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 Π^* 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 Ω and returns a sample drawn from Π^* when it first hits set Ω. We start with introducing the new first hitting model.

*Corresponding author. Email: maoye21@utexas.edu

NeurIPS 2022 Workshop on Score-Based Methods.

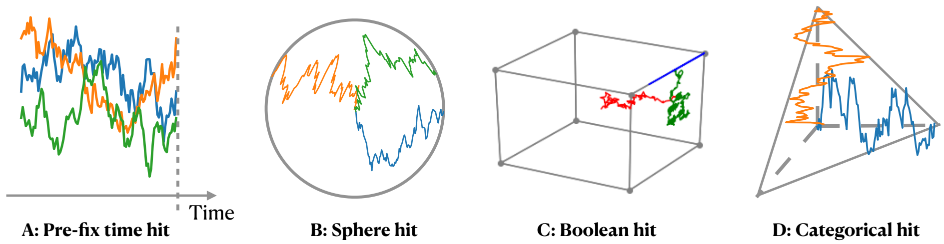

Figure 1: The four hitting schemes introduced in this paper. A: fixed-time hit, the process terminates at a fixed time; B: Sphere hit, hitting the boundary of a sphere from inside; C: Boolean hit, each coordinate terminates when it hits 0 or 1 and the whole process terminates when all of its coordinates terminate; D: Categorical hit, hitting the one-hot codes based on a conditioned process.

Let $Z \coloneqq \{Z_t \colon t \in [0, +\infty)\}$ be a continuous-time Markov process with probability law $\mathbb{Q}$ taking value in a set V that contains Ω as a subset. Here $\mathbb{Q}$ is a probability measure defined on the space of all continuous trajectories $C([0, +\infty),\ \mathbb{R}^d)$. We use $\mathbb{Q}_t$ 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 Ω. Denote by τ the *first hitting time* of Z_t on Ω, that is, $\tau = \inf_t\{t \geq 0 \colon Z_t \in \Omega\}$. We call that Z_t is absorbing to set Ω if

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

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

We define the *Poisson kernel* of $\mathbb{Q}$ as the conditional distribution of Z_τ given $Z_t = z$, denoted by $\mathbb{Q}_\Omega(\mathrm{d}x \mid Z_t = z) \coloneqq \mathbb{Q}(Z_\tau = \mathrm{d}x \mid Z_t = z)$. The marginal distribution of Z_τ, which we write as $\mathbb{Q}_\Omega(\mathrm{d}x) = \mathbb{Q}(Z_\tau = \mathrm{d}x)$, is called the *exit distribution, or harmonic measure*. Note that $\mathbb{Q}_\Omega(\mathrm{d}x) = \int_V \mathbb{Q}_\Omega(\mathrm{d}x \mid Z_0 = z)\mathbb{Q}_0(\mathrm{d}z)$. The crux of our framework is to leverage the exit distribution $\mathbb{Q}_\Omega$ 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 $V = \{x \in \mathbb{R}^d \colon \|x\| \leq 1\}$ be the unit ball and $\Omega = S_d \coloneqq \partial V$ the unit sphere. Let Z be a Brownian motion starting from $z \in V$ and stopped once it hits the boundary Ω. It is written as*

$$\mathbb{Q}^{S_d}: \qquad \mathrm{d}Z_t = \mathbb{I}(\|Z_t\| < 1)\mathrm{d}W_t, \quad Z_0 \in V, \tag{1}$$

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

$$\mathbb{Q}_\Omega^{S_d}(\mathrm{d}x \mid Z_t = z) \propto \frac{1 - \|z\|^2}{\|x - z\|^d} \times \mu_\Omega(\mathrm{d}z), \quad \textit{where } \mu_\Omega \textit{ is the surface measure on } \Omega = S_d. \tag{2}$$

Example 2.2 (Boolean Hitting). *As shown in Figure 1-C, let $V = [0, 1]^d$ be the unit cube and $\Omega = B_d \coloneqq \{0, 1\}^d$ 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:*

$$\mathbb{Q}^{B_d}: \qquad \mathrm{d}Z_{t,i} = \mathbb{I}(Z_{t,i} \in (0, 1))\mathrm{d}W_{t,i}, \quad \forall i \in \{1, 2, \cdots, d\},$$

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 or 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 $\mathbb{Q}^{B_d}$ is a simple product of Bernoulli distributions:*

$$\mathbb{Q}_\Omega^{B_d}(x \mid Z_t = z) = \mathrm{Ber}(x|z) \coloneqq \prod_{i=1}^{d} \mathrm{Ber}(x_i|z_i), \quad \textit{where} \quad \mathrm{Ber}(x_i|z_i) = x_i z_i + (1 - x_i)(1 - z_i);$$

$\mathrm{Ber}(x_i|z_i)$ *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 $\bar{Z}_t = (t, Z_t)$ be a stochastic process Z_t with law $\mathbb{Q}$ 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 Ω is a vertical plane on the augmented space. Then the hitting time τ equals t deterministically, and the exit distribution equals the marginal distribution of Z_t at time t. See Figure 1-A, for illustration.*

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 Ω, denoted as $\mathrm{Ito}_\Omega(b, \sigma)$,

$$\mathbb{Q} \sim \mathrm{Ito}_\Omega(b, \sigma): \qquad \mathrm{d}Z_t = b_t(Z_t)\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \quad \forall t \in [0, +\infty), \qquad Z_0 \sim \mathbb{Q}_0, \quad (3)$$

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 σ 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 $\mathbb{Q}$ conditioned on a future event, e.g., the event of hitting a particular value x at exit, that is, $\{Z_\tau = x\}$. 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 $\mathbb{Q}(\cdot \mid Z_\tau = x)$ can be shown to be the law of the following diffusion process [Doob and Doob, 1984, Särkkä and Solin, 2019]:

$$\mathbb{Q}(\cdot|Z_\tau = x): \quad \mathrm{d}Z_t = \left(b_t(Z_t) + \sigma_t^2(Z_t)\nabla_{Z_t}\log q_\Omega(x \mid Z_t)\right)\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \;\; Z_0 \sim \mu_{0|x}, \quad (4)$$

where $q_\Omega(x \mid z)$ is the density function of the Poisson kernel $\mathbb{Q}_\Omega(\mathrm{d}x \mid Z_t = z)$ w.r.t. a reference measure μ_Ω on Ω, and σ^2 is the matrix square of σ, and the conditional initial distribution $\mu_{0|x} = \mathbb{Q}_0(\cdot \mid Z_\tau = x)$ is the posterior probability of Z_0 given $Z_\tau = x$.

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

Proposition 2.5. *For $\mathbb{Q}^{S_d}$, the process conditioned on $Z_\tau = x \in S_d$ at exit is*

$$\mathbb{Q}^{S_d}(\cdot \mid Z_\tau = x): \qquad \mathrm{d}Z_t = \mathbb{I}(\|Z_t\| < 1)\left(\nabla_{Z_t}\log\frac{1 - \|Z_t\|^2}{\|x - Z_t\|^d}\mathrm{d}t + \mathrm{d}W_t\right). \quad (5)$$

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 $Z_\tau = x$ when we exit the unit ball.

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

$$\mathbb{Q}^{B_d}(\cdot|Z_\tau = x): \quad \mathrm{d}Z_{t,i} = \mathbb{I}(Z_{t,i} \in (0,1))\left(\frac{2x_i - 1}{x_i z_i + (1 - x_i)(1 - z_i)}\mathrm{d}t + \mathrm{d}W_{t,i}\right), \quad \forall i. \quad (6)$$

The additional drift term (colored in blue) enforces that $Z_{\tau,i} = x_i$ 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 $\mathbb{Q}^T$ be the standard Brownian motion $\mathrm{d}Z_t = \mathrm{d}W_t$ stopped at a fixed time $t = T$, then $\mathbb{Q}$ conditioned on $\mathbb{Q}^T(Z|Z_T = x)$ is*

$$\mathbb{Q}^T(\cdot|Z_\tau = x): \qquad \mathrm{d}Z_t = \mathbb{I}(t \le T)\left(\frac{Z_t - x}{T - t}\mathrm{d}t + \mathrm{d}W_t\right). \quad (7)$$

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 $\mathbb{Q}_\Omega$ matches the desirable target distribution Π^*. Doob's h-transform Doob and Doob [1984] provides a simple general procedure to do so. Note that by disintegration theorem, we have $\mathbb{Q}(\mathrm{d}Z) = \int \mathbb{Q}_\Omega(\mathrm{d}x)\mathbb{Q}(\mathrm{d}Z \mid Z_\tau = x)$, which factorizes $\mathbb{Q}$ into the product of the exit distribution and the conditional process given a fixed exit location $Z_\tau = x$. To modify the exit distribution of $\mathbb{Q}$ to Π^*, we can simply replace $\mathbb{Q}_\Omega$ with Π^* in the disintegration theorem, yielding

$$\mathbb{Q}^{\Pi^*}(\mathrm{d}Z) := \int \Pi^*(\mathrm{d}x)\mathbb{Q}(\mathrm{d}Z \mid Z_\tau = x) = \pi^*(Z_\tau)\mathbb{Q}(\mathrm{d}Z), \quad \text{with} \;\; \pi^*(Z_\tau) := \frac{\mathrm{d}\Pi^*}{\mathrm{d}\mathbb{Q}_\Omega}(Z_\tau), \quad (8)$$

where $\pi^* = \frac{\mathrm{d}\Pi^*}{\mathrm{d}\mathbb{Q}_\Omega}$ is the Radon–Nikodym derivative (or density ratio) between Π^* and $\mathbb{Q}_\Omega$, and $\mathbb{Q}^{\Pi^*}$ is called an h-transform of $\mathbb{Q}$. Intuitively, $\mathbb{Q}^{\Pi^*}$ is the distribution of trajectories $Z \sim \mathbb{Q}(\cdot|Z_\tau = x)$ 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_τ, and $\mathbb{Q}^{\Pi^*}$ is obtained by reweighing (or tilting) the probability of each trajectory based on its score.

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

$$\mathbb{Q}^{\Pi^*}: \qquad \mathrm{d}Z_t = \left(b_t(Z_t) + \sigma_t^2(Z_t)\nabla_z \log h_t^{\Pi^*}(Z_t)\right)\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \quad Z_0 \sim \mathbb{Q}_0^{\Pi^*} \tag{9}$$

where the initial distribution $\mathbb{Q}_0^{\Pi^*}$ and h^{Π^*} in the drift term are defined as

$$\mathbb{Q}_0^{\Pi^*}(\mathrm{d}z) = \int_\Omega \pi^*(x)\mathbb{Q}(Z_\tau = \mathrm{d}x, Z_0 = \mathrm{d}z) \tag{10}$$

$$h_t^{\Pi^*}(z) = \mathbb{E}_{\mathbb{Q}}[\pi^*(Z_\tau) \mid Z_t = z] = \int_\Omega \pi^*(x)\mathbb{Q}(Z_\tau = \mathrm{d}x \mid Z_t = z). \tag{11}$$

It is clear that h coincides with π^* 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^{Π^*} is a (space-time) harmonic function w.r.t. $\mathbb{Q}$ in the light of a mean value property: $h_t^{\Pi^*}(z) = \mathbb{E}_{\mathbb{Q}}[h_{t+s}^{\Pi^*}(Z_{t+s}) \mid Z_t = z], \forall s,t > 0$. $\mathbb{Q}^{\Pi^*}$ yields a simple variational representation in terms of Kullback–Leibler (KL) divergence.

2.3 Learning First Hitting Diffusion Models

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

$$\mathbb{P}^\theta: \qquad \mathrm{d}Z_t = s_t^\theta(Z_t)\mathrm{d}t + \sigma_t(Z_t)\mathrm{d}W_t, \qquad Z_0 \sim \mathbb{P}_0^\theta, \tag{12}$$

such that the exit distribution $\mathbb{P}_\Omega^\theta$ matches the unknown Π^*. Here $s_t^\theta(z)$ is a deep neural network with input (z,t) and parameters θ. We should design s^θ and σ properly to ensure the absorbing property.

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

$$\min_\theta \left\{ \mathcal{L}(\theta) := \mathcal{KL}(\mathbb{Q}^{\Pi^*} \mid\mid \mathbb{P}^\theta) \equiv -\mathbb{E}_{Z\sim\mathbb{Q}^{\Pi^*}}\left[\log p^\theta(Z)\right] + const, \right\},$$

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

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

Proposition 2.8. *Assume $\mathbb{Q}$ in (3), and $\mathbb{P}^\theta$ in (12) are absorbing to Ω. We have*

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}_{\mathbb{Q}^{\Pi^*}}\left[\int_0^\tau \left\|\sigma_t(Z_t)^{-1}(s_t^\theta(Z_t) - b_t(Z_t \mid Z_\tau))\right\|^2 \mathrm{d}t - \log p_0^\theta(Z_0)\right] + const, \tag{13}$$

where $b_t(z|x) := b_t(z) + \sigma_t^2(z)\nabla_z \log p_\Omega(x|z)$ is the drift of the conditioned process $\mathbb{Q}(\cdot|Z_\tau = x)$ in (4), and p_0^θ is the probability density function of the initial distribution $\mathbb{P}_0^\theta$. In addition, θ^* achieves the global minimum of $\mathcal{L}(\theta)$ if

$$s_t^{\theta^*}(z) = \mathbb{E}_{Z\sim\mathbb{Q}^{\Pi^*}}[b_t(z|Z_\tau) \mid Z_t = z], \qquad \mathbb{P}_0^{\theta^*} = \mathbb{Q}_0^{\Pi^*} = \mathbb{E}_{x\sim\Pi^*}[\mathbb{Q}_0^x(\cdot)].$$

Therefore, the optimal drift term $s_t^{\theta^*}$ should match the conditional expectation of $b_t(z|x)$ with $x \sim \mathbb{Q}_\Omega(\cdot|Z_t = z)$, which coincides with the drift of $\mathbb{Q}^{\Pi^*}$ in (9). The initial distribution of $\mathbb{P}^\theta$ should obviously match the initial distribution of $\mathbb{Q}^{\Pi^*}$. In practice, we recommend eliminating the need of estimating $\mathbb{P}\theta_0$ by starting $\mathbb{Q}$ from a deterministic point $Z_0 = z_0$, in which case $\mathbb{P}^\theta$ should initialize from the same deterministic point.

References

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021.

Joseph L Doob and JI Doob. *Classical potential theory and its probabilistic counterpart*, volume 549. Springer, 1984.

Robert Shevilevich Liptser and Al'bert Nikolaevich Shiriaev. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

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