \\
&\qquad + G(\mu_s(y); \widehat{\mu}^\theta_{\lfloor s_n \rfloor}(y)) - G(\mu_{s_n}(y); \widehat{\mu}^\theta_{\lfloor s \rfloor}(y)) \Bigg) \nu(\mathrm{d}y)\mathrm{d}s.
\end{aligned}$$

Therefore, the overall error is bounded by

$$\begin{aligned}
&D_{\mathrm{KL}}(\overleftarrow{p}_{0:T-\delta} \| \widehat{q}_{0:T-\delta}) \\
\lesssim& D_{\mathrm{KL}}(\overleftarrow{p}_0 \| \widehat{q}_0) + \mathbb{E}\left[\int_0^{T-\delta} \int_{\mathbb{X}} \left(\mu_{s_n}(y) \log \frac{\mu_{s_n}(y)}{\widehat{\mu}^\theta_{\lfloor s_n \rfloor}(y)} - \mu_{s_n}(y) + \widehat{\mu}^\theta_{\lfloor s_n \rfloor}(y) \right) \nu(\mathrm{d}y)\mathrm{d}t \right] \\
&+ \int_{s_n}^{s_{n+1}} \int_{\mathbb{X}} \left| G(\mu_s(y); \widehat{\mu}^\theta_{\lfloor s_n \rfloor}(y)) - G(\mu_{s_n}(y); \widehat{\mu}^\theta_{\lfloor s \rfloor}(y)) \right| \nu(\mathrm{d}y)\mathrm{d}s \\
\lesssim& D_{\mathrm{KL}}(p_{T-\delta} \| p_\infty) \\
&+ \mathbb{E}\Bigg[\sum_{n=0}^{N-1} (s_{n+1} - s_n) \\
&\qquad \int_{\mathbb{X}} \left(\overleftarrow{s}_{s_n}(\overleftarrow{x}_{s_n^-}, y) \log \frac{\overleftarrow{s}_{s_n}(\overleftarrow{x}_{s_n^-}, y)}{\overleftarrow{s}^\theta_{s_n}(\overleftarrow{x}_{s_n^-}, y)} - \overleftarrow{s}_{s_n}(\overleftarrow{x}_{s_n^-}, y) + \overleftarrow{s}^\theta_{s_n}(\overleftarrow{x}_{s_n^-}, y) \right) \widetilde{Q}(\overleftarrow{x}_{s_n^-}, y) \nu(\mathrm{d}y) \Bigg] \\
&+ \sum_{n=0}^{N-1} (\log C + \log M) \, \overline{D}^2 \kappa (s_{n+1} - s_n) \\
\lesssim& \begin{cases} \exp(-\rho T) \log |\mathbb{X}| + \epsilon + \overline{D}^2 \kappa T, & \gamma < 1, \\ \exp(-\rho T) \log |\mathbb{X}| + \epsilon + \overline{D}^2 \kappa \left(T + \log^2 \delta^{-1} \right), & \gamma = 1, \end{cases}
\end{aligned}$$

where in the last inequality we used results for the first term (Truncation error, *cf.* Theorem E.5), the second term (Approximation error, *cf.* Assumption E.4) and the third term (Discretization error, *cf.* Proposition E.9).

By taking

$$T = \mathcal{O}\left(\frac{\log\left(\epsilon^{-1} \log |\mathbb{X}|\right)}{\rho} \right), \quad \kappa = \mathcal{O}\left(\frac{\epsilon \rho}{\overline{D}^2 \log\left(\epsilon^{-1} \log |\mathbb{X}|\right)} \right),$$

deploying the time discretization scheme with $\gamma < \eta \lesssim 1 - T^{-1}$ when $\gamma < 1$, and $\eta = 1$ when $\gamma = 1$, and performing early stopping as

$$\delta = \begin{cases} 0, & \gamma < 1, \\ \Omega\left(\exp(-\sqrt{T}) \right), & \gamma = 1, \end{cases}$$

we have $D_{\mathrm{KL}}(\overleftarrow{p}_{T-\delta} \| \widehat{q}_T) \le D_{\mathrm{KL}}(\overleftarrow{p}_{0:T-\delta} \| \widehat{q}_{0:T-\delta}) \lesssim \epsilon$ with

$$N \lesssim \kappa^{-1} T = \mathcal{O}\left(\frac{\overline{D}^2 \log^2\left(\epsilon^{-1} \log |\mathbb{X}|\right)}{\epsilon \rho^2} \right)$$

steps. □

Remark E.11 (Remark on Early Stopping)**.** *As in Assumption E.2, the true score function may exhibit singular behavior as $s \to T$, due to possible vacancy in the target distribution $\boldsymbol{p}_0$. To handle this singularity, two different regimes are considered for the time discretization scheme depending on the continuity parameter γ of the score function (Assumption E.3). The main intuition is that (a) in the worse case $\gamma = 1$, early stopping at time $s = T - \delta$ is necessary; (b) if the target distribution $\boldsymbol{p}_0$ is such well-posed* (e.g. *both upper and lower bounded) and the rate matrix $\boldsymbol{Q}$ is constructed in a way that the score exhibits certain (local) continuity reflected by $\gamma < 1$, one may choose an appropriate shrinkage η, with which finite discretization error can be achieved with finite steps.*

Proof of Theorem 4.8. Due to the equivalence of the stochastic integral formulation (4.2) of the uniformization scheme and the approximate backward process (C.13) established in Proposition 4.2, the error for the uniformization scheme is directly bounded by the error (E.2) in Corollary C.11, *i.e.*

$$\begin{aligned}
&D_{\mathrm{KL}}(\overleftarrow{p}_{0:T-\delta}\|q_{0:T-\delta})\\
\leq&D_{\mathrm{KL}}(\overleftarrow{p}_0\|q_0)+\mathbb{E}\left[\int_0^{T-\delta}\int_{\mathbb{X}}\left(\mu_s(y)\log\frac{\mu_s(y)}{\widehat{\mu}_s^\theta(y)}-\mu_s(y)+\widehat{\mu}_s^\theta(y)\right)\nu(\mathrm{d}y)\mathrm{d}t\right]\\
\lesssim&D_{\mathrm{KL}}(p_T\|p_\infty)\\
+&\mathbb{E}\left[\int_0^{T-\delta}\int_{\mathbb{X}}\left(\overleftarrow{s}_s(\overleftarrow{x}_{s^-},y)\log\frac{\overleftarrow{s}_s(\overleftarrow{x}_{s^-},y)}{\overleftarrow{s}_s^\theta(\overleftarrow{x}_{s^-},y)}-\overleftarrow{s}_s(\overleftarrow{x}_{s^-},y)+\overleftarrow{s}_s^\theta(\overleftarrow{x}_{s^-},y)\right)\widetilde{Q}(\overleftarrow{x}_{s^-},y)\nu(\mathrm{d}y)\mathrm{d}s\right]\\
\lesssim&|\mathbb{X}|\exp(-\rho T)+\epsilon.
\end{aligned}$$

The expectation of the number of steps N is bounded by

$$\begin{aligned}
\mathbb{E}[N]=\mathbb{E}\left[\sum_{b=0}^{B-1}\mathcal{P}\left(\overline{\lambda}_{s_{b+1}}(s_{b+1}-s_b)\right)\right]=\sum_{b=0}^{B-1}\overline{\lambda}_{s_{b+1}}(s_{b+1}-s_b)\\
\lesssim\sum_{b=0}^{B-1}\overline{D}(1\vee(T-s_{b+1}))^{-1}(s_{b+1}-s_b)\\
\lesssim\overline{D}\left(T+\int_\delta^1 t^{-1}\mathrm{d}t\right)=\overline{D}(T+\log\delta^{-1}).
\end{aligned}$$

By taking

$$T=\mathcal{O}\left(\frac{\log\left(\epsilon^{-1}\log|\mathbb{X}|\right)}{\rho}\right),\quad \delta=\Omega\left(\exp(-T)\right)$$

we have $D_{\mathrm{KL}}(\overleftarrow{p}_{T-\delta}\|q_{T-\delta})\leq D_{\mathrm{KL}}(\overleftarrow{p}_{0:T-\delta}\|q_{0:T-\delta})\lesssim\epsilon$ with

$$\mathbb{E}[N]=\mathcal{O}\left(\frac{\overline{D}\log\left(\epsilon^{-1}\log|\mathbb{X}|\right)}{\rho}\right)$$

steps. □