Introduction to Discrete Diffusion

Leo Zhang

University of Oxford

July 30, 2025

Contents

1. **Continuous Time Markov Chains (CTMCs)**
2. **Historical Development**
3. **Further Topics**
4. **Applications**

Continuous Time Markov Chains (CTMCs)

Overview

There are many representations of CTMCs:

- Dynamics
- Densities
- Generator/Rate Matrix

However, I don't think the literature around discrete diffusion does a great job at showing the connection between these different viewpoints. Furthermore, the literature around CTMCs mainly only explores the time-homogeneous case, whereas discrete diffusions are time-inhomogeneous.

Time-Homogeneous CTMCs

Definition

Let $S := [N] = \{1, \ldots, N\}$ be a finite set. A Q-matrix/rate matrix on S is defined as any matrix Q of the form

1. $Q_{ij} \geq 0$ for all $i \neq j$;
2. $0 \leq -Q_{ii} < \infty$ for all i
3. $\sum_{j \in S} Q_{ij} = 0 \implies Q_{ii} = -\sum_{j \neq i} Q_{ij}$.

Time-Homogeneous CTMCs

Definition

A (càdlàg) time-homogeneous CTMC $(X_t)_{t\geq 0}$ is defined by the dynamics:

1. Sample $\epsilon_n \sim \text{Exp}(-Q_{ii})$ and set $T_{n+1} := T_n + \epsilon_n$;
2. For all $T_n \leq s < T_{n+1}$, we set $X_s := X_{T_n}$;
3. At time T_{n+1}, we sample the next state according to the probabilities $K(i, j) = Q_{i,j} / - Q_{i,i}$ for $j \neq i$,

with the initial conditions $X_0 = x_0$ and $T_0 = 0$.

Time-Homogeneous CTMCs

From this we can derive further results about $(X_t)_{t\geq 0}$ Norris (1998):

- The transition probabilities only depend on relative times (time-homogeneous): $\mathbb{P}(X_t = x' | X_s = x) = \mathbb{P}(X_{t-s} = x' | X_0 = x)$;
- The ODE governing the probability flow $p_t = \exp(tQ)$ induced by $(X_t)_{t\geq 0}$;
- First order approximations of transition kernels in terms of Q;
- The time-reversal of $(X_t)_{t\geq 0}$;
- The strong Markov property.

However, this still raises the question: where does this "rate matrix" come from and how does it encode information about the law of $(X_t)_{t\geq 0}$?

Markov Processes

Definition

A stochastic process $(X_t)_{t\geq 0}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ living in the state space $(S, \tilde{\mathcal{F}})$ is a Markov process if for all $s, t \geq 0$:

$$\mathbb{P}(X_{s+t} \in A | \mathcal{F}_s) = \mathbb{P}(X_{s+t} \in A | X_s),$$

where $A \in \tilde{\mathcal{F}}$ and $\mathcal{F}_s = \sigma(X_r : 0 \leq r \leq s)$ is the σ-algebra/filtration generated $(X_t)_{t\geq 0}$ up to and including time $s \geq 0$.

Remark

We say a Markov process $(X_t)_{t\geq 0}$ is time-homogeneous if for all $s, t \geq 0$:

$$\mathbb{P}(X_{s+t} \in A | X_s) = \mathbb{P}(X_t \in A | X_0).$$

Markov Processes

Markov processes are usually studied in terms of their transition kernels $K_t(x, dy)$.

- However, for what follows, it will be useful to consider the abstract study of Markov processes through Markov semigroups.

Semigroups

Definition

A family of linear operators $(P_t)_{t\geq 0}$ on a Banach space $(\mathcal{B}, ||\cdot||)$ is called a one-parameter semigroup if the following conditions are satisfied (Guionnet and Zegarlinksi, 2004):

1. $P_0 = \mathrm{id}$;
2. $P_t P_s = P_{s+t}$ for any $s, t \geq 0$.
3. The mapping $t \mapsto P_t$ is continuous in the sense that for all $f \in \mathcal{B}$, $t \mapsto P_t f$ is continuous.

Remark

For most purposes, we can consider $\mathcal{B}$ to be the set $\mathcal{C}(S)$ of real-valued bounded continuous functions on a Polish space S equipped with the uniform norm (this covers both the continuous and discrete settings).

Markov Semigroups

Definition

A one-parameter semigroup $(P_t)_{t \geq 0}$ is called a Markov semigroup if the following conditions hold for any $t \geq 0$:

1. $P_t \mathbf{1} = \mathbf{1}$;
2. $P_t f \geq 0$ if $f \geq 0$;
3. $||P_t f|| \leq ||f||$.

Feller Semigroups

Definition

A one-parameter semigroup $(P_t)_{t\geq 0}$ is called a Feller semigroup if the following conditions hold for every $f \in C_0(S)$ and any $t \geq 0$:

1. $P_t f \in C_0(S)$;
2. $P_t f \geq 0$ if $f \geq 0$;
3. $||P_t f||_\infty \leq ||f||_\infty$
4. The following limit holds

$$\lim_{h \to 0^+} ||P_h f - f||_\infty = 0$$

Kernel Representation

Definition

Given a (time-homogeneous) Markov process $(X_t)_{t\geq 0}$, the associated semigroup is given by the linear operator:

$$P_t f(x) = \mathbb{E}[f(X_t)|X_0 = x],$$

where $f : S \to \mathbb{R}$ belongs to some suitable set of functions.

Chapman-Kolmogorov

Remark

We note that this is not an arbitrary choice, indeed, under fairly weak conditions, any (abstract) Markov semigroup admits a kernel representation (Bakry et al., 2013). Moreover, this usually can also be expressed in terms of a (probability) density:

$$P_t f(x) = \int_S p_t(x, y) f(y) dm(y).$$

Theorem

As a consequence, we recover the Chapman-Kolmogorov equations:

$$p_{t+s}(x, y) = \int_S p_s(x, z) p_t(z, y) dm(y).$$

This determines the finite-dimensional distributions of $(X_t)_{t \geq 0}$ *providing a link between abstract semigroups and Markov processes (see Kolmogorov extension theorem).*

Infinitesimal Generator

Definition

The infinitesimal generator $\mathcal{L}$ of a semigroup $(P_t)_{t\geq 0}$ is defined by

$$\mathcal{L}f = \lim_{t\to 0} \frac{P_t f - f}{t},$$

for any function f for which the limit makes sense. The domain $\mathcal{D}(\mathcal{L})$ of $\mathcal{L}$ is the set of functions of $\mathcal{B}$ for which the limit makes sense.

Hille-Yoshida Theorem

Theorem

(Guionnet and Zegarlinksi, 2004) A linear operator $\mathcal{L}$ is the infinitesimal generator of a Markov semi-group $(P_t)_{t\geq 0}$ on $\mathcal{B}$ if and only if

- $\mathbf{1} \in \mathcal{D}(\mathcal{L})$ *and* $\mathcal{L}\mathbf{1} = 0$*;*
- $\mathcal{D}(\mathcal{L})$ *is dense in* $\mathcal{B}$*;*
- $\mathcal{L}$ *is closed*
- *For any* $\lambda > 0$*,* $(\lambda \mathrm{id} - \mathcal{L})$ *is invertible and its inverse* $(\lambda \mathrm{id} - \mathcal{L})^{-1}$ *is bounded with*

$$\sup_{||f||\leq 1} \Big| |(\lambda \mathrm{id} - \mathcal{L})^{-1} f \Big| | \leq \frac{1}{\lambda},$$

  and preserves positivity - i.e. for all $f \geq 0$*,* $(\lambda \mathrm{id} - \mathcal{L})^{-1} f \geq 0$*.*

Kolmogorov Equations

Remark

One of the confusing parts of trying to understand the forward and backward equations presented in discrete diffusion papers is that they are shown on densities, whereas the wider literature on Markov processes presents this at the level of semigroups.

Time-Homogeneous Backward Equation

Theorem

The backward equation for a time-homogeneous Markov process is given by:

$$\begin{aligned}\partial_t P_t f &= \lim_{h\to 0} \frac{P_{t+h} f - P_t f}{h} \\ &= \lim_{h\to 0} \left(\frac{P_h - \mathrm{id}}{h}\right) P_t f \\ &= \mathcal{L} P_t f.\end{aligned}$$

Remark

When S is discrete, P_t and $\mathcal{L}$ can be expressed as a matrix (as f can be represented as a column vector) and so the limits here are well defined. In the continuous case, we can use the density representation of P_t to justify the above limits.

Time-Homogeneous Backward Equation

Theorem

At the level of densities, we have

$$\partial_t p_t(x, y) = \mathcal{L}_x p_t(x, y),$$

where $\mathcal{L}_x$ is the generator $\mathcal{L}$ acting on the x argument (considering the y argument to be fixed).

Time-Homogeneous Backward Equation

Proof

For the first claim, from the backward equation, we have

$$\begin{aligned}
\mathcal{L}(P_t f)(x) &= \lim_{h\to 0} \frac{1}{h}(P_h(P_t f)(x) - P_t f(x)) \\
&= \lim_{h\to 0} \frac{1}{h}\left(\int p_h(x,z) P_t f(z) dz - P_t f(x)\right) \\
&= \lim_{h\to 0} \frac{1}{h}\left(\int p_h(x,z)\left(\int p_t(z,y) f(y) dy\right) dz - \int p_t(x,y) f(y) dy\right) \\
&= \int \left(\lim_{h\to 0} \frac{\int p_h(x,z) p_t(z,y) dz - p_t(x,y)}{h}\right) f(y) dy
\end{aligned}$$

Time-Homogeneous Backward Equation

Proof

$$\implies \mathcal{L}(P_t f)(x) = \int \mathcal{L}_x p_t(x,y) f(y) dy$$

We also have $\partial_t P_t f(x) = \int \partial_t p_t(x,y) f(y) dy$ so by the backward equation:

$$\int \left(\partial_t p_t(x,y) - \mathcal{L}_x p_t(x,y)\right) f(y) dy = 0$$

from which we can conclude as f belongs in a richer enough class of functions.

Time-Homogeneous Backward Equation

Remark

This looks different than the usual presentation of the Kolmogorov backward equation which is usually presented for the time-inhomogeneous case.

Adjoint Operator

Definition

Given some reference measure m on S, we can form the standard inner product of functions in $L^2(S, m)$ by

$$\langle f, g \rangle = \int_S f(x)g(x)dm(x).$$

We define the set $\mathcal{D}(\mathcal{L}*)$ as the space of functions $g \in L^2(S, \mu)$ where for a suitable set of $f \in \mathcal{D}(\mathcal{L})$ (Ren et al., 2025), there exists a unique function $\mathcal{L}^* g \in L^2(S, m)$, for which the following relationship holds

$$\langle \mathcal{L} f, g \rangle = \langle f, \mathcal{L}^* g \rangle$$

We refer to $\mathcal{L}^*$ as the adjoint (linear) operator to $\mathcal{L}$.

Adjoint Operator

Example

If S is a discrete set, we can take m to be the counting measure on S so that $f \in L^2(S, m)$ are represented by finite vectors of length $|S|$. In this case, the inner product reduces to the standard Euclidean inner product.

- In this setting, $\mathcal{L}$ can be represented by a matrix. This allows us to compute the adjoint as simply just $\mathcal{L}^* = \mathcal{L}^\top$.

Time-Homogeneous Forward Equation

Theorem

The forward equation for a time-homogeneous Markov process is given by:

$$\begin{aligned}\partial_t P_t f &= \lim_{h\to 0} \frac{P_{t+h} f - P_t f}{h} \\ &= \lim_{h\to 0} P_t \left(\frac{P_h f - f}{h} \right) \\ &= P_t \mathcal{L} f.\end{aligned}$$

Time-Homogeneous Forward Equation

Theorem

At the level of densities, we have

$$\partial_t p_t(x, y) = \mathcal{L}_y^* p_t(x, y),$$

where $\mathcal{L}_y^*$ is the adjoint to $\mathcal{L}$ acting on the y argument (considering the y argument to be fixed). Moreover, this implies $\partial_t p_t(x) = \mathcal{L}^* p_t(x)$.

Time-Homogeneous Forward Equation

Proof

From the forward equation, we have

$$\begin{aligned}
P_t\mathcal{L}f(x) &= \int p_t(x,y)\mathcal{L}f(y)dy \\
&= \langle p_t(x,\cdot), \mathcal{L}f(\cdot)\rangle \\
&= \langle \mathcal{L}^* p_t(x,\cdot), f(\cdot)\rangle \\
&= \int \mathcal{L}_y^* p_t(x,y)f(y)dy
\end{aligned}$$

We also have $\partial_t P_t f(x) = \int \partial_t p_t(x,y)f(y)dy$ so by the backward equation:

$$\int \Big(\partial_t p_t(x,y) - \mathcal{L}_y^* p_t(x,y)\Big) f(y)dy = 0$$

from which we can conclude as f belongs in a richer enough class of functions.

Time-Homogeneous Forward Equation

Proof

Let $p_t(x) = \mathbb{P}(X_t = x)$, then by the linearity of $\mathcal{L}_y^*$, we have conclude by marginalisation that

$$\partial_t p_t(x) = \mathcal{L}^* p_t(x),$$

remembering that $p_t(x, y) = \mathbb{P}(X_t = y | X_0 = y)$ (somewhat bad choice of notation here).

Application to Time-Homogeneous CTMCs

Example

Recall the setting when $S = [N]$ is discrete, then P_t is represented by a matrix. Then from taking $f(x) = \delta_i(x)$ and its column vector representation, we can see that P_t is formed of transition probabilities:

$$P_t = (p_{ij}(t))_{i,j \in S} \quad \text{where} \quad p_{ij}(t) = \mathbb{P}(X_t = j | X_0 = i).$$

- Moreover, the forward and backward equations (on semigroups) allow us to conclude:

  $$P_t = \exp(t\mathcal{L})$$

  (where $\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}$ is the matrix exponential) from standard ODE theory (applied column-wise to the matrix equation $\partial_t P_t = QP_t$ and $P_0 = \mathrm{id}$).

Application to Time-Homogeneous CTMCs

Example

We will see later how the form of the Q matrix pops out from the definition of the CTMC's dynamics later when we look at the more general time-inhomogeneous case.

Remark

- The objects presented above are not necessarily causally related; they can be defined from a range of starting points - i.e. starting from Markovian transition kernel to define semigroups instead of starting with abstract semigroups.
- In addition, there tends to be lots of technical conditions to these results which vary from different sources; we have omitted over some of these to transmit the essential ideas.
- Of course, this does make learning about the subject more confusing.

Time-Inhomogeneous CTMCs

Definition

A time-inhomogeneous CTMC is a (càdlàg) time-inhomogeneous Markov process $(X_t)_{t\geq 0}$ defined by the dynamics (Del Moral and Penev, 2017):

1. For all $T_n \leq s < T_{n+1}$, we set $X_s := X_{T_n}$ where

$$T_{n+1} := \inf\left\{ t \geq T_n : \int_{T_n}^{t} \lambda_s(X_s) \geq \mathsf{Exp}(1) \right\};$$

2. At time T_{n+1}, we sample the next state by $X_{T_{n+1}} \sim K_{T_{n+1}}(X_{T_{n+1}-}, dx)$,

with the initial conditions $X_0 = x_0$ and $T_0 = 0$.

- Furthermore, we have $\lambda_t(x) \geq 0$, $K_t(x, x) = 0$ and $t \mapsto \lambda_t(x), t \mapsto K_t(f)(x)$ are Lipschitz continuous for any x and for any (suitable) function f.

Time-Inhomogeneous CTMCs

Remark

We recover the time-homogeneous case by choosing $\lambda_t(i) = -Q_{ii}$ and $K_t(i, j) = -Q_{ij}/Q_{ii}$.

Time-Inhomogeneous CTMCs

Theorem

The transition kernels have the form:

$$\mathbb{P}(T_{n+1} \in dt, X_{T_{n+1}} \in dy | T_n = s, X_{T_n} = x) = \lambda_t(x) \exp\left(-\int_s^t \lambda_r(x) dr\right) dt \quad K_t(x, dy).$$

Time-Inhomogeneous CTMCs

Proof

Define $F(t) = \int_{T_n}^{t} \lambda_r(X_{T_n})dr$ for $t \geq T_n$. We note that by assumption $F(t)$ is monotonically increasing and hence is invertible.

$$\begin{aligned}
\mathbb{P}(T_{n+1} \leq t | T_n) = 1 - \mathbb{P}(T_{n+1} > t | T_n) &= 1 - \mathbb{P}(F^{-1}(\mathsf{Exp}(1)) > t) \\
&= 1 - \mathbb{P}(\mathsf{Exp}(1) > F(t)) \\
&= 1 - \exp\left(-\int_{T_n}^{t} \lambda_r(X_{T_n})dr\right).
\end{aligned}$$

To recover the density, we just take the derivative.

Time-Inhomogeneous CTMCs

Theorem

Moreover, when only conditioning on $X_s = x$, the distribution of the next jump time $T^{(s)}$ is given by

$$\mathbb{P}(T^{(s)} \in dt | X_s = x) = \lambda_t(x) \exp\left(-\int_s^t \lambda_r(x) dr\right) dt.$$

This is a consequence of the memoryless property of the T_n distributions.

Time-Evolution Operators

Definition

A family of linear operators $(P_{s,t})_{s \leq t, t \in \mathbb{T}}$ on a Banach space $(\mathcal{B}, ||\cdot||)$ is called a time-evolution operator family, or simply an evolution system, if the following conditions hold (Ren et al., 2025):

1. $P_{t,t} = \mathrm{id}$ for any $t \in \mathbb{T}$;
2. $P_{s,t} = P_{s,r} P_{r,t}$ for any $s \leq r \leq t$ with $s, r, t \in \mathbb{T}$.

Remark

Ren et al. (2025) takes the state space $(S, \mathcal{B}(S), \mu)$ where S is locally compact, separable Hausdorff with Radon measure μ, and bounded Borel measurable functions for the space $\mathcal{B}$ with the supremum norm $||\cdot||_\infty$.

Time-Evolution Operators

Remark

Similarly, the associated evolution system for a Markov process $(X_t)_{t\in\mathbb{T}}$ is

$$P_{s,t}f(x) = \mathbb{E}[f(X_t)|X_s = x].$$

This possesses the further properties (could be taken as definitions in an abstract treatment) which hold for any $s \leq t, s, t \in \mathbb{T}$:

1. $P_{s,t}\mathbf{1} = \mathbf{1}$;
2. $P_{s,t}f \geq 0$ if $f \geq 0$;
3. $||P_{s,t}f|| \leq ||f||$;

Time-Evolution Operators

Note

We stress the direction of the evolution property is $P_{s,t}f = P_{s,r}P_{r,t}f$ and not $P_{s,t}f = P_{r,t}P_{s,t}f$ which might be expected from the intuition that we should apply the transition for the first time interval $[s,r]$ to the function f first.

- In the setting where $S = [N]$ is discrete and so $P_{s,t}$ is represented by the matrix $(\mathbb{P}(X_t = j | X_s = i))_{ij}$, we see the direction makes sense by

$$\begin{aligned}
(P_{s,r}P_{r,t})_{ij} &= \sum_{k \in S} (P_{s,r})_{ik}(P_{r,t})_{kj} \\
&= \sum_{k \in S} \mathbb{P}(X_r = k | X_s = i)\mathbb{P}(X_t = j | X_r = k) \\
&= \mathbb{P}(X_t = j | X_s = i) = (P_{s,t})_{ij},
\end{aligned}$$

from using Chapman-Kolmogorov.

Feller Evolution System

Definition

An evolution system $(P_{s,t})_{s\leq t, s,t\in\mathbb{T}}$ is called a Feller evolution system if for any $f \in C_0(S)$ and for any $s \leq t$, $s, t \in \mathbb{T}$, the following holds:

1. $P_{s,t}f \in C_0(S)$;
2. $P_{s,t}f \geq 0$ if $f \geq 0$;
3. $||P_{s,t}f||_\infty \leq ||f||_\infty$;
4. The following limit holds:

$$\lim_{(\sigma,\tau)\to(s,t)} ||P_{\sigma,\tau}f - P_{s,t}f||_\infty = 0.$$

Infinitesimal Generator

Definition

The right generator of the evolution system is defined by the limit

$$\mathcal{L}_t f = \lim_{h \to 0+} \frac{P_{t,t+h} f - f}{h},$$

with domain $\mathcal{D}(\mathcal{L}_t)$. Similarly, the left generator is defined as

$$\mathcal{L}_t^- f = \lim_{h \to 0+} \frac{P_{t-h,t} f - f}{h},$$

with corresponding domain $\mathcal{D}(\mathcal{L}_t^-)$.

Remark

We note that $\partial_s P_{s,t} f|_{s=t} = -\mathcal{L}_t^-$.

Infinitesimal Generator

Remark

In the case of a time-homogeneous evolution system - i.e. $P_{s,t} = P_{t-s}$ where $(P_t)_{t \geq 0}$ is a semigroup, we have that the right and left generators coincide.

- In the time-inhomogeneous case, they may not coincide, however, Böttcher (2014) provides conditions for when they do (e.g. when the operator continuously depends on time and has bounded coefficients).

Time-Inhomogeneous CTMCs

Theorem

The time-evolution operator for a time-inhomogeneous CTMC is given by

$$P_{s,t}f(x) = f(x)e^{-\int_s^t \lambda_r(x)dr} + \int_s^t \lambda_r(x)e^{-\int_s^r \lambda_u(x)du} K_r P_{r,t} f(x) dr,$$

where $K_t f(x) = \int K_t(x, dy) f(y)$.

Time-Inhomogeneous CTMCs

Proof

Let $T_{s,1}$ be the first jump from the time s. Then

$$\begin{aligned} P_{s,t}f(x) &= \mathbb{E}[f(X_t)|X_s = x] \\ &= \mathbb{E}[f(X_t)\mathbf{1}_{T_{s,1}>t}|X_s = x] + \mathbb{E}[\mathbb{E}[f(X_t)|T_{s,1} = r, X_r = y]\mathbf{1}_{T_{s,1}\leq t}|X_s = x]], \end{aligned}$$

where

$$\mathbb{E}[f(X_t)\mathbf{1}_{T_{s,1}>t}|X_s = x] = f(x)e^{-\int_s^t \lambda_r(x)dr},$$

and

$$\mathbb{E}[f(X_t)|T_{s,1} = r, X_r = y] = P_{r,t}f(y).$$

We apply the form of the transition kernel to conclude.

Time-Inhomogeneous Backward Equation

Theorem

The backward equation for a time-inhomogeneous Markov process is given by:

$$\begin{aligned}
\partial_s P_{s,t} f &= \lim_{h \to 0^+} \frac{P_{s+h,t}(f) - P_{s,t} f}{h} \\
&= \lim_{h \to 0^+} \frac{P_{s+h,t} f - P_{s,s+h} P_{s+h,t} f}{h} \\
&= -\lim_{h \to 0^+} \left(\frac{P_{s,s+h} - \mathrm{id}}{h} \right) P_{s+h,t} f \\
&= -\mathcal{L}_s P_{s,t} f.
\end{aligned}$$

Remark

This is the more familiar version of the backward equation. Note that we have the implicit boundary condition that for $s = t$, $P_{s,t} f(x) = f(x)$. This equation is usually used for understanding the evolution of statistics of the stochastic process.

Time-Inhomogeneous Backward Equation

Remark

We note the previous results only considers the right derivative of $P_{s,t}$. By looking at the left derivative, we have the following expression in terms of the left generator:

$$\begin{aligned}\partial_s P_{s,t} f &= \lim_{h \to 0^+} \frac{P_{s-h,t}(f) - P_{s,t} f}{-h} \\ &= -\lim_{h \to 0^+} \frac{P_{s-h,s} P_{s,t} f - P_{s,t} f}{h} \\ &= -\lim_{h \to 0^+} \left(\frac{P_{s-h,s} - \mathrm{id}}{h} \right) P_{s,t} f \\ &= -\mathcal{L}_s^- P_{s,t} f.\end{aligned}$$

For the time-inhomogeneous CTMC defined previously, we also have $\mathcal{L}_s = \mathcal{L}_s^-$.

Time-Inhomogeneous Backward Equation

Theorem

At the level of densities, where we define $p_{s,t}(x,y) = \mathbb{P}(X_t = y | X_s = x)$, we have

$$\partial_s p_{s,t}(x,y) = -\mathcal{L}_{s,x} p_{s,t}(x,y),$$

where $\mathcal{L}_{s,x}$ is the generator $\mathcal{L}_s$ acting on the x argument (considering the y argument to be fixed).

Time-Inhomogeneous Backward Equation

Proof

The proof proceeds in the same manner as in the time-homogeneous case. To simplify the proof, apply the definition of $\mathcal{L}_s^-$ instead of $\mathcal{L}_s$ (we note that for most practical cases, the two operators will coincide).

Time-Inhomogeneous Forward Equation

Theorem

The forward equation for a time-inhomogeneous Markov process is given by

$$\begin{aligned}\partial_t P_{s,t} f &= \lim_{h \to 0^+} \frac{P_{s,t+h} f - P_{s,t} f}{h} \\ &= \lim_{h \to 0^+} \frac{P_{s,t} P_{t,t+h} f - P_{s,t} f}{h} \\ &= \lim_{h \to 0^+} P_{s,t} \left(\frac{P_{t,t+h} f - f}{h} \right) \\ &= P_{s,t} \mathcal{L}_t f.\end{aligned}$$

Time-Inhomogeneous Forward Equation

Remark

We note the previous results only considers the right derivative of $P_{s,t}$. By looking at the left derivative, we have the following expression in terms of the left generator:

$$\begin{aligned}
\partial_t P_{s,t} f &= \lim_{h \to 0+} \frac{P_{s,t-h} f - P_{s,t} f}{-h} \\
&= \lim_{h \to 0^+} \frac{P_{s,t-h} f - P_{s,t-h} P_{t-h,t} f}{-h} \\
&= \lim_{h \to 0^+} P_{s,t-h} \left(\frac{P_{t-h,t} f - f}{h} \right) \\
&= P_{s,t} \mathcal{L}_t^- f.
\end{aligned}$$

For the time-inhomogeneous CTMC defined previously, we also have $\mathcal{L}_t = \mathcal{L}_t^-$.

Time-Inhomogeneous Forward Equation

Theorem

At the level of densities, where we define $p_{s,t}(x,y) = \mathbb{P}(X_t = y | X_s = x)$*, we have*

$$\partial_t p_{s,t}(x,y) = \mathcal{L}^*_{t,y} p_{s,t}(x,y),$$

where $\mathcal{L}^*_{t,x}$ *is the adjoint to the generator* $\mathcal{L}_t$ *acting on the* y *argument (considering the* x *argument to be fixed). Moreover, this implies that* $\partial_t p_t(x) = \mathcal{L}^*_t p_t(x)$*.*

Time-Inhomogeneous Forward Equation

Proof

The proof proceeds in the same manner as in the time-homogeneous case, where we employ the notion of the adjoint to separate out f to conclude. For the second claim, we use the fact that $\mathcal{L}^*_{t,y}$ is linear and marginalisation to conclude.

- An alterative approach is to fix some constant $T > 0$ and define $u(x, s) = P_{s,T}f(x)$. By the Markov property, we also have the expression $u(x, s) = P_{s,t}P_{t,T}f(x) = P_{s,t}u(t, x)$ for $s \leq t \leq T$. Then by considering $\partial_t u(x, s)$, we have

  $$0 = \partial_t u(x, s) = \int \partial_t p_{s,t}(x, y)u(y, t) + p_{s,t}(x, y)\partial_t u(y, t)dy,$$

  where we can use the backward equation and the definition of the adjoint to conclude.

Generator for Time-Inhomogeneous CTMCs

Theorem

Returning to the definition of a time-inhomogeneous CTMC $(X_t)_{t\geq 0}$, the generator has the following form

$$\mathcal{L}_t f(x) = \mathcal{L}_t^- f(x) = \lambda_t(x) \int [f(y) - f(x)] K_t(x, dy).$$

Proof

Differentiate $P_{s,t}$ and use the Leibniz integral rule to handle the integral.

Generator for Time-Inhomogeneous CTMCs

Example

Under the discrete setting, $\mathcal{L}_t$ can be represented by a matrix $(\mathcal{L}_t)_{ij}$ and $f : S \to \mathbb{R}$ can be represented as a vector $f = (f_1, \ldots, f_{|S|})^\top \in \mathbb{R}^{|S|}$.

- To see how the form of the Q-matrix arises, we look at the matrix representation of $\mathcal{L}_t$:

$$\begin{aligned}\mathcal{L}_t f(i) = (\mathcal{L}_t f)_i = \sum_{j \in S} (\mathcal{L}_t)_{ij} f_j = \lambda_t(i) \sum_{j \in S} [f_j - f_i] K_t(i,j) \\ = \sum_{j \neq i} [\lambda_t(i) K_t(i,j) f_j] - \lambda_t(i) f_i,\end{aligned}$$

  remembering that $K_t(i,i) = 0$ by definition. Hence, from matching coefficients, we can conclude $\mathcal{L}_t$ has the form of Q-matrix:

$$\forall j \neq i, (\mathcal{L}_t)_{ij} = \lambda_t(i) K_t(i,j) \;\text{ and }\; (\mathcal{L}_t)_{ii} = -\lambda_t(i).$$

Generator for Time-Inhomogeneous CTMCs

Example

Let $R_t(i, j)$ represent the rate matrix/generator for a CTMC. The forward equation has a nice interpretation by the following form:

$$\partial_t p_t(i) = \sum_{j \neq i} p_t(j) R_t(j, i) - q_t(i) R_t(i, j),$$

where the rate of change of probability mass is equal to the difference between the rate of probability mass moving into state i and out of state i.

- This intuition matches the form of the continuity equation used in flow matching (due to their related origins (Holderrieth et al., 2024)).

First-Order Approximation

Remark

From a elementary Taylor expansion, we have for $h \geq 0$:

$$\begin{aligned} P_{t,t+h}(f) &= P_{t,t}f + h\partial_t P_{t,s}(f)|_{s=t^+} + O(h^2) \\ &= \mathrm{id}(f) + h\mathcal{L}_t f + O(h^2) \\ \implies P_{t,t+h} &= \mathrm{id} + h\mathcal{L}_t + O(h^2), \end{aligned}$$

so in the CTMC setting, we have the first-order approximation:

$$\mathbb{P}(X_{t+h} = j | X_t = i) = \delta_i(j) + h(\mathcal{L}_t)_{ij} + O(h^2).$$

First-Order Approximation

Remark

Moreover, we can also look at the left limits to give

$$\begin{aligned} P_{t-h,t}(f) &= P_{t,t}f - h\partial_s P_{s,t}(f)|_{s=t^-} + O(h^2) \\ &= \mathrm{id}(f) + h\mathcal{L}_t^- f + O(h^2) \\ \implies P_{t-h,t} &= \mathrm{id} + h\mathcal{L}_t^- + O(h^2), \end{aligned}$$

so in the CTMC setting, we have the first-order approximation:

$$\mathbb{P}(X_t = j | X_{t-h} = i) = \delta_i(j) + h(\mathcal{L}_t)_{ij} + O(h^2).$$

A confusing aspect about some discrete diffusion papers is that they use the rate matrix to approximate transition probabilities in different directions (e.g. Campbell et al. (2022)).

Time Conversion

Definition

(Wentzell, 2022) Let $(X_t)_{t\in\mathbb{T}}$ be a Markov process governed by the evolution system $(P_{s,t})_{s\leq t,s,t\in\mathbb{T}}$ with a right generator $\mathcal{L}_t$. We define the augmented process $(\tilde{X}_t)_{t\in\mathbb{T}}$ in the augmented probability space $(\tilde{\Omega},\tilde{\mathcal{F}},\tilde{\mathbb{P}})$ by

1. The augmented state $\tilde{x}$ is defined as $\tilde{x} := (s,x) \in \mathbb{T}\times S$ with $\mathbb{T}\times S$ equipped with the σ-algebra:

$$\tilde{\mathcal{B}} := \left\{\tilde{B} \subset \mathbb{T}\times S : \tilde{B}_s := \left\{x \in S : \tilde{x} = (s,x) \in \tilde{B}\right\} \in \mathcal{B}(S), \forall s \in \mathbb{T}\right\};$$

2. The augmented event $\tilde{\omega} := (s,\omega) \in \mathbb{T}\times\Omega := \tilde{\Omega}$ on the augmented state space equipped with the σ-algebra:

$$\tilde{\mathcal{F}} := \left\{\tilde{F} \subset \tilde{\Omega} : \tilde{F}_s := \left\{\omega \in \Omega : \tilde{\omega} = (s,\omega) \in \tilde{F}\right\} \in \mathcal{F}, \forall s \in \mathbb{T}\right\};$$

Time Conversion

Definition

3 The augmented process $(\tilde{X}_t)_{t\in\mathbb{T}}$ is defined as

$$\tilde{X}_t(\tilde{\omega}) = \tilde{X}_t(s,\omega) := (s+t, X_{s+t}(\omega)),$$

and the augmented probability measure is defined so that the following holds:

$$\begin{aligned}\tilde{\mathbb{P}}\left(\tilde{X}_t(\tilde{\omega}) \in \tilde{B} | \tilde{X}_0(\tilde{\omega}) = \tilde{x}\right) &= \tilde{P}\left((s+t, X_{s+t}(\omega)) \in \tilde{B} | (s, X_s(\omega)) = (s,x)\right) \\ &:= \mathbb{P}(X_{s+t}(\omega) \in \tilde{B}_{s+t} | X_s(\omega) = x),\end{aligned}$$

for any $\tilde{x} = (s,x) \in \mathbb{T} \times S, t \geq 0$ and $\tilde{B} \in \tilde{\mathcal{B}}$.

Time Conversion

Remark

We have the following relationships between the time-inhomogeneous process $(X_t)_{t\in\mathbb{T}}$ and the augmented process:

- If $(X_t)_{t\in\mathbb{T}}$ is Markov then $(\tilde{X}_t)_{t\geq 0}$ is Markov;
- We have the relationship $\tilde{\mathcal{L}}f(\tilde{x}) = \partial_s f(s,x) + \mathcal{L}_s f(s,x)$ where $\tilde{\mathcal{L}}$ is the generator of $(\tilde{X}_t)_{t\geq 0}$;
- The Feller property carries over in both directions.

Dynkin's Formula

Theorem

Under the assumptions presented in (Ren et al., 2025), the following relation holds for a time-inhomogeneous processes:

$$\mathbb{E}\left[f_t(X_t)|X_0 = x\right] = f_0(x) + \mathbb{E}\left[\int_0^t \partial_s f_s(X_s) + \mathcal{L}_s f_s(X_s) ds|X_0 = x\right].$$

Remark

This result is used in order to derive the ELBO in Campbell et al. (2022), specifically to rewrite the

Path Measure

Definition

For a time-inhomogeneous CTMC $(X_t)_{t\in[0,T]}$ with rate matrix R_t, consider the path space view given by the trajectories $W : \omega \mapsto (t \in [0,T] \mapsto X_t(\omega))$ of $(X_t)_{t\in[0,T]}$.

- Each trajectory of W can be described in terms of an initial point W_0, a sequence of jump times and new states $\{(T_i, W_{T_i})\}_{i=1}^n$ and the condition $T_{n+1} \geq T$.
- We define the path measure as the induced probability measure assigned to trajectories by $(X_t)_{t\in[0,T]}$:

$$\mathbb{P}(W \in d\omega) = \mathbb{P}(W_0 \in d\omega_0, (T_1, W_{T_1}) \in d(t_1, \omega_1), \ldots, (T_n, W_{T_n}) \in d(t_n, \omega_n), T_{n+1} \geq T)$$

Path Measure

Theorem

The path measure has the form:

$$\mathbb{P}(W \in d\omega) = p_0(W_0) \exp\left(-\int_0^T R_s(W_s^-, W_s^-) ds\right) \prod_{s: W_s \neq W_s^-} R_s(W_s^-, W_s),$$

where $W_s^- := \lim_{t \to s^-} W_t$ *is the left limit of* W_s*. Moreover, the Radon-Nikodym derivative between the path measures of two CTMCs easily follows.*

Remark

This result is used in order to derive the ELBO in Campbell et al. (2022).

Path Measure

Proof

This is a simple consequence of the Markov property and the transition probabilities provided previously.

Historical Development

Timeline

Discrete diffusion has followed the progression:

- Discrete time Markov processes (Austin et al., 2021)
- Continuous time Markov processes (Campbell et al., 2022)
- "Score"-based formulation (Lou et al., 2023)
- Simplification of the objective for masking discrete diffusion (Shi et al., 2024; Sahoo et al., 2024)
- Flow matching formulation (Campbell et al., 2024; Gat et al., 2024)

Austin et al. (2021)

Recall the ELBO objective for DDPM (Ho et al., 2020):

$$\begin{aligned} L_{\mathsf{vb}}(x_0) =& D_{KL}(q(x_T|x_0)) + \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[D_{KL}(q(x_{t-1}|x_t, x_0)|p_\theta(x_{t-1}|x_t)\right] \\ & - \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)]. \end{aligned}$$

We see that this decomposition works for any Markov forward noising process $q(x_{1:T}|x_0) = \prod q(x_t|x_{t-1})$ and Markov reverse process $p_\theta(x_{0:T}) = p(x_T) \prod p_\theta(x_{t-1}|x_t)$.

Austin et al. (2021)

Therefore to generalise this to discrete objects, we can just take (where x_t are one-hot row vectors from representing elements from some discrete set S):

$$q(x_t|x_{t-1}) = \mathsf{Cat}(x_t; x_{t-t}Q_t),$$

and all quantities of interest such as $q_t(x_t|x_0)$ and $q(x_{t-1}|x_t, x_0)$ are tractable, and the KL can easily be computed here.

Austin et al. (2021)

Austin et al. (2021) explores different noising processes for $q(x_t|x_0) = \text{Cat}(x_t|x_0\bar{Q}_t)$ where $\bar{Q}_t = Q_1 Q_2 \ldots Q_t$ which have tractble stationary distributions:

- Uniform
- Masking
- Discretized Gaussian
- Token embedding distance

Campbell et al. (2022)

We have standard issues with fixed schedule generative models:

- We are forced to pre-specify the noise discretisation we want to train on;
- We are limited to simple ancestral sampling/inflexible sampling;
- From the continuous case, we see that working in continuous time can bring a lot of benefits.

Campbell et al. (2022)

Campbell et al. (2022) proposes defining the forward process as a time-inhomogeneous CTMC $(X_t)_{t\in[0,T]}$ with rate matrix R_t and marginal distributions q_t.

- The corresponding time-reversal $(Y_t)_{t\in[0,T]} = (X_{T-t})_{t\in[0,T]}$ is also a CTMC but with the rate matrix $\hat{R}_t$:

  $$\hat{R}_t(x,\tilde{x}) = R_t(\tilde{x},x)\frac{q_t(\tilde{x})}{q_t(x)} = R_t(\tilde{x},x)\sum_{x_0}\frac{q_{t|0}(\tilde{x}|x_0)}{q_{t|0}(x|x_0)}q_{0|t}(x_0|x) \text{ for } x \neq \tilde{x},$$

  This follows from the forward/backward equations and Bayes' rule.
- The reverse process is parametrised in terms of $p_\theta(x_0|x)$ being substituted above.

Remark

Note that the convention used in Campbell et al. (2022) is that $\hat{R}_t$ denotes the rate of $Y_{T-t} = X_t$, so the forward equation of the backward process is $\partial_t q_{T-t} = \hat{R}_t^\top q_{T-t}$.

Campbell et al. (2022)

Remark

Note that Campbell et al. (2022) uses the approximation

$$q_{t|t-\Delta t}(x|\tilde{x}) = \delta_{x,\tilde{x}} + R_t(\tilde{x}, x)\Delta t + O(\Delta t^2).$$

We recall from before that the same rate matrix R_t can be used in the approximation of $q_{t+\Delta t|t}(x|\tilde{x})$ as well - i.e. R_t is not biased in any direction as the left and right generators coincide.

Campbell et al. (2022)

Theorem

The negative ELBO has the form:

$$\mathcal{L}(\theta) = T\mathbb{E}_{t\in\mathcal{U}(0,T),q_t(x),r_t(\tilde{x}|x)}\left[\left\{\sum_{x'\neq x}\hat{R}_t^\theta(x,x')\right\} - \mathcal{Z}^t(x)\log\left(\hat{R}_t^\theta(\tilde{x},x)\right)\right] + C,$$

where $\mathcal{Z}^t(x) = \sum_{x'\neq x} R_t(x,x')$ *and* $r_t(\tilde{x}|x) = (1-\delta_{\tilde{x},x})R_t(x,\tilde{x})\mathcal{Z}^t(x)$.

Remark

We see the correspondence: $\mathcal{Z}^x = \lambda_t(x)$ and $r_t(\tilde{x}|x) = K_t(x,\tilde{x})$ using the previous notation.

Campbell et al. (2022)

To avoid evaluating the neural network twice for each gradient step at $x, \tilde{x}$, Campbell et al. (2022) proposes the following approximation to the objective:

$$\mathcal{L}^{\mathsf{approx}}(\theta) = T\mathbb{E}_{t\in\mathcal{U}(0,T),q_t(x),r_t(\tilde{x}|x)}\left[\left\{\sum_{x'\neq\tilde{x}}\hat{R}_t^{\theta}(\tilde{x},x')\right\} - \mathcal{Z}^t(x)\log\left(\hat{R}_t^{\theta}(\tilde{x},x)\right)\right] + C,$$

as the distributions of x and $\tilde{x}$ are approximately the same.

Campbell et al. (2022)

For D dimensional data $X^{1:D} \in S^D$ - i.e. sequences with D tokens - instead of considering the total possible $|S|^D$ transitions, Campbell et al. (2022) considers a sparse rate matrix to ensure transitions only ever involves a change in exactly one token:

- This results in choosing a forward process which factorises as $q_{t|s}(x_t^{1:D}|x_s^{1:D}) = \prod_{d=1}^{D} q_{t|s}(x_t^d|x_s^d)$;
- The forward and reverse rates are of the form:

$$R_t^{1:D}(\tilde{x}^{1:D}, x^{1:D}) = \sum_{d=1}^{D} R_t^d(\tilde{x}^d, x^d)\delta_{x^{1:D\backslash d}, \tilde{x}^{1:D\backslash d}}$$

$$\hat{R}_t(x^{1:D}, x^{1:D}) = \sum_{d=1}^{D} R_t^d(\tilde{x}^d, x^d)\delta_{x^{1:D\backslash d}, \tilde{x}^{1:D\backslash d}} \sum_{x_0^d} q_{0|t}(x_0^d|x^{1:D}) \frac{q_{t|0}(\tilde{x}^d|x_0^d)}{q_{t|0}(x^d|x_0^d)}$$

Campbell et al. (2022)

Further introduces the sampling schemes:

- τ-leaping to approximate CTMC trajectories
- Predictor-corrector scheme

Lou et al. (2023)

Lou et al. (2023) proposes learning the score function $s(x, t)$ defined as

$$s(x,t) = \left(\frac{p_t(y)}{p_t(x)}\right)_{y \in S},$$

instead of the density $p_{0|t}$.

- Recall the form of the time-reversal rate matrix which is a function of $s(x, t)$.x‘
- We can view this as a generalisation of the score to discrete spaces by considering the “concrete” score:

$$\frac{\Delta p_t(x)}{p_t(x)} = \left(\frac{p_t(y) - p_t(x)}{p_t(x)}\right)_{y \in S}.$$

Remark

Confusingly, the paper works with $\partial_t p_t = Q_t p_t$ as the forward equation for the forward process, so that the (forward) rate matrix is $Q_t^\top$.

Lou et al. (2023)

Definition

The denoising score entropy loss is given by

$$\mathcal{L}_{\mathsf{DSE}}(\theta) = \mathbb{E}_{x_0, x_t}\left[\sum_{y \neq x_t} w_{x_t,y}\left(s_\theta(x_t)_y - \frac{p_t(y|x_0)}{p_t(x_t|x_0)} \log s_\theta(x_t)_y\right)\right],$$

which has the minimiser $s_\theta = s(x,t)$.

Remark

This can be shown as equivalent (up to a constant) to the score entropy loss:

$$\mathbb{E}_{x_t}\left[\sum_{y \neq x_t} w_{x_t,y}\left(s_\theta(x_t)_y - \frac{p_t(y)}{p_t(x_t)} \log s_\theta(x_t)_y\right)\right].$$

with a similar trick as the score matching loss from using $p_t(x) = \sum_{x_0} p(x_t|x_0)p(x_0)$.

Shi et al. (2024); Sahoo et al. (2024)

Shi et al. (2024) considers a masking discrete diffusion on the state space $S = [m] = \{0, 1, \ldots, m\}$ where m denotes the special mask token.

- The forward process is given by (where x are one-hot column vector in this notation):

$$q(x_t|x_0) = \mathsf{Cat}(x_t; \bar{Q}(t)^\top x_0) \quad \text{where} \quad \bar{Q}(t) = \alpha_t I + (1 - \alpha_t)\mathbf{1}e_m^\top,$$

  and e_m denotes the vector of unit entry in the mth position and zeros elsewhere.
- Essentially, this encodes the process where x_t remains at x_0 with probability α_t or transitions to m with probability $1 - \alpha_t$; and when $x_t = m$, the process stays at the masked state.
- We take α_t to monotonically decrease from 1 to 0.

Shi et al. (2024); Sahoo et al. (2024)

Remark

This amounts to the choice of generator:

$$Q(t) = \beta(t)(\mathbf{1}e_m^\top - I) = \beta(t)Q,$$

for some choice of $\beta \geq 0$.

- By the forward equation $\partial_t q_t = Q(t)^\top q_t$, we can use the commutativity of Q - i.e. $Q(t)Q(s) = Q(s)Q(t)$ - to show the solution is given by the matrix exponential:

$$q_t = \exp\left(Q \int_0^t \beta(s)ds\right)^\top q_0 = \bar{Q}(t)^\top q_0 \text{ where } \alpha_t = \int_0^t \beta(s)ds.$$

Shi et al. (2024); Sahoo et al. (2024)

Remark

Due to simplicity of the forward process, we can compute the reverse process CTMC $q(x_s|x_t, x_0)$ conditioned on x_0 in closed form. In particular, the reverse transition matrix (i.e. evolution system) is

$$\bar{R}^{x_0}(t,s) = I + \frac{\alpha_s - \alpha_t}{1-\alpha_t} e_m (x_0 - e_m)^\top.$$

- Moreover, we can compute the generator for the reverse process conditioned on x_0 as

$$R^{x_0}(t) = -\frac{\alpha_t'}{1-\alpha_t} e_m (x_0 - e_m)^\top.$$

Shi et al. (2024); Sahoo et al. (2024)

Theorem

Consider the (mean) parametrisation $p_\theta(x_s|x_t) := q(x_s|x_t, \mu_\theta(x_t, t))$*, where*

$$\mu_\theta(x_t, t) = \begin{cases} \textit{softmax}(f_\theta(x_t, t)) & x_t = m \\ x_t & x_t \neq m, \end{cases}$$

The negative ELBO has the simple form:

$$\mathcal{L}(\theta) = \int_0^1 \frac{\alpha_t'}{1 - \alpha_t} \mathbb{E}_{q(x_t|x_0)} \left[\delta_{x_t, m} \cdot x_0^\top \log \mu_\theta(x_t, t) \right] dt.$$

Shi et al. (2024); Sahoo et al. (2024)

Remark

We note that Shi et al. (2024) also demonstrates how the objectives from Campbell et al. (2022) and Sahoo et al. (2024) can be derived from this objective,

- In particular, the relationship between the score and mean parametrisations:

$$s(m,t)_j = \frac{\alpha_t}{1-\alpha_t}\mathbb{P}(x_0 = j | x_t = m) \text{ which satisfies } \sum_{j \neq m} s(m,t)_j = \frac{\alpha_t}{1-\alpha_t},$$

where $s(x,t)_j = \frac{q_t(j)}{q_t(x)}$ is the score function.

Campbell et al. (2024); Gat et al. (2024)

The standard flow matching recipe also carries over to the discrete case due to the same marginalisation relationship:

- Define some simple conditional path $p_{t|1}(x_t|x_1)$ which satisfies $p_{1|1}(x_t|x_1) = \delta_{x_1}(x_t)$ where $x_1 \sim p_{\text{data}}$ and $p_{0|t}(x_0|x_1) = p_{\text{noise}}(x_0)$
- Given the rate matrix $R_t(x, y|x_1)$ generating a CTMC producing the probability path $p_{t|1}$, the following rate matrix:

$$R_t(x_t, j) := \mathbb{E}_{p_{1|t}(x_1|x_t)}\left[R_t(x_t, j|x_1)\right],$$

  generates the marginal probability path $p_t(x_t) = \mathbb{E}_{p_{\text{data}}}[p_{t|1}(x_t|x_1)]$ bridging p_{data} and p_{noise}.

Remark

We note that the notation of Gat et al. (2024) defines the "probability velocity" u_t as the transpose of the rate matrix of the underlying CTMC.

Campbell et al. (2024); Gat et al. (2024)

Remark

We note that the relationship between the conditional rate matrix and marginal rate matrix also holds for the general case of arbitrary generators (under reasonable assumptions) (Holderrieth et al., 2024):

$$\mathcal{L}_t f(x) = \mathbb{E}_{z \sim p_{1|t}(\cdot|x)} \left[\mathcal{L}_t^z f(x) \right].$$

Campbell et al. (2024); Gat et al. (2024)

In order to learn the marginal rate matrix R_t, Campbell et al. (2024) proposes the cross-entropy objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{p_{\mathsf{data}}(x_1),\mathcal{U}(t;0,1),p_{t|1}(x_t|x_1)}\left[\log p^{\theta}_{1|t}(x_1|x_t)\right].$$

We can then sample from the model from estimating $\mathbb{E}_{p_{1|t}(x_1|x_t)}\left[R_t(x_t, j|x_1)\right]$.

Remark

The ELBO decomposition from Campbell et al. (2022) still holds in this setting, but Campbell et al. (2024) justifies this new objective by decomposing the ELBO further and arguing that only optimising the cross-entropy component is beneficial.

Ren et al. (2025)

Definition

Let $(X_t)_{t\in[0,T]}$ be a Markov process. The associated time-reversal process $(\overleftarrow{X}_t)_{t\in[0,T]}$ is defined by

$$\overleftarrow{X}_t := \lim_{h\to 0^+} X_{T-t-h} = X_{(T-t)^-},$$

and by setting $\overleftarrow{X}_T := X_0$.

Theorem

Under suitable assumptions, we have the following relation for the time-reversal generator $\overleftarrow{\mathcal{L}}_t$:

$$p_t \overleftarrow{\mathcal{L}}_{T-t} f = \mathcal{L}_t^*(p_t f) - f\mathcal{L}_t^* p_t,$$

where p_t denotes the marginal distribution of $(X_t)_{t\in[0,T]}$ at time t.

Ren et al. (2025)

Theorem

Recall the generator for a CTMC can be expressed as $\mathcal{L}_t f(i) = \sum_{j \in S}[f(j) - f(i)]Q_t(x, y)$. We can express the time-reversal generator for a CTMC by

$$\overleftarrow{\mathcal{L}}_{T-t} f(i) = \frac{1}{p_t(i)} \sum_{j \in S} (p_t(j) f(j) Q_t(i, j) - p_t(i) f(i) Q_t(j, i)) \\ - \frac{f(i)}{p_t(i)} \sum_{j \in S} (p_t(j) Q_t(i, j) - p_t(i) Q_t(j, i)) \\ = \sum_{j \in S} [f(j) - f(i)] s_t(i, j) Q_t(i, j),$$

where $s_t(i, j) = \frac{p_t(j)}{p_t(i)}$ is the score function.

Ren et al. (2025)

Remark

- We see the time-reversal must also be a CTMC with the corresponding generator.
- Moreover, Ren et al. (2025) shows that when parametrising the time-reversal generator in a natural way, we can recover the objective from Lou et al. (2023),
- This relies on considering the KL divergence of between the two path measures, where we note the parametrised generator differs from the true time-reversal by some carré du champ operator term.

Note

Similar ideas have been presented in Benton et al. (2024) and Holderrieth et al. (2024).

Further Topics

Further Topics

Other topics regarding discrete diffusion:

- Sampling (Liu et al., 2024b,a; Kim et al., 2025);
- Connection between masking discrete diffusion and masking any-order autoregressive models (Ou et al., 2024);
- Inference scaling and data scaling compared to autoregressive models (Prabhudesai et al., 2025; Swerdlow et al., 2025);
- Guidance (Schiff et al., 2024; He et al., 2025)
- Other forward processes (e.g. uniform) (Sahoo et al., 2025).

Applications

Molecular Generation

Molecular generation relies on modelling discrete atom types and bond types:

- Continuous diffusion with quantisation (Hoogeboom et al., 2022; Schneuing et al., 2024);
- Continuous diffusion on the simplex (Davis et al., 2024; Stark et al., 2024);
- Discrete diffusion (Vignac et al., 2023; Irwin et al., 2024).

Thank you!

leo.zhang@stx.ox.ac.uk

`https://leozhangml.github.io/`

References I

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. (2021). Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Bakry, D., Gentil, I., and Ledoux, M. (2013). *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media.

Benton, J., Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. (2024). From denoising diffusions to denoising markov models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):286–301.

Böttcher, B. (2014). Feller evolution systems: Generators and approximation. *Stochastics and Dynamics*, 14(03):1350025.

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. (2022). A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279.

References II

Campbell, A., Yim, J., Barzilay, R., Rainforth, T., and Jaakkola, T. (2024). Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*.

Davis, O., Kessler, S., Petrache, M., Ceylan, I., Bronstein, M., and Bose, J. (2024). Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084.

Del Moral, P. and Penev, S. (2017). *Stochastic processes: From applications to theory*. Chapman and Hall/CRC.

Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R. T., Synnaeve, G., Adi, Y., and Lipman, Y. (2024). Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385.

Guionnet, A. and Zegarlinksi, B. (2004). Lectures on logarithmic sobolev inequalities. In *Séminaire de probabilités XXXVI*, pages 1–134. Springer.

References III

He, J., Hernández-Lobato, J. M., Du, Y., and Vargas, F. (2025). Rne: a plug-and-play framework for diffusion density estimation and inference-time control. *arXiv preprint arXiv:2506.05668*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Holderrieth, P., Havasi, M., Yim, J., Shaul, N., Gat, I., Jaakkola, T., Karrer, B., Chen, R. T., and Lipman, Y. (2024). Generator matching: Generative modeling with arbitrary markov processes. *arXiv preprint arXiv:2410.20587*.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. (2022). Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR.

References IV

Irwin, R., Tibo, A., Janet, J. P., and Olsson, S. (2024). Semlaflow–efficient 3d molecular generation with latent attention and equivariant flow matching. *arXiv preprint arXiv:2406.07266*.

Kim, J., Shah, K., Kontonis, V., Kakade, S., and Chen, S. (2025). Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *arXiv preprint arXiv:2502.06768*.

Liu, A., Broadrick, O., Niepert, M., and Broeck, G. V. d. (2024a). Discrete copula diffusion. *arXiv preprint arXiv:2410.01949*.

Liu, S., Nam, J., Campbell, A., Stärk, H., Xu, Y., Jaakkola, T., and Gómez-Bombarelli, R. (2024b). Think while you generate: Discrete diffusion with planned denoising. *arXiv preprint arXiv:2410.06264*.

Lou, A., Meng, C., and Ermon, S. (2023). Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*.

References V

Norris, J. R. (1998). *Markov chains*. Number 2. Cambridge university press.

Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., and Li, C. (2024). Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*.

Prabhudesai, M., Wu, M., Zadeh, A., Fragkiadaki, K., and Pathak, D. (2025). Diffusion beats autoregressive in data-constrained settings. *arXiv preprint arXiv:2507.15857*.

Ren, Y., Rotskoff, G. M., and Ying, L. (2025). A unified approach to analysis and design of denoising markov models. *arXiv preprint arXiv:2504.01938*.

Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., and Kuleshov, V. (2024). Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Sahoo, S. S., Deschenaux, J., Gokaslan, A., Wang, G., Chiu, J., and Kuleshov, V. (2025). The diffusion duality. *arXiv preprint arXiv:2506.10892*.

References VI

Schiff, Y., Sahoo, S. S., Phung, H., Wang, G., Boshar, S., Dalla-torre, H., de Almeida, B. P., Rush, A., Pierrot, T., and Kuleshov, V. (2024). Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*.

Schneuing, A., Harris, C., Du, Y., Didi, K., Jamasb, A., Igashov, I., Du, W., Gomes, C., Blundell, T. L., Lio, P., et al. (2024). Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. (2024). Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167.

Stark, H., Jing, B., Wang, C., Corso, G., Berger, B., Barzilay, R., and Jaakkola, T. (2024). Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*.

References VII

Swerdlow, A., Prabhudesai, M., Gandhi, S., Pathak, D., and Fragkiadaki, K. (2025). Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*.

Vignac, C., Osman, N., Toni, L., and Frossard, P. (2023). Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 560–576. Springer.

Wentzell, A. D. (2022). *Theorie zufälliger Prozesse*, volume 50. Walter de Gruyter GmbH & Co KG.