First Hitting Diffusion Models for Generating Manifold, Graph and Categorical Data

Mao Ye, Lemeng Wu, Qiang Liu
Department of Computer Science
The University of Texas at Austin

Abstract

We propose a family of First Hitting Diffusion Models (FHDM), deep generative models that generate data with a diffusion process that terminates at a random first hitting time. This yields an extension of the standard fixed-time diffusion models that terminate at a pre-specified deterministic time. Although standard diffusion models are designed for continuous unconstrained data, FHDM is naturally designed to learn distributions on continuous as well as a range of discrete and structure domains.

1 Introduction

Standard diffusion processes used in ML can be classified into two categories: 1) infinite (or mixing) time diffusion processes such as Langevin dynamics, which requires the process to run sufficiently long to converge to the invariant distribution, whose property is leveraged for the purpose of learning and inference; and 2) fixed time diffusion processes such as DDPM, SMLD, and Schrodinger bridges [De Bortoli et al., 2021], which are designed to output the desirable results at a pre-fixed time. Although fixed-time diffusion has been show to surpass infinite time diffusion on both speed and quality, it still yield slow speed for modern applications due to the need of a pre-specified time and the incapability to adapt the time based on the difficulty of instances and problems. Moreover, standard diffusion models are naturally designed on $\mathbb{R}^d$, and can not work for discrete and structured data without special cares.

In this work, we study and explore a different first hitting time diffusion model that terminates at the first time as it hits a given domain, and leverages the distribution of the exit location (known as exit distribution, or harmonic measure [Oksendal, 2013]) as a tool for learning and inference. We provide the basic framework and tools for first hitting diffusion models. We leverage our framework to develop a general approach for learning deep generative models based on first hitting diffusion. This approach generalizes SMLD and its SDE extensions but can be attractively applied to a range of discrete and structured domains. This contrasts with the standard diffusion models, which are restricted to continuous $\mathbb{R}^d$ data. In particular, we instantiate our framework to three cases, yielding new diffusion models for learning 1) spherical, 2) binary and 3) categorical data.

2 Main Framework

2.1 First Hitting Diffusion Processes

Let $\Pi^*$ be a distribution of interest on a domain $\Omega \subset \mathbb{R}^d$. The goal is to construct a first hitting stochastic process, which starts from a point outside of $\Omega$ and returns a sample drawn from $\Pi^*$ when it first hits set $\Omega$. We start with introducing the new first hitting model.

*Corresponding author. Email: maoye21@utexas.edu

NeurIPS 2022 Workshop on Score-Based Methods.
Let $Z := \{ Z_t : t \in [0, +\infty) \}$ be a continuous-time Markov process with probability law $Q$ taking value in a set $V$ that contains $\Omega$ as a subset. Here $Q$ is a probability measure defined on the space of all continuous trajectories $\mathcal{C}([0, +\infty), \mathbb{R}^d)$. We use $Q_t$ to denote the marginal distribution of $Z_t$ at time $t$. We assume that the process is initialized from a point $Z_0$ outside of $\Omega$. Denote by $\tau$ the first hitting time of $Z_t$ on $\Omega$, that is, $\tau = \inf\{ t \geq 0 : Z_t \in \Omega \}$. We call that $Z_t$ is absorbing to set $\Omega$ if

i) The process enters $\Omega$ in finite time almost surely when initialized from anywhere in $V$, that is, $Q(\tau < +\infty | Z_0 = z) = 1, \forall z \in V$.

ii) The process stops to move once it arrives at $\Omega$, that is, $Q(Z_{t+s} = Z_t | Z_t \in \Omega) = 1, \forall s, t \geq 0$.

We define the Poisson kernel of $Q$ as the conditional distribution of $Z_\tau$ given $Z_t = z$, denoted by $Q_{\Omega t}(dz | Z_t = z) := Q(Z_\tau = dz | Z_t = z)$. The marginal distribution of $Z_\tau$, which we write as $Q_{\Omega}(dz) = Q(Z_\tau = dz)$, is called the exit distribution, or harmonic measure. Note that $Q_{\Omega t}(dz) = \int_V Q_{\Omega}(dz | Z_0 = z)Q_t(dz)$. The crux of our framework is to leverage the exit distribution $Q_{\Omega}$ as a tool for statistical learning and inference, which is different from traditional frameworks that exploit the properties of the distributions at a fixed time or at convergence.

**Example 2.1 (Sphere Hitting).** As shown in Figure 1-B, let $V = \{ x \in \mathbb{R}^d : \|x\| \leq 1 \}$ be the unit ball and $\Omega = S_d := \partial V$ the unit sphere. Let $Z$ be a Brownian motion starting from $z \in V$ and stopped once it hits the boundary $\Omega$. It is written as

$$Q_{\Omega}^{S_d} : \quad dZ_t = I(\|Z_t\| < 1)dW_t, \quad Z_0 \in V,$$

where $W_t$ is a Wiener process; the indicator function $I(\|Z_t\| < 1)$ sets the velocity to zero and hence stops the process once $Z_t$ hits $\Omega$. The Poisson kernel in this case is a textbook result:

$$Q_{\Omega t}^{S_d}(dz | Z_t = z) \propto \frac{1 - \|z\|^2}{\|x - z\|^d} \times \mu_\Omega(dz), \quad \text{where} \ \mu_\Omega \text{is the surface measure on} \ \Omega = S_d. \quad (2)$$

**Example 2.2 (Boolean Hitting).** As shown in Figure 1-C, let $V = [0, 1]^d$ be the unit cube and $\Omega = B_d := \{ 0, 1 \}^d$ the Boolean cube. Let $Z$ be a Brownian motion starting from $Z_0 \in V$ and confined inside the cube $V$ in the following way:

$$Q_{\Omega}^{B_d} : \quad dZ_{t,i} = I(Z_{t,i} \in (0, 1))dW_{t,i}, \quad \forall i \in \{ 1, 2, \cdots , d \},$$

where $Z_{t,i}$ is the $i$-th element of $Z$. Here, each coordinate $Z_{t,i}$ stops to move once it hits one of the end points $(0, 1)$. It can be viewed as a particle flying in a room that sticks on a wall once it hits it.

**Proposition 2.3.** The Poisson kernel of $Q_{\Omega}^{B_d}$ is a simple product of Bernoulli distributions:

$$Q_{\Omega t}^{B_d}(x | Z_t = z) = \text{Ber}(x | z) := \prod_{i=1}^d \text{Ber}(x_i | z_i), \quad \text{where} \ \text{Ber}(x_i | z_i) = x_i z_i + (1 - x_i)(1 - z_i);$$

Ber$(x_i | z_i)$ is the likelihood function of observing $x_i \in \{ 0, 1 \}$ under Bernoulli$(z_i)$ with $z_i \in \{ 0, 1 \}$.

**Example 2.4 (Fixed Time Hitting).** Our first hitting framework includes the more standard models with fixed terminal time. To see this, let $Z_t = (t, Z_t)$ be a stochastic process $Z_t$ with law $Q$ augmented with time $t$ as one of its coordinates. Let $V = [0, t] \times \mathbb{R}^d$ and $\Omega = \{ t \} \times \mathbb{R}^d$, where $\Omega$ is a vertical plane on the augmented space. Then the hitting time $\tau$ equals $t$ deterministically, and the exit distribution equals the marginal distribution of $Z_t$ at time $t$. See Figure 1-A, for illustration.

![Figure 1](image-url)
2.2 Diffusion Process Tools: Conditioning and $h$-transform

We introduce some basic tools for diffusion processes, including how to conduct conditioning, and exponential tilting (via $h$-transform) on diffusion processes. We apply these tools to the first hitting models we have. The readers can find related background in Oksendal [2013], Särkkä and Solin [2019].

Assume $Z$ is a general Ito diffusion process in $V$ that is absorbed to $\Omega$, denoted as $\text{It}\alpha_\Omega(b, \sigma)$,

$$Q \sim \text{It}\alpha_\Omega(b, \sigma) : \quad dZ_t = b_t(Z_t)dt + \sigma_t(Z_t)dW_t, \quad \forall t \in [0, +\infty), \quad Z_0 \sim Q_0,$$

(3)

where $b_t(x) \in \mathbb{R}^d$ is the drift term and $\sigma_t(x) \in \mathbb{R}^{d \times d}$ is a positive definite diffusion matrix. We always assume that $b$ and $\sigma$ are sufficiently regular to yield a unique weak solution of (3).

**Conditioning**

A step in our work is to find the distribution of the trajectories of a process $Q$ conditioned on a future event, e.g., the event of hitting a particular value $x$ at exit, that is, $\{Z_T = x\}$.

A notable result is that the conditioned diffusion processes are also diffusion processes. Given a point $x \in \Omega$ on the exit surface, the process of $Q(\cdot \mid Z_T = x)$ can be shown to be the law of the following diffusion process [Doob and Doob, 1984, Särkkä and Solin, 2019]:

$$Q(\cdot \mid Z_T = x) : \quad dZ_t = (b_t(Z_t) + \sigma_t^2(Z_t)\nabla Z_t \log q_\Omega(x \mid Z_t)) dt + \sigma_t(Z_t)dW_t, \quad Z_0 \sim \mu_{0|x},$$

(4)

where $q_\Omega(x \mid z)$ is the density function of the Poisson kernel $Q_\Omega(dx \mid Z_T = z)$ w.r.t. a reference measure $\mu_{0|\Omega}$ on $\Omega$, and $\sigma^2$ is the matrix square of $\sigma$, and the conditional initial distribution $\mu_{0|x} = Q_0(\cdot \mid Z_T = x)$ is the posterior probability of $Z_0$ given $Z_T = x$.

Intuitively, the additional drift term $\nabla Z_t \log q_\Omega(x \mid Z_t)$ plays the role of steering the process towards the target $x$, with an increasing magnitude as $Z_t$ approaches $\Omega$ (because $P_\Omega(\cdot \mid Z_T = z)$ converges to a delta measure centered at $x$ when $z$ approaches $\Omega$). This process is known as a diffusion bridge, because it is guaranteed to achieve $Z_T = x$ at the first hitting time with probability one.

**Proposition 2.5.** For $Q^S_d$, the process conditioned on $Z_T = x \in S_d$ at exit is

$$Q^S_d(\cdot \mid Z_T = x) : \quad dZ_t = I(||Z_t|| < 1) \left(\nabla Z_t \log \frac{1 - ||Z_t||^2}{||x - Z_t||^2}dt + dW_t\right).$$

(5)

Here the additional drift term (colored in blue) grows to infinity if $||Z_t|| \to 1$ but $||Z_t - x||$ is large, and hence enforces that $Z_T = x$ when we exit the unit ball.

**Proposition 2.6.** For $Q^I_d$, the process conditioned on $Z_T = x \in \{0, 1\}^d$ at exit is

$$Q^I_d(\cdot \mid Z_T = x) : \quad dZ_{t,i} = I(\mathbf{Z}_{t,i} \in (0, 1)) \left(\frac{2x_i - 1}{x_i z_i + (1 - x_i)(1 - z_i)}dt + dW_{t,i}\right), \quad \forall i.$$

(6)

The additional drift term (colored in blue) enforces that $Z_{T,i} = x_i$ at the exit time as the drift would be infinite if $z_i$ is still far from $x_i$ when $z_i$ is close to $\{0, 1\}$.

**Proposition 2.7.** For the fixed time diffusion in Example 2.4, let $Q^T$ be the standard Brownian motion $dZ_t = dW_t$, stopped at a fixed time $t = T$, then $Q$ conditioned on $Q^T(\cdot \mid Z_T = x)$ is

$$Q^T(\cdot \mid Z_T = x) : \quad dZ_t = I(t \leq T) \left(\frac{Z_t - x}{T - t}dt + dW_t\right).$$

(7)

The additional drift (colored in blue) forces $Z_T = x$ as it grows to infinity if $Z_t \neq x$ while $t \to T$.

**$h$-Transform**

Assume we want to modify the Markov process $Z$ such that its exit distribution $Q_0$ matches the desirable target distribution $\Pi^*$. Doob's $h$-transform Doob and Doob [1984] provides a simple general procedure to do so. Note that by disintegration theorem, we have $Q(dx) = \int Q_\Omega(dx)Q(dZ \mid Z_T = x)$, which factorizes $Q$ into the product of the exit distribution and the conditional process given a fixed exit location $Z_T = x$. To modify the exit distribution of $Q$ to $\Pi^*$, we can simply replace $Q_\Omega$ with $\Pi^*$ in the disintegration theorem, yielding

$$Q^{\Pi^*}(dZ) := \int \Pi^*(dx)Q(dZ \mid Z_T = x) = \pi^*(Z_T)Q(dZ), \quad \text{with } \pi^*(Z_T) := \frac{d\Pi^*}{dQ_\Omega}(Z_T).$$

(8)
where $\pi^* = \frac{d\Pi^*}{d\Omega^*}$ is the Radon–Nikodym derivative (or density ratio) between $\Pi^*$ and $\Omega^*$, and $\Pi^*$ is called an $h$-transform of $Q$. Intuitively, $\Pi^*$ is the distribution of trajectories $Z \sim Q(\cdot | Z_\tau = x)$ when the exit location $x$ is randomly drawn from $x \sim \Pi^*$. We can also view $\pi^*(Z_\tau)$ as an importance score of each trajectory $Z$ based on its terminal state $Z_\tau$, and $\Pi^*$ is obtained by reweighting (or tilting) the probability of each trajectory based on its score.

If $Q$ is a diffusion process, then $\Pi^*$ is also a diffusion process. In addition, $\Pi^*$ is the law of the following diffusion process:

$$\Pi^* : \quad dZ_t = \left( b_t(Z_t) + \sigma_t^2(Z_t) \nabla_x \log h_t^\Pi(Z_t) \right) dt + \sigma_t(Z_t)dW_t, \quad Z_0 \sim Q_0^{\Pi^*}$$

(9)

where the initial distribution $Q_0^{\Pi^*}$ and $h^\Pi_t$ in the drift term are defined as

$$Q_0^{\Pi^*}(dz) = \int_\Omega \pi^*(x)Q(Z_\tau = dx, Z_0 = dz)$$

(10)

$$h_t^\Pi(z) = E_Q[\pi^*(Z_\tau) | Z_t = z] = \int_{\Omega} \pi^*(x)Q(Z_\tau = dx | Z_t = z).$$

(11)

It is clear that $h$ coincides with $\pi^*$ on the boundary, that is, $h_{\pi^*}(x, t) = \pi^*(x)$ for all $x \in \Omega, t \geq 0$. The name of $h$-transform comes from the fact that $h^\Pi_t$ is a (space-time) harmonic function w.r.t. $Q$ in the light of a mean value property: $h^\Pi_t(z) = E_Q[h^\Pi_{t+s}(Z_{t+s}) | Z_t = z], \forall s, t > 0$. $\Pi^*$ yields a simple variational representation in terms of Kullback–Leibler (KL) divergence.

2.3 Learning First Hitting Diffusion Models

Assume $\Pi^*$ is unknown and we observe it through an i.i.d. sample $\{x^{(i)}\}_{i=1}^n$ drawn from $\Pi^*$. We want to fit the data with a parametric diffusion process $\Pi_{\theta}(\cdot | \sigma, \theta)$ in $V$ that is absorbing to $\Omega$,

$$\Pi^\theta : \quad dZ_t = s^\theta_t(Z_t)dt + \sigma_t(Z_t)dW_t, \quad Z_0 \sim \Pi_{\theta}^0,$$

(12)

such that the exit distribution $\Pi^\theta$ matches the unknown $\Pi^*$. Here $s^\theta_t(z)$ is a deep neural network with input $(z, t)$ and parameters $\theta$. We should design $s^\theta$ and $\sigma$ properly to ensure the absorbing property.

The standard approach to estimate $\Pi^*$ is maximum likelihood estimation, which can be viewed as approximatively solving $\min_\theta \mathcal{K}\mathcal{L}(\Pi^* || \Pi^\theta)$. However, calculating the likelihood of the exit distribution $\Pi^\theta$ of a general general diffusion process is computationally intractable. To address this problem, we fix $Q$ as a “prior” process, and augment the data distribution $\Pi^*$ to the $h$-transform $\Pi^*$, whose exit distribution $Q_0^\Pi$ matches $\Pi^*$ by definition. Note that we can draw i.i.d. sample from $Q^\Pi$ in a “backwar” way: first drawing an exit location $x \sim \Pi^*$ from the data, and then draw the trajectory $Z$ from $Q(\cdot | Z_\tau = x)$ with the fixed exit point. To train a generative model, we train $\Pi^\theta$ to fit it with the data drawn from $Q^\Pi$ by maximum likelihood estimation:

$$\min_\theta \left\{ \mathcal{L}(\theta) := \mathcal{K}\mathcal{L}(\Pi^\theta || \Pi^\theta) = -E_{Z \sim \Pi^\theta} \left[ \log p^\theta(Z) \right] + \text{const.} \right\},$$

(13)

where $p^\theta = \frac{d\Pi^\theta}{d\Pi^\pi^*}$ is Radon–Nikodym density function of $\Pi^\theta$ relative to $\Pi^{\Pi^*}$. By the chain rule of KL divergence (??), we have $\mathcal{K}\mathcal{L}(\Pi^* || \Pi^\theta) \leq \mathcal{K}\mathcal{L}(\Pi^\theta || \Pi^\theta)$. Therefore, if minimizing the KL divergence allows us to achieve $\Pi^\theta \approx Q^\Pi$, we should also have $Q_0^\Pi \approx Q_0^\Pi = \Pi^*$.

Using Girsanov theorem [Liptser and Shiriaev, 1977], we can calculate the density function $p^\theta$ and hence the loss function.

**Proposition 2.8.** Assume $Q$ in (3), and $\Pi^\theta$ in (12) are absorbing to $\Omega$. We have

$$\mathcal{L}(\theta) = \frac{1}{2} E_{Q^\Pi} \left[ \int_0^T \left\| \sigma_t(Z_t) - b_t(Z_t | Z_\tau) \right\|^2 dt - \log p^\theta_t(Z_0) \right] + \text{const.}$$

(13)

where $b_t(z|x) := b_t(z) + \sigma_t(Z_t)\nabla_z \log p^{\Pi^*}(x|z)$ is the drift of the conditioned process $Q(\cdot | Z_\tau = x)$ in (4), and $p^\theta_t$ is the probability density function of the initial distribution $P^\theta_0$. In addition, $\Pi^*$ achieves the global minimum of $\mathcal{L}(\theta)$ if

$$s^\Pi_t(z) = E_{Z \sim Q^\Pi}[b_t(z | Z_\tau) | Z_t = z], \quad \Pi^\Pi_0 = Q^{\Pi^* | \cdot}.$$

(13)
Therefore, the optimal drift term $s_t^\theta$ should match the conditional expectation of $b_t(z|x)$ with $x \sim Q_{\Omega}(\cdot|Z_t = z)$, which coincides with the drift of $Q^{\Pi^*}$ in (9). The initial distribution of $P_\theta$ should obviously match the initial distribution of $Q^{\Pi^*}$. In practice, we recommend eliminating the need of estimating $P_\theta$ by starting $Q$ from a deterministic point $Z_0 = z_0$, in which case $P_\theta$ should initialize from the same deterministic point.
References


