Statistical Convergence of First Hitting Diffusion Models

joint work with Lukas Trottner

Simon Bienewald

University of Stuttgart
Department of Mathematics
Institute for Stochastics and Applications

Motivation

Goal: Generate data on the unit sphere $\partial B \subset \mathbb{R}^d$ according to a distribution Π^* based on n data points (e.g. climate data).

Motivation

Goal: Generate data on the unit sphere $\partial B \subset \mathbb{R}^d$ according to a distribution Π^* based on n data points (e.g. climate data).

Standard diffusion models are not tailored to this setting:

Motivation

Goal: Generate data on the unit sphere $\partial B \subset \mathbb{R}^d$ according to a distribution Π^* based on n data points (e.g. climate data).

Standard diffusion models are not tailored to this setting:

- **Issue 1: No guarantee to stop on** ∂B

Motivation

Goal: Generate data on the unit sphere $\partial B \subset \mathbb{R}^d$ according to a distribution Π^* based on n data points (e.g. climate data).

Standard diffusion models are not tailored to this setting:

- **Issue 1: No guarantee to stop on** ∂B
- **Issue 2: Number of noising steps need to be** $\mathcal{O}(\log(n)/d)$ **for optimal convergence** (Oko, Akiyama, and Suzuki 2023)

Motivation

Goal: Generate data on the unit sphere $\partial B \subset \mathbb{R}^d$ according to a distribution Π^* based on n data points (e.g. climate data).

Standard diffusion models are not tailored to this setting:

- **Issue 1: No guarantee to stop on** ∂B
- **Issue 2: Number of noising steps need to be** $\mathcal{O}(\log(n)/d)$ **for optimal convergence** (Oko, Akiyama, and Suzuki 2023)
- **Issue 3: The score** $\nabla \log p_t$ **is time-dependent**

Motivation

Goal: Generate data on the unit sphere $\partial B \subset \mathbb{R}^d$ according to a distribution Π^* based on n data points (e.g. climate data).

Standard diffusion models are not tailored to this setting:

- **Issue 1: No guarantee to stop on** ∂B
- **Issue 2: Number of noising steps need to be** $\mathcal{O}(\log(n)/d)$ **for optimal convergence** (Oko, Akiyama, and Suzuki 2023)
- **Issue 3: The score** $\nabla \log p_t$ **is time-dependent**

$\rightarrow$ Idea by Ye, Wu, and Liu (2022): First Hitting Diffusion Models based on Doob's h-transform.

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

Generating process ("backward" process):

$$\mathrm{d}Z_t^h = \nabla \log h(Z_t^h)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z_0^h = 0.$$

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

Generating process ("backward" process):

$$\mathrm{d}Z_t^h = \nabla \log h(Z_t^h)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z_0^h = 0.$$

By construction,

$$\mathbb{P}(Z_{\tau^h}^h \in \cdot) = \Pi^*,$$

for $\tau^h = \inf\{t \geq 0 : \|Z_t^h\| = 1\}$.

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

Generating process ("backward" process):

$$\mathrm{d}Z_t^h = \nabla \log h(Z_t^h)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z_0^h = 0.$$

By construction,

$$\mathbb{P}(Z_{\tau^h}^h \in \cdot) = \Pi^*,$$

for $\tau^h = \inf\{t \geq 0 : \|Z_t^h\| = 1\}$.

- **Improvement 1: Stops exactly on ∂B**

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

Generating process ("backward" process):

$$\mathrm{d}Z_t^h = \nabla \log h(Z_t^h)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z_0^h = 0.$$

By construction,

$$\mathbb{P}(Z_{\tau^h}^h \in \cdot) = \Pi^*,$$

for $\tau^h = \inf\{t \geq 0 : \|Z_t^h\| = 1\}$.

- **Improvement 1: Stops exactly on ∂B**
- **Improvement 2: Average simulation time $\mathbb{E}[\tau^h] = 1/d$**

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

Generating process ("backward" process):

$$\mathrm{d}Z_t^h = \nabla \log h(Z_t^h)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z_0^h = 0.$$

By construction,

$$\mathbb{P}(Z_{\tau^h}^h \in \cdot) = \Pi^*,$$

for $\tau^h = \inf\{t \geq 0 : \|Z_t^h\| = 1\}$.

- **Improvement 1: Stops exactly on ∂B**
- **Improvement 2: Average simulation time $\mathbb{E}[\tau^h] = 1/d$**
- **Improvement 3: $\nabla \log h$ is time-independent**

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

Generating process ("backward" process) with neural network s:

$$\mathrm{d}Z_t^s = s(Z_t^s)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z_0^s = 0.$$

By construction,

$$\mathbb{P}(Z_{\tau^s}^s \in \cdot) \approx \Pi^*,$$

for $\tau^s = \inf\{t \geq 0 : \|Z_t^s\| = 1\}$.

- **Improvement 1: Stops exactly on ∂B**
- **Improvement 2: Average simulation time $\mathbb{E}[\tau^h] = 1/d$**
- **Improvement 3: $\nabla \log h$ is time-independent**

First Hitting Diffusion Models (FHDMs)

Define

$$h(z) = \mathbb{E}^z[\pi^*(W_\tau)], \qquad z \in B.$$

Generating process ("backward" process) with neural network s:

$$\mathrm{d}Z_t^s = s(Z_t^s)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z_0^s = 0.$$

By construction,

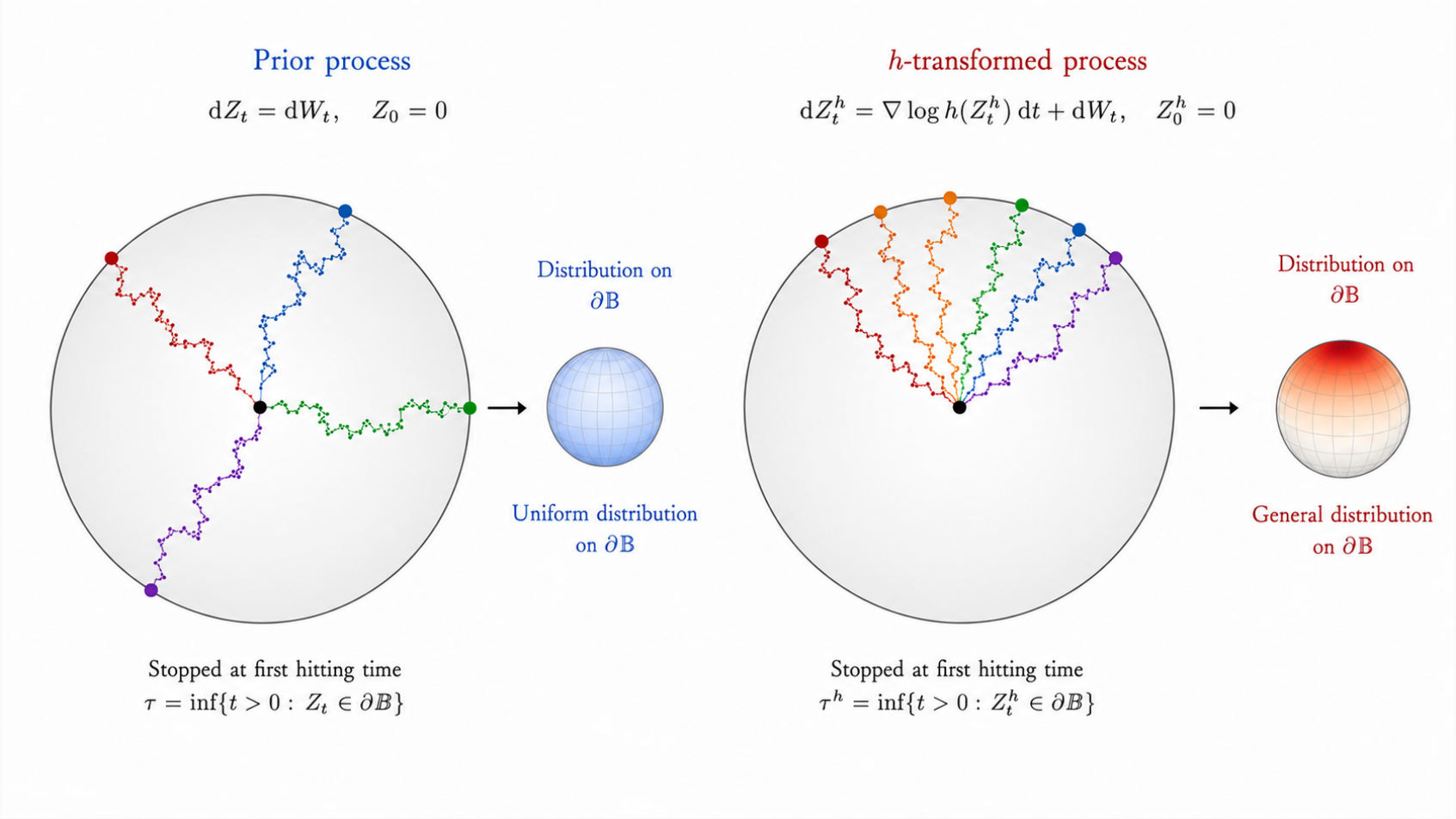
$$\mathbb{P}(Z^s_{\tau^s_{1-\varepsilon}}/(1-\varepsilon) \in \cdot) \approx \Pi^*,$$

for $\tau^s_{1-\varepsilon} = \inf\{t \geq 0 : \|Z_t^s\| = 1-\varepsilon\}$.

- **Improvement 1: Stops exactly on ∂B**
- **Improvement 2: Average simulation time $\mathbb{E}[\tau^h] = 1/d$**
- **Improvement 3: $\nabla \log h$ is time-independent**

First Hitting Diffusion Models (FHDMs)

Graphical example:



Assumptions

$(\mathcal{H}1)$ $d \geq 3$.

Assumptions

$(\mathcal{H}1)$ $d \geq 3$.

$(\mathcal{H}2)$ π^* is bounded from below by $\pi_{\min} > 0$.

Assumptions

$(\mathcal{H}1)$ $d \geq 3$.

$(\mathcal{H}2)$ π^* is bounded from below by $\pi_{\min} > 0$.

$(\mathcal{H}3)$ $\pi^* \in H^\alpha(\partial B)$ for $\alpha \in \mathbb{N}$ and $\alpha > (d-1)/2$.

Assumptions

$(\mathcal{H}1)$ $d \geq 3$.

$(\mathcal{H}2)$ π^* is bounded from below by $\pi_{\min} > 0$.

$(\mathcal{H}3)$ $\pi^* \in H^\alpha(\partial B)$ for $\alpha \in \mathbb{N}$ and $\alpha > (d-1)/2$.

Minimax lower bound [Yang and Barron (1999)]

The minimax lower bound for the class of densities $\pi \in H^\alpha(\partial B)$ with $\|\pi\|_{H^\alpha(\partial B)} \leq K$ over the set $\mathcal{E}_n$ of all density-valued estimators based on n observations is

$$\inf_{\hat{\rho} \in \mathcal{E}_n} \sup_{\pi \in \mathcal{H}^\alpha(\partial B, K)} \mathbb{E}_{\bigotimes_{i=1}^n \pi}[\mathrm{TV}(\pi, \hat{\rho}(X))] \gtrsim n^{-\alpha/(2\alpha+d-1)}.$$

Main result

Theorem [S.B. and Trottner (2026)]

Under assumptions $(\mathcal{H}1)$-$(\mathcal{H}3)$ there exists neural network sizes

$$\mathsf{L} \lesssim \log^2 n, \qquad \mathsf{W} \lesssim n^{(d-1)/(2\alpha+d-1)} \log^2 n,$$
$$\mathsf{S} \lesssim n^{(d-1)/(2\alpha+d-1)} \log^3 n, \qquad \mathsf{B} \lesssim \mathrm{Poly}(n),$$

and $\varepsilon = \mathrm{Poly}(n^{-1})$ such that the ERM $\hat{s}$ of the score matching loss over a neural network class $\mathcal{S} = \mathcal{S}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ fulfills

$$\mathbb{E}\Big[\mathrm{TV}\big(\Pi^*, Z^{\hat{s}}_{\tau^{\hat{s}}_{1-\varepsilon}}/(1-\varepsilon)\big)\Big] \lesssim n^{-\alpha/(2\alpha+d-1)} \log^{7/2}(n).$$

Main result

Theorem [S.B. and Trottner (2026)]

Under assumptions $(\mathcal{H}1)$-$(\mathcal{H}3)$ there exists neural network sizes

$$\mathsf{L} \lesssim \log^2 n, \qquad \mathsf{W} \lesssim n^{(d-1)/(2\alpha+d-1)} \log^2 n,$$
$$\mathsf{S} \lesssim n^{(d-1)/(2\alpha+d-1)} \log^3 n, \qquad \mathsf{B} \lesssim \mathrm{Poly}(n),$$

and $\varepsilon = \mathrm{Poly}(n^{-1})$ such that the ERM $\hat{s}$ of the score matching loss over a neural network class $\mathcal{S} = \mathcal{S}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ fulfills

$$\mathbb{E}\Big[\mathrm{TV}\big(\Pi^*, Z^{\hat{s}}_{\tau^{\hat{s}}_{1-\varepsilon}}/(1-\varepsilon)\big)\Big] \lesssim n^{-\alpha/(2\alpha+d-1)} \log^{7/2}(n).$$

$\rightarrow$ FHDMs adapt to low dimensional support of Π^*.

References

- Oko, Kazusato, Shunta Akiyama, and Taiji Suzuki (2023). "Diffusion Models are Minimax Optimal Distribution Estimators". In: *International Conference on Machine Learning*.
- S.B. and Lukas Trottner (2026). *Statistical Convergence of Spherical First Hitting Diffusion Models*. arXiv: 2605.07625 [math.ST].
- Yang, Yuhong and Andrew Barron (1999). "Information-theoretic determination of minimax rates of convergence". In: *Ann. Statist.* 27.5, pp. 1564–1599.

 Ye, Mao, Lemeng Wu, and Qiang Liu (2022). “First Hitting Diffusion Models for Generating Manifold, Graph and Categorical Data”. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 27280–27292.