Statistical guarantees for denoising reflected diffusion models

MFO Mini-Workshop on Statistical Challenges for Deep Generative Models

Claudia Strauch and Lukas Trottner
joint work with Asbjørn Holk Thomsen

17 February 2025

Heidelberg University University of Birmingham Aarhus University

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

UNIVERSITY OF BIRMINGHAM

Motivation:

"Creating noise from data is easy; creating data from noise is **generative modeling**.*"*

Source: Song et al. (2021). Score based generative modeling through stochastic differential equations. *ICLR*.

Generative modelling

- ⇝ involves **learning the underlying distribution** of a dataset to **generate new, similar data points**
- ⇝ aims to model the data distribution $p(x)$ for observed data x, allowing the **generation of new samples** x' that resemble the original data
- ⇝ essential in applications like **image synthesis**, text generation, and data augmentation

Core tasks:

1. **Density estimation:** Learning the probability distribution $p(x)$ or its properties.
2. **Sampling:** Drawing new samples x' from the learned $p(x)$.

Examples: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and normalizing flows

Generative modelling

- ⇝ involves **learning the underlying distribution** of a dataset to **generate new, similar data points**
- ⇝ aims to model the data distribution $p(x)$ for observed data x, allowing the **generation of new samples** x' that resemble the original data
- ⇝ essential in applications like **image synthesis**, text generation, and data augmentation

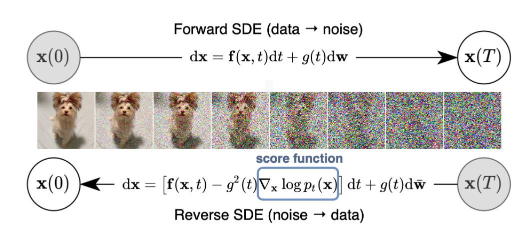
Core tasks:

1. **Density estimation:** Learning the probability distribution $p(x)$ or its properties.
2. **Sampling:** Drawing new samples x' from the learned $p(x)$.

Examples: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and normalizing flows

Denoising diffusion models (DDMs)

- provide an **iterative generative algorithm** to create new samples that approximately match the target distribution p_0, given a finite number of samples corresponding to an unknown p_0
- **general idea:** find a **stochastic process** that perturbs p_0 to a new distribution p_T such that
  1) p_T or a good approximation thereof is **easy to sample from**, and
  2) the perturbation is **reversible** in the sense that we know how to **simulate the time-reversed process**



Source: Song et al. (2021). Score based generative modeling through stochastic differential equations. *ICLR.*

Statistical challenges in generative modelling

Unifying principle of generative modelling:

transform noise to **create new data** that matches a given training data set

⇝ transformations must **adapt to the information contained in the training data**, which is **high-dimensional** in typical machine learning applications

Generative models have demonstrated remarkable **empirical success** across diverse domains, including images, videos, and text, despite their differences in methodology:

- models like **Generative Adversarial Networks (GANs)** aim to **directly approximate** the transformation from noise to data using adversarial training
- **Denoising Diffusion Models (DDMs) dynamically evolve noise into data** by approximating the characteristics of a **stochastic process**

Under what conditions do these models ensure that the generated distribution converges to the target distribution at a (minimax) optimal rate?

<u>Statistical</u> challenges in generative modelling

Unifying principle of generative modelling:

transform noise to **create new data** that matches a given training data set

$\rightsquigarrow$ transformations must **adapt to the information contained in the training data**, which is **high-dimensional** in typical machine learning applications

Generative models have demonstrated remarkable **empirical success** across diverse domains, including images, videos, and text, despite their differences in methodology:

- models like **Generative Adversarial Networks (GANs)** aim to **directly approximate** the transformation from noise to data using adversarial training
- **Denoising Diffusion Models (DDMs) dynamically evolve noise into data** by approximating the characteristics of a **stochastic process**

Under what conditions do these models ensure that the generated distribution converges to the target distribution at a (minimax) optimal rate?

Classical Denoising Diffusion Models (DDMs)

- for some fixed time $T > 0$ and suitable drift $b\colon [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and dispersion $\sigma\colon [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$, consider the **forward model**

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}W_t, \qquad t \in [0,T], X_0 \sim p_0,$$

  $W = (W_t)_{t\in[0,T]}$ some standard d-dimensional Brownian motion

- under sufficient regularity conditions, the forward model has a solution $X = (X_t)_{t\in[0,T]}$ with marginal densities $(p_t)_{t\in[0,T]}$ such that the **time-reversed process** $\overleftarrow{X}_t = X_{T-t}$, $t \in [0,T]$, solves

$$\mathrm{d}\overleftarrow{X}_t = -\overline{b}(T-t, \overleftarrow{X}_t)\,\mathrm{d}t + \sigma(T-t, \overleftarrow{X}_t)\,\mathrm{d}\overline{W}_t, \qquad t \in [0,T], \overleftarrow{X}_0 \sim p_T,$$

  for some Brownian motion $(\overline{W}_t)_{t\in[0,T]}$ and drift $\overline{b}\colon [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ given by

$$\overline{b}_i(t,x) = b_i(t,x) - \frac{1}{p_t(x)} \sum_{j,k=1}^{d} \frac{\partial}{\partial x_j}\left[p_t(x)\sigma_{ik}(t,x)\sigma_{jk}(t,x)\right], \qquad i = 1, \dots, d$$

⇝ time-reversed process solves a **time-inhomogeneous SDE**, now with drift $-\overline{b}(T-\cdot,\cdot)$ and dispersion $\sigma(T-\cdot,\cdot)$

Classical Denoising Diffusion Models (DDMs)

- for some fixed time $T > 0$ and suitable drift $b\colon [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ and dispersion $\sigma\colon [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$, consider the **forward model**

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}W_t, \qquad t \in [0, T], X_0 \sim p_0,$$

  $W = (W_t)_{t\in[0,T]}$ some standard d-dimensional Brownian motion

- under sufficient regularity conditions, the forward model has a solution $X = (X_t)_{t\in[0,T]}$ with marginal densities $(p_t)_{t\in[0,T]}$ such that the **time-reversed process** $\overleftarrow{X}_t = X_{T-t}$, $t \in [0, T]$, solves

$$\mathrm{d}\overleftarrow{X}_t = -\overline{b}(T-t, \overleftarrow{X}_t)\,\mathrm{d}t + \sigma(T-t, \overleftarrow{X}_t)\,\mathrm{d}\overline{W}_t, \qquad t \in [0, T], \overleftarrow{X}_0 \sim p_T,$$

  for some Brownian motion $(\overline{W}_t)_{t\in[0,T]}$ and drift $\overline{b}\colon [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ given by

$$\overline{b}_i(t, x) = b_i(t, x) - \frac{1}{p_t(x)} \sum_{j,k=1}^{d} \frac{\partial}{\partial x_j}\big[p_t(x)\sigma_{ik}(t, x)\sigma_{jk}(t, x)\big], \qquad i = 1, \dots, d$$

⇝ time-reversed process solves a **time-inhomogeneous SDE**, now with drift $-\overline{b}(T-\cdot,\cdot)$ and dispersion $\sigma(T-\cdot,\cdot)$

Classical DDMs

- **standard convention:** set $\sigma(t,x) = \gamma(t)\mathbb{I}_d$ for some scalar function γ

⇝ **forward model** is given by a (possibly time-inhomogeneous) **Ornstein–Uhlenbeck process** with explicit transition densities, and the backward drift becomes

$$\overline{b}(t,x) = b(t,x) - \gamma^2(t) \underbrace{\nabla \log p_t(x)}_{\text{“score” of the forward model}}$$

⇝ **backwards process** has the dynamics

$$\mathrm{d}\overleftarrow{X}_t = \big(- b(T-t, \overleftarrow{X}_t) + \gamma^2(T-t)\nabla \log p_{T-t}(\overleftarrow{X}_t)\big)\,\mathrm{d}t + \gamma(T-t)\,\mathrm{d}\overline{W}_t \quad t \in [0,T], \overleftarrow{X}_0 \sim p_T$$

- as $t \to T$, the density of $\overleftarrow{X}_t$ approaches $p_0 \implies$ **simulating the reverse process** generates new data **samples corresponding to the target** p_0
- **note:** we are free to choose the coefficients of our forward process (i.e., b and σ), but the **score function** $\nabla \log p_t$ depends on p_0 ⇝ needs to be estimated from the data (“**score matching**”)

Classical DDMs

- **standard convention:** set $\sigma(t,x) = \gamma(t)\mathbb{I}_d$ for some scalar function γ

⇝ **forward model** is given by a (possibly time-inhomogeneous) **Ornstein–Uhlenbeck process** with explicit transition densities, and the backward drift becomes

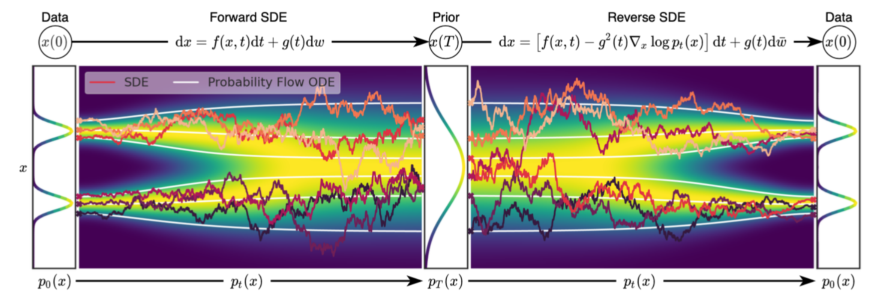
$$\overline{b}(t,x) = b(t,x) - \gamma^2(t)\underbrace{\nabla \log p_t(x)}_{\text{“score” of the forward model}}$$

⇝ **backwards process** has the dynamics

$$\mathrm{d}\overleftarrow{X}_t = \big(-b(T-t,\overleftarrow{X}_t) + \gamma^2(T-t)\nabla \log p_{T-t}(\overleftarrow{X}_t)\big)\,\mathrm{d}t + \gamma(T-t)\,\mathrm{d}\overline{W}_t \quad t \in [0,T], \overleftarrow{X}_0 \sim p_T$$

- as $t \to T$, the density of $\overleftarrow{X}_t$ approaches $p_0 \implies$ **simulating the reverse process** generates new data **samples corresponding to the target p_0**
- **note:** we are free to choose the coefficients of our forward process (i.e., b and σ), but the **score function $\nabla \log p_t$** depends on p_0 ⇝ needs to be estimated from the data (“**score matching**”)

Statistical aspects of denoising diffusion models



Source: Song et al. (2021). Score based generative modeling through stochastic differential equations. *ICLR*.

Statistical questions:

1. are diffusion models **minimax learners** (in terms of smoothness assumptions on p_0)?
2. how can empirical lack of curse of dimensionality be explained? $\rightsquigarrow$ **submanifold hypothesis**
3. alternative model designs with better theoretical/experimental justification?

Diffusion models are minimax optimal distribution estimators[1]

"Is diffusion modeling a **good distribution estimator***? In other words, how can the* **estimation error of the generated data distribution** *be explicitly bounded by the number of the training data and in a data structure dependent way?"*

Assumptions on the **initial distribution with density p_0** can be summarised by three key components:

(i) p_0 is **compactly supported** on a d-dimensional hypercube;

(ii) p_0 is **bounded away from zero** on its support;

(iii) p_0 has **Besov smoothness of order s away from the support boundary** (where s is allowed to be sufficiently small to not necessarily imply continuity of p_0) and is **infinitely differentiable close to the boundary**.

Under these conditions, Oko et al. (2024) show that generated data distribution achieves the **nearly minimax optimal estimation rate $n^{-\frac{s}{2s+d}}(\log n)^8$** in total variation distance.

[1] K. Oko, S. Akiyama, and T. Suzuki (2023). **Diffusion Models are Minimax Optimal Distribution Estimators**. *ICML*.

Convergence of diffusion models under the manifold hypothesis

Convergence rates (even optimal ones) expressed **in terms of the ambient dimension d** fall short of capturing the empirical success of DDMs

$\rightsquigarrow$ gap is related to the **manifold hypothesis**: real-world high-dimensional data often reside on **lower-dimensional manifolds**, to which well-trained **generative models are believed to adapt**

Tang and Yang (2024)[2] establish (up to log factors) the **minimax convergence rate** $C(d)n^{-\frac{s+1}{2s+\widetilde{d}}}$ in Wasserstein-1 distance for distributions p_0 such that

(i) p_0 is supported on a **compact and β-smooth $\widetilde{d}$-dimensional** submanifold $\mathcal{M}$, where $\beta \geq 2$;

(ii) p_0 is **bounded away from zero** on $\mathcal{M}$;

(iii) p_0 has **smoothness of order** $s \in [0, \beta - 1]$ w.r.t. the volume measure on $\mathcal{M}$.

[2] R. Tang and Y. Yang (2024). **Adaptivity of Diffusion Models to Manifold Structures**. *AISTATS*.

Convergence of diffusion models under the manifold hypothesis

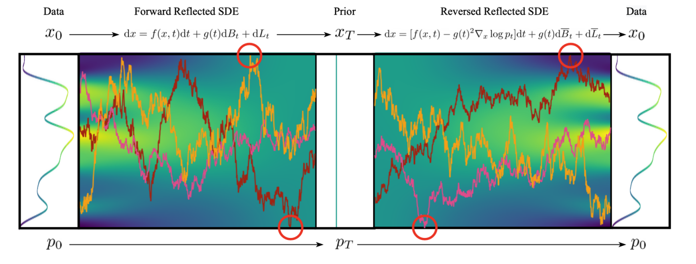
Convergence rates (even optimal ones) expressed **in terms of the ambient dimension** d fall short of capturing the empirical success of DDMs

$\rightsquigarrow$ gap is related to the **manifold hypothesis**: real-world high-dimensional data often reside on **lower-dimensional manifolds**, to which well-trained **generative models are believed to adapt**

- **multiplicative factor** $C(d)$ in Tang and Yang's convergence rate is **of order** $d^{s+\widetilde{d}/2}$ and thus potentially very large for high ambient dimension d
- most recently, Azangulov et al. (2024)[2] show that this **multiplicative factor can be significantly reduced to the order** $\sqrt{d}$

[2] I. Azangulov, G. Deligiannidis and J. Rousseau (2024). **Convergence of Diffusion Models Under the Manifold Hypothesis in High-Dimensions**. `arXiv:2409.18804`

Denoising reflected diffusion models



Source: Lou and Ermon (2023). Reflected Diffusion Models. *ICML*.

Questions:

1. are diffusion models minimax learners (in terms of smoothness assumptions on p_0)?
2. how can empirical lack of curse of dimensionality be explained? ⇝ submanifold hypothesis
3. alternative model designs with better theoretical/experimental justification?

Denoising Reflected Diffusion Models (DRDMs) in a nutshell

- extend DDMs by constraining both forward and backward processes to a bounded domain $D \subset \mathbb{R}^d$
- forward process includes a reflection term to enforce boundary constraints,

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t + \nu(X_t)\,\mathrm{d}\ell_t, \quad X_0 \in \overline{D},$$

  where ℓ_t is the local time at the boundary ∂D and ν determines the direction of reflection
- under technical conditions (Cattiaux, 1988)[3], time reversed process is reflected at the boundary as well and solves

$$\mathrm{d}\overleftarrow{X}_t = -\overline{b}(t, \overleftarrow{X}_t)\,\mathrm{d}t + \sigma(\overleftarrow{X}_t)\,\mathrm{d}\overline{W}_t + \nu(\overleftarrow{X}_t)\,\mathrm{d}\overline{\ell}_t, \quad \overleftarrow{X}_0 \sim p_T$$

- retains Markov properties with constrained state space and specific Neumann boundary conditions

[3] Cattiaux (1988). **Time reversal of diffusion processes with a boundary condition**. *SPA*

Denoising Reflected Diffusion Models (DRDMs) in a nutshell

- extend DDMs by constraining both forward and backward processes to a bounded domain $D \subset \mathbb{R}^d$
- forward process includes a reflection term to enforce boundary constraints,

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t + \nu(X_t)\,\mathrm{d}\ell_t, \quad X_0 \in \overline{D},$$

  where ℓ_t is the local time at the boundary ∂D and ν determines the direction of reflection
- under technical conditions (Cattiaux, 1988)[3], time reversed process is reflected at the boundary as well and solves

$$\mathrm{d}\overleftarrow{X}_t = -\overline{b}(t, \overleftarrow{X}_t)\,\mathrm{d}t + \sigma(\overleftarrow{X}_t)\,\mathrm{d}\overline{W}_t + \nu(\overleftarrow{X}_t)\,\mathrm{d}\overline{\ell}_t, \quad \overleftarrow{X}_0 \sim p_T$$

- retains Markov properties with constrained state space and specific Neumann boundary conditions

[3] Cattiaux (1988). **Time reversal of diffusion processes with a boundary condition**. *SPA*

Comparison: DDMs versus their reflected counterparts

- **domain:** DDMs operate on $\mathbb{R}^d$, while DRDMs are constrained to a bounded domain $D \subset \mathbb{R}^d$

$\rightsquigarrow$ DRDMs include reflection terms to ensure dynamics remain in $\overline{D}$, while DDMs do not account for spatial constraints

- **implementation complexity:** DRDMs require managing boundary local times and Neumann conditions, introducing additional complexity
- **applications:** DRDMs are better suited for generating data confined to specific domains or bounded physical spaces

Source: Lou and Ermon (2023). Reflected Diffusion Models. *ICML*.

Generative modelling with reflected diffusions: Forward process

- assume that $D \subseteq \mathbb{R}^d$ is an open, connected and bounded set with $\mathcal{C}^\infty$ boundary ∂D
- consider the reflected time-homogeneous forward model

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t + \nu(X_t)\,\mathrm{d}\ell_t, \quad X_0 \in \overline{D},$$

with smooth and bounded coefficients $b\colon \overline{D} \to \mathbb{R}^d$, $\sigma\colon \overline{D} \to \mathbb{R}^{d\otimes d}$ and conormal reflection determined by

$$\nu(x) := \frac{1}{2}\sigma\sigma^\top(x)n(x), \quad x \in \partial D$$

⇝ n is the inward unit normal vector at the boundary ∂D, $(\ell_t)_{t\geq 0}$ is the local time at ∂D satisfying

$$\ell_t = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\int_0^t \mathbf{1}_{(\partial D)_\varepsilon}(X_s)\,\mathrm{d}s$$

⇝ boundary reflection process reflects X in a conormal direction whenever it hits the boundary ∂D, thus constraining the state space of the diffusion to the compact set $\overline{D}$

Forward process and SDE

- **key requirement:** precise understanding of the forward process's limiting behaviour to determine the runtime needed for the backward initialisation to approximate the true terminal forward distribution p_T
- **subsequent simplification:** choose $b = \nabla f$ and $\sigma = \sqrt{2f}\mathbb{I}_{d\times d}$ for some diffusivity $\mathcal{C}^\infty(\overline{D}) \ni f : \mathbb{R}^d \to [f_{\min}, \infty) \subset (0, \infty)$

$\rightsquigarrow$ time-homogeneous forward dynamics are described by the divergence form L^2-generator

$$\mathcal{A} = \nabla \cdot f\nabla = \langle \nabla f, \nabla \cdot \rangle + f\Delta,$$

corresponding to the constrained SDE

$$\mathrm{d}X_t = \nabla f(X_t)\,\mathrm{d}t + \sqrt{2f(X_t)}\,\mathrm{d}W_t + \nu(X_t)\,\mathrm{d}\ell_t$$

$\rightsquigarrow$ both the reflected forward and backward SDEs exhibit normal reflection at the boundary

- divergence theorem $\implies$ invariant distribution of the forward Markov process X is the (easy-to-sample-from) uniform distribution on $\overline{D}$, i.e., $\mu = \mathrm{Leb}\,|_{\overline{D}} / \mathrm{Leb}(\overline{D})$

Forward process and SDE

- **key requirement:** precise understanding of the forward process's limiting behaviour to determine the runtime needed for the backward initialisation to approximate the true terminal forward distribution p_T
- **subsequent simplification:** choose $b = \nabla f$ and $\sigma = \sqrt{2f}\mathbb{I}_{d\times d}$ for some diffusivity $\mathcal{C}^\infty(\overline{D}) \ni f : \mathbb{R}^d \to [f_{\min}, \infty) \subset (0, \infty)$

⇝ time-homogeneous forward dynamics are described by the divergence form L^2-generator

$$\mathcal{A} = \nabla \cdot f\nabla = \langle \nabla f, \nabla \cdot \rangle + f\Delta,$$

corresponding to the constrained SDE

$$\mathrm{d}X_t = \nabla f(X_t)\,\mathrm{d}t + \sqrt{2f(X_t)}\,\mathrm{d}W_t + \nu(X_t)\,\mathrm{d}\ell_t$$

⇝ both the reflected forward and backward SDEs exhibit normal reflection at the boundary

- divergence theorem $\Longrightarrow$ invariant distribution of the forward Markov process X is the (easy-to-sample-from) uniform distribution on $\overline{D}$, i.e., $\mu = \mathrm{Leb}\,|_{\overline{D}} / \mathrm{Leb}(\overline{D})$

Spectral properties

- under the given assumptions, there exist orthonormal eigenpairs $(\lambda_j, e_j)_{j\geq 0}$ of the operator $-\nabla \cdot f\nabla$ satisfying

$$0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots, \quad \lambda_j \asymp j^{2/d}$$

- transition densities can be expressed as

$$q_t(x,y) = \sum_{j\geq 0} \mathrm{e}^{-t\lambda_j} e_j(x) e_j(y), \quad x, y \in D$$

- bounds on eigenfunctions:

$$\|e_j\|_{H^k} \lesssim \lambda_j^{k/2} \asymp j^{k/d}, \quad j \geq 1, \qquad \|e_j\|_\infty \lesssim j^\tau, \text{ for } \tau > 1/2$$

$\rightsquigarrow$ smoothing property of densities:

$$\|p_t\|_{H^k} \lesssim \|p_0\|_\infty \mathrm{e}^{-tj^{2/d}} j^{\tau + k/d}, \quad t > 0,$$

for arbitrary $\tau > 1/2 \rightsquigarrow p_t \in \mathcal{C}^\infty(D)$ for any bounded initial density p_0

Backward process and score approximation

- backward dynamics becomes

$$\mathrm{d}\overleftarrow{X}_t = \left(\nabla f(\overleftarrow{X}_t) + 2f(\overleftarrow{X}_t)\nabla \log p_{T-t}(\overleftarrow{X}_t)\right)\mathrm{d}t + \sqrt{2f(\overleftarrow{X}_t)}\,\mathrm{d}\overline{W}_t + \nu(\overleftarrow{X}_t)\,\mathrm{d}\bar{\ell}_t,$$

  with initialisation $\overleftarrow{X}_0 \sim p_T$

- spectral decomposition of the transition densities $\implies$ score is explicitly given by

$$\nabla \log p_t(x) = \frac{\sum_{j\geq 0} \mathrm{e}^{-t\lambda_j}\langle p_0, e_j\rangle_{L^2}\nabla e_j(x)}{\sum_{j\geq 0} \mathrm{e}^{-t\lambda_j}\langle p_0, e_j\rangle_{L^2} e_j(x)}, \quad x \in D, t > 0$$

$\rightsquigarrow$ will be instrumental in analysing the score approximation properties of neural networks underlying the algorithm

Neural network classes

- use ReLU activation function $\sigma(y) = y \vee 0$, and, for any $b, x \in \mathbb{R}^m$, let $\sigma_b(x) = \begin{bmatrix} \sigma(x_1 - b_1) \\ \sigma(x_2 - b_2) \\ \vdots \\ \sigma(x_m - b_m) \end{bmatrix}$
- consider functions of the form

$$\varphi(x) = A_L \sigma_{b_L} A_{L-1} \sigma_{b_{L-1}} \cdots A_1 \sigma_{b_1} A_0 x,$$

where $A_i \in \mathbb{R}^{W_{i+1} \times W_i}$, $b_i \in \mathbb{R}^{W_{i+1}}$ for $i = 0, \ldots, L$, and where there are at most a total of S non-zero entries of the A_i's and b_i's and all entries are numerically at most B

$\rightsquigarrow$ class of networks

$$\Phi(L, W, S, B) := \left\{ \begin{array}{l} A_L \sigma_{b_L} A_{L-1} \sigma_{b_{L-1}} \cdots A_1 \sigma_{b_1} A_0 \mid A_i \in \mathbb{R}^{W_{i+1} \times W_i}, b_i \in \mathbb{R}^{W_{i+1}}, \\ \sum_{i=0}^{L} (\|A_i\|_0 + \|b_i\|_0) \leq S, \max_{i \in \{0, \ldots, L\}} (\|A_i\|_\infty \vee \|b_i\|_\infty) \leq B \end{array} \right\}$$

Generative modeling with reflected diffusions

- denote the true score by $\mathfrak{s}^\circ(x,t) := \nabla \log p_t(x)$, and assume we are given data samples $(X_{0,i})_{i\in[n]} \overset{\text{i.i.d.}}{\sim} p_0$
- for a hypothesis class $\mathcal{S}$ of neural networks and $\mathfrak{s} \in \mathcal{S} \cup \{\mathfrak{s}^\circ\}$, define

$$L_{\mathfrak{s}}(x) := \mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} |\mathfrak{s}(X_t,t) - \nabla_y \log q_t(x,X_t)|^2 \mid X_0 = x\Big]$$

  - ⇝ $\overline{T}$ is the terminal runtime of the reflected forward process
  - ⇝ $\underline{T} \in (0,\overline{T})$ is such that we run the reflected generative process, which is initialised with distribution $\mathcal{U}(D)$, until $\overline{T} - \underline{T}$
- denote the empirical denoising score matching loss associated to $\mathfrak{s}$ by

$$\hat{L}_{\mathfrak{s},n} := \frac{1}{n}\sum_{i=1}^{n} L_{\mathfrak{s}}(X_{0,i}),$$

  and define the empirical score minimiser by

$$\hat{\mathfrak{s}}_n := \operatorname*{arg\,min}_{\mathfrak{s}\in\mathcal{S}} \hat{L}_{\mathfrak{s},n}$$

- let $\overline{X}^{\mathfrak{s}}$ be a solution of the reflected SDE

$$\mathrm{d}\overline{X}_t^{\mathfrak{s}} = \big(\nabla f(\overline{X}_t^{\mathfrak{s}}) + 2f(\overline{X}_t^{\mathfrak{s}})s(\overline{X}^{\mathfrak{s}}, t)\big)\,\mathrm{d}t + \sqrt{2f(\overline{X}_t^{\mathfrak{s}})}\,\mathrm{d}\overline{W}_t + \nu(\overline{X}_t^{\mathfrak{s}})\,\mathrm{d}\bar{\ell}_t, \quad t \in [0, \overline{T}, -\underline{T}],$$
$$\overline{X}_0^{\mathfrak{s}} \sim \mathcal{U}(\overline{D}),$$

for some Brownian motion $(\overline{W}_t)_{t\in[0,\overline{T}-\underline{T}]}$ and local time $(\bar{\ell}_t)_{t\in[0,\overline{T}-\underline{T}]}$ at the boundary ∂D, and denote its density at time t by $\overleftarrow{p}_t^{\mathfrak{s}}$

- initialisation $X_0^{\hat{\mathfrak{s}}_n} \sim \mathcal{U}(\overline{D})$

⇝ $(\overleftarrow{p}_t^{\hat{\mathfrak{s}}_n})_{t\in[0,\overline{T})}$ are the densities of the backward process driven by the score estimate $\hat{\mathfrak{s}}_n$

⇝ assessing the quality of the generated samples boils down to analysing the distance between the distribution induced by p_0 and the (random) distribution induced by $\overleftarrow{p}_{\overline{T}-\underline{T}}^{\hat{\mathfrak{s}}_n}$

- let $\overline{X}^{\mathfrak{s}}$ be a solution of the reflected SDE

$$\mathrm{d}\overline{X}_t^{\mathfrak{s}} = \big(\nabla f(\overline{X}_t^{\mathfrak{s}}) + 2f(\overline{X}_t^{\mathfrak{s}})s(\overline{X}^{\mathfrak{s}}, t)\big)\,\mathrm{d}t + \sqrt{2f(\overline{X}_t^{\mathfrak{s}})}\,\mathrm{d}\overline{W}_t + \nu(\overline{X}_t^{\mathfrak{s}})\,\mathrm{d}\bar{\ell}_t, \quad t \in [0, \overline{T}, -\underline{T}],$$
$$\overline{X}_0^{\mathfrak{s}} \sim \mathcal{U}(\overline{D}),$$

for some Brownian motion $(\overline{W}_t)_{t\in[0,\overline{T}-\underline{T}]}$ and local time $(\bar{\ell}_t)_{t\in[0,\overline{T}-\underline{T}]}$ at the boundary ∂D, and denote its density at time t by $\overleftarrow{p}_t^{\mathfrak{s}}$

- initialisation $X_0^{\hat{\mathfrak{s}}_n} \sim \mathcal{U}(\overline{D})$

⇝ $(\overleftarrow{p}_t^{\hat{\mathfrak{s}}_n})_{t\in[0,\overline{T})}$ are the densities of the backward process driven by the score estimate $\hat{\mathfrak{s}}_n$

⇝ assessing the quality of the generated samples boils down to analysing the distance between the distribution induced by p_0 and the (random) distribution induced by $\overleftarrow{p}_{\overline{T}-\underline{T}}^{\hat{\mathfrak{s}}_n}$

Main result

Theorem (Holk, CS and LT (2024))

Assume $p_0 = \tilde{p}_0 + \alpha$ with $\tilde{p}_0 \in H^s_c(D)$, $\alpha > 0$, $s \in \mathbb{N} \cap (d/2, \infty)$, and let

$$\underline{T} \asymp n^{-\frac{2s}{((2-d/s)\wedge 1)(2s+d)}}, \quad \overline{T} = \frac{s}{\lambda_1(2s+d)} \log n.$$

Then, there exists a class of feed forward ReLU neural networks $\mathcal{S}$, with explicit size constraints in terms of n, d and s, such that

$$\mathbb{E}\left[\mathrm{TV}(p_0, \overleftarrow{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\right] \lesssim n^{-\frac{s}{2s+d}} (\log n)^3 (\log\log n)^{1/2}.$$

Letting $\overline{p}^{\mathfrak{s}}_t$ be the density at time t of the time-reversed forward process

$$\mathbb{E}\left[\mathrm{TV}(p_0, \overleftarrow{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\right] \leq \underbrace{\mathrm{TV}(p_0, p_{\underline{T}})}_{=: \mathbf{(I)}} + \underbrace{\mathrm{TV}(\mathbb{P}(X_{\overline{T}} \in \cdot \mid X_0 \sim p_0), \mathcal{U}(\overline{D}))}_{=: \mathbf{(II)}} + \underbrace{\mathbb{E}\left[\mathrm{TV}(\overline{p}^{\mathfrak{s}^\circ}_{\overline{T}-\underline{T}}, \overline{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\right]}_{=: \mathbf{(III)}}.$$

Main result

Theorem (Holk, CS and LT (2024))

Assume $p_0 = \tilde{p}_0 + \alpha$ with $\tilde{p}_0 \in H^s_c(D)$, $\alpha > 0$, $s \in \mathbb{N} \cap (d/2, \infty)$, and let

$$\underline{T} \asymp n^{-\frac{2s}{((2-d/s)\wedge 1)(2s+d)}}, \quad \overline{T} = \frac{s}{\lambda_1(2s+d)} \log n.$$

Then, there exists a class of feed forward ReLU neural networks $\mathcal{S}$, with explicit size constraints in terms of n, d and s, such that

$$\mathbb{E}\left[\mathrm{TV}(p_0, \overleftarrow{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\right] \lesssim n^{-\frac{s}{2s+d}} (\log n)^3 (\log \log n)^{1/2}.$$

Letting $\overline{p}^{\mathfrak{s}}_t$ be the density at time t of the time-reversed forward process

$$\mathbb{E}\left[\mathrm{TV}(p_0, \overleftarrow{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\right] \le \underbrace{\mathrm{TV}(p_0, p_{\underline{T}})}_{=: \mathbf{(I)}} + \underbrace{\mathrm{TV}(\mathbb{P}(X_{\overline{T}} \in \cdot \mid X_0 \sim p_0), \mathcal{U}(\overline{D}))}_{=: \mathbf{(II)}} + \underbrace{\mathbb{E}\left[\mathrm{TV}(\overline{p}^{\mathfrak{s}^\circ}_{\overline{T}-\underline{T}}, \overline{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\right]}_{=: \mathbf{(III)}}.$$

Error decomposition

$$\mathbb{E}\big[\mathrm{TV}(p_0, \overleftarrow{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\big] \leq \underbrace{\mathrm{TV}(p_0, p_{\underline{T}})}_{=:\,\mathbf{(I)}} + \underbrace{\mathrm{TV}(\mathbb{P}(X_{\overline{T}} \in \cdot \mid X_0 \sim p_0), \mathcal{U}(\overline{D}))}_{=:\,\mathbf{(II)}} + \underbrace{\mathbb{E}\big[\mathrm{TV}(\overline{p}^{\mathfrak{s}^\circ}_{\overline{T}-\underline{T}}, \overline{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\big]}_{=:\,\mathbf{(III)}}$$

(I) represents the error induced by stopping early the backward process initialised by the true forward terminal density $p_{\overline{T}}$ at time $\overline{T} - \underline{T}$

- ⇝ controlled via small time heat kernel bounds for the transition densities
- ⇝ relies on Hölder continuity of p_0: for $\beta \in [(2 - d/s) \wedge 1, 1]$,

$$|p_0(x) - p_0(y)| \leq c_\beta |x-y|^\beta, \quad x, y \in D$$

Lemma (Holk, CS and LT (2024))

There exists a constant C depending only on f, d, D, β and c_β such that

$$\mathrm{TV}(p_0, p_{\underline{T}}) \leq C\underline{T}^{\beta/2}, \qquad \underline{T} \leq 1.$$

Error decomposition

$$\mathbb{E}\Big[\mathrm{TV}\big(p_0, \overleftarrow{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}}\big)\Big] \leq \underbrace{\mathrm{TV}(p_0, p_{\underline{T}})}_{=:\,(\mathbf{I})} + \underbrace{\mathrm{TV}(\mathbb{P}(X_{\overline{T}} \in \cdot \mid X_0 \sim p_0), \mathcal{U}(\overline{D}))}_{=:\,(\mathbf{II})} + \underbrace{\mathbb{E}\Big[\mathrm{TV}\big(\overline{p}^{\mathfrak{s}^\circ}_{\overline{T}-\underline{T}}, \overline{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}}\big)\Big]}_{=:\,(\mathbf{III})}$$

(II) is the error associated to starting the backward process in its stationary distribution instead of $p_{\overline{T}}$

$\rightsquigarrow$ controlled in terms of the spectral gap λ_1 of $\mathcal{A}$, which can be lower bounded by $\lambda_1 \geq f_{\min}/C_{\mathrm{P}}(D)$, where $C_{\mathrm{P}}(D)$ is the Poincaré constant of the domain D

Lemma (Holk, CS and LT (2024))

It holds that

$$\mathrm{TV}(\mathbb{P}(X_{\overline{T}} \in \cdot \mid X_0 \sim p_0), \mathcal{U}(\overline{D})) \leq \frac{\sqrt{\mathrm{Leb}(D)}}{2} \|p_0\|_{L^2} \mathrm{e}^{-\lambda_1 \overline{T}}, \quad \overline{T} > 0.$$

Error decomposition

$$\mathbb{E}\big[\mathrm{TV}(p_0, \overleftarrow{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\big] \leq \underbrace{\mathrm{TV}(p_0, p_{\underline{T}})}_{=: \textbf{(I)}} + \underbrace{\mathrm{TV}(\mathbb{P}(X_{\overline{T}} \in \cdot \mid X_0 \sim p_0), \mathcal{U}(\overline{D}))}_{=: \textbf{(II)}} + \underbrace{\mathbb{E}\big[\mathrm{TV}(\overline{p}^{\mathfrak{s}^\circ}_{\overline{T}-\underline{T}}, \overline{p}^{\hat{\mathfrak{s}}_n}_{\overline{T}-\underline{T}})\big]}_{=: \textbf{(III)}}$$

(III) quantifies the error coming from running the backward process with the drift determined by the estimated score $\hat{\mathfrak{s}}_n$ instead of the true score $\mathfrak{s}^\circ$

$\rightsquigarrow$ by Girsanov's theorem and Pinsker's inequality, controlled by

$$\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} \int_D |\hat{\mathfrak{s}}(x,t) - \nabla \log p_t(x)|^2 p_t(x) \, \mathrm{d}x \, \mathrm{d}t\Big]$$

$\rightsquigarrow$ key to bounding it: equivalence between explicit and denoising score matching, i.e.,

$$\int_{\underline{T}}^{\overline{T}} \int_D |\mathfrak{s}(y,t) - \nabla \log p_t(y)|^2 p_t(y) \, \mathrm{d}y \, \mathrm{d}t = \mathbb{E}[L_{\mathfrak{s}}(X_0)] + C,$$

where $C \leq 0$ is a constant that is independent of $\mathfrak{s}$

Bounding the score matching error (III)

Generalisation loss can be bounded in terms of the minimal score approximation error over the class $\mathcal{S}$ and the complexity of the induced function class $\mathcal{L} := \{L_{\mathfrak{s}} : \mathfrak{s} \in \mathcal{S}\}$ for a desired precision level δ:

Theorem (Oko et. al (2023))

Suppose that $\sup_{\mathfrak{s}\in\mathcal{S}} \|L_{\mathfrak{s}} - L_{\mathfrak{s}^\circ}\|_\infty \leq C(\mathcal{L}) < \infty$. Then, for any $\delta > 0$ such that $\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta) \geq 3$, it holds that

$$\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} \int_{\underline{D}} |\hat{\mathfrak{s}}(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big]$$
$$\leq 2 \inf_{\mathfrak{s}\in\mathcal{S}} \int_{\underline{T}}^{\overline{T}} \int_{D} |\mathfrak{s}(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t + 2\frac{C(\mathcal{L})}{n}\Big(\frac{37}{9}\log \mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta) + 32\Big) + 3\delta.$$

- $\rightsquigarrow$ **next step:** control both the uniform loss upper bound $C(\mathcal{L})$ and the covering number $\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta)$ for $\delta = n^{-2s/(2s+d)}$ $\Big[\checkmark$ since $\log \mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta) \leq \log \mathcal{N}\big(\Phi(L, W, S, B), \|\cdot\|_\infty, \frac{\delta}{C\overline{T}}\big) \lesssim LS \log\Big(\frac{LWB\overline{T}}{\delta}\Big)\Big]$
- $\rightsquigarrow$ **final and most fundamental question**: treatment of the explicit score approximation error

Strategy for bounding the approximation error

1. Truncate

$$p_t(x) = \sum_{j=0}^{\infty} \mathrm{e}^{-\lambda_j t} \langle p_0, e_j \rangle e_j(x) \approx \sum_{j=0}^{N} \mathrm{e}^{-\lambda_j t} \langle p_0, e_j \rangle e_j(x) =: h_N(x, t)$$

   and use $\nabla h_N(x,t) \approx \nabla p_t(x)$

$$\rightsquigarrow \int_{\underline{T}}^{\overline{T}} \int_{\mathbb{R}^d} \left| s^\circ(x,t) - \frac{\nabla h_N(x,t)}{h_N(x,t)} \right|^2 p_t(x) \,\mathrm{d}x \,\mathrm{d}t \lesssim N^{-2s/d} \log \underline{T} \implies N \asymp n^{d/(2s+d)}$$

2. for an appropriately chosen discrete set of time points $\{t_i\}$, use the spatial smoothness of $h_N(x, t_i)$ induced by the Sobolev smoothness of p_0 to obtain an efficient neural network approximation of $h_N(\cdot, t_i)$, based on general approximation results from Suzuki (2019)[4];
3. approximate the space-time functions $h_N(x,t)$, $\nabla h_N(x,t)$ by constructing a neural network approximation of a polynomial time interpolation of the neural networks from Step 2., where the interpolation degree is adapted to the parameters N, s and d.

[4] Suzuki (2019). **Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality**. *ICLR*

Strategy for bounding the approximation error

1. Truncate

$$p_t(x) = \sum_{j=0}^{\infty} \mathrm{e}^{-\lambda_j t} \langle p_0, e_j \rangle e_j(x) \approx \sum_{j=0}^{N} \mathrm{e}^{-\lambda_j t} \langle p_0, e_j \rangle e_j(x) =: h_N(x,t)$$

   and use $\nabla h_N(x,t) \approx \nabla p_t(x)$

$$\rightsquigarrow \int_{\underline{T}}^{\overline{T}} \int_{\mathbb{R}^d} \left| s^\circ(x,t) - \frac{\nabla h_N(x,t)}{h_N(x,t)} \right|^2 p_t(x)\, \mathrm{d}x\, \mathrm{d}t \lesssim N^{-2s/d} \log \underline{T} \implies N \asymp n^{d/(2s+d)}$$

2. for an appropriately chosen discrete set of time points $\{t_i\}$, use the spatial smoothness of $h_N(x, t_i)$ induced by the Sobolev smoothness of p_0 to obtain an efficient neural network approximation of $h_N(\cdot, t_i)$, based on general approximation results from Suzuki (2019)[4];
3. approximate the space-time functions $h_N(x,t)$, $\nabla h_N(x,t)$ by constructing a neural network approximation of a polynomial time interpolation of the neural networks from Step 2., where the interpolation degree is adapted to the parameters N, s and d.

[4] Suzuki (2019). **Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality**. *ICLR*

Strategy for bounding the approximation error

1. Truncate

$$p_t(x) = \sum_{j=0}^{\infty} \mathrm{e}^{-\lambda_j t}\langle p_0, e_j\rangle e_j(x) \approx \sum_{j=0}^{N} \mathrm{e}^{-\lambda_j t}\langle p_0, e_j\rangle e_j(x) =: h_N(x,t)$$

   and use $\nabla h_N(x,t) \approx \nabla p_t(x)$

$$\rightsquigarrow \int_{\underline{T}}^{\overline{T}} \int_{\mathbb{R}^d} \left| s^\circ(x,t) - \frac{\nabla h_N(x,t)}{h_N(x,t)} \right|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \lesssim N^{-2s/d} \log \underline{T} \implies N \asymp n^{d/(2s+d)}$$

2. for an appropriately chosen discrete set of time points $\{t_i\}$, use the spatial smoothness of $h_N(x,t_i)$ induced by the Sobolev smoothness of p_0 to obtain an efficient neural network approximation of $h_N(\cdot, t_i)$, based on general approximation results from Suzuki (2019)[4];
3. approximate the space-time functions $h_N(x,t)$, $\nabla h_N(x,t)$ by constructing a neural network approximation of a polynomial time interpolation of the neural networks from Step 2., where the interpolation degree is adapted to the parameters N, s and d.

[4] Suzuki (2019). **Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality**. *ICLR*

Theorem (Holk, CS and LT (2024))

Let $0 < \underline{T} < \overline{T}$ and $n \in \mathbb{N}$ sufficiently large be given with $\underline{T} \in \mathrm{Poly}(n^{-1})$. Then, there exists a neural network $\mathfrak{s} \in \Phi(L(n), W(n), S(n), B(n))$ satisfying

$$\int_{\underline{T}}^{\overline{T}} \int_D \left|\mathfrak{s}(x,t) - \nabla_x \log p_t(x)\right|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \lesssim n^{-\frac{2s}{2s+d}} (\log n)^2 (\overline{T} + \log(\underline{T}^{-1})).$$

The size of the network is evaluated as

$$\begin{aligned} L(n) &\lesssim \log n \log\log n, \\ \|W(n)\|_\infty &\lesssim M n^{\frac{d}{2s+d}} \log n, \\ S(n) &\lesssim M n^{\frac{d}{2s+d}} (\log n)^2, \quad \text{and} \\ B(n) &\lesssim n^{\frac{1}{2s+d}} \vee \frac{1}{\underline{T}}, \end{aligned}$$

where $M \in O(|\log \frac{\overline{T}}{\underline{T}}|)$. Furthermore, the network can be chosen such that there exists a constant $C < \infty$ depending only on p_0 and D such that $|\mathfrak{s}(x,t)| \le \frac{C}{\sqrt{t}}$ for all $t \in [\underline{T}, \overline{T}]$ and $x \in D$.

Future Research

- extending the DRDM framework to data supported on lower-dimensional submanifolds ⇝ challenging because L^2-techniques don't translate naturally, but explicit form of Skorokhod map for reflected BM in $D = [0,1]^d$ gives a possible starting point
- Use score approximation techniques to unify the statistical analysis in the framework of denoising Markov models (Benton et. al., 2024)[5] for appropriate self-adjoint forward Markov processes

⇝ efficient sampling methods for the generative reflected process with estimated score?

  - natural sampling schemes for reflected diffusions combine an Euler–Maryuama discretisation with a projection (Słomiński, 1994)[6] or rejection (Fishman et al., 2024)[7] step

- how to alter the training objective to avoid training data replication while maintaining statistical optimality?
  - Vardanyan et. al (2024)[8] suggest penalised Wasserstein-1-GAN generator obtained from

$$\hat{g}_n \in \arg\min_{g \in \mathcal{G}} \Big\{ W_1\Big(g \sharp \mathcal{U}_d, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\Big) + \lambda \min_{h \in \mathcal{H}} \int_{[0,1]^d} |h \circ g(u) - u|^2 \, \mathrm{d}u \Big\},$$

[5] Benton, Shi, De Bortoli, Delegiannidis and Doucet (2024). **From denoising diffusions to denoising Markov models**. *JRSS B*

[6] Słomiński (1994). **On approximation of solutions of multidimensional SDE's with reflecting boundary conditions**. *SPA*

[7] Fishman, Klarner, Mathieu, Hutchinson and De Bortoli (2024). **Metropolis Sampling for Constrained Diffusion Models**. *NeurIPS*

[8] Vardanyan, Hunanyan, Galstyan, Minasyan and Dalalyan (2024). **Statistically Optimal Generative Modeling with Maximum Deviation from the Empirical Distribution**. *ICML*

Future Research

- extending the DRDM framework to data supported on lower-dimensional submanifolds $\rightsquigarrow$ challenging because L^2-techniques don't translate naturally, but explicit form of Skorokhod map for reflected BM in $D = [0,1]^d$ gives a possible starting point
- Use score approximation techniques to unify the statistical analysis in the framework of denoising Markov models (Benton et. al., 2024)[5] for appropriate self-adjoint forward Markov processes

$\rightsquigarrow$ efficient sampling methods for the generative reflected process with estimated score?

  - natural sampling schemes for reflected diffusions combine an Euler–Maryuama discretisation with a projection (Słomiński, 1994)[6] or rejection (Fishman et al., 2024)[7] step

- how to alter the training objective to avoid training data replication while maintaining statistical optimality?
  - Vardanyan et. al (2024)[8] suggest penalised Wasserstein-1-GAN generator obtained from

$$\hat{g}_n \in \arg\min_{g \in \mathcal{G}} \Big\{ W_1\Big(g \sharp \mathcal{U}_d, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\Big) + \lambda \min_{h \in \mathcal{H}} \int_{[0,1]^d} |h \circ g(u) - u|^2 \, \mathrm{d}u \Big\},$$

[5] Benton, Shi, De Bortoli, Delegiannidis and Doucet (2024). **From denoising diffusions to denoising Markov models**. *JRSS B*

[6] Słomiński (1994). **On approximation of solutions of multidimensional SDE's with reflecting boundary conditions**. *SPA*

[7] Fishman, Klarner, Mathieu, Hutchinson and De Bortoli (2024). **Metropolis Sampling for Constrained Diffusion Models**. *NeurIPS*

[8] Vardanyan, Hunanyan, Galstyan, Minasyan and Dalalyan (2024). **Statistically Optimal Generative Modeling with Maximum Deviation from the Empirical Distribution**. *ICML*

Future Research

- extending the DRDM framework to data supported on lower-dimensional submanifolds $\rightsquigarrow$ challenging because L^2-techniques don't translate naturally, but explicit form of Skorokhod map for reflected BM in $D = [0,1]^d$ gives a possible starting point
- Use score approximation techniques to unify the statistical analysis in the framework of denoising Markov models (Benton et. al., 2024)[5] for appropriate self-adjoint forward Markov processes

$\rightsquigarrow$ efficient sampling methods for the generative reflected process with estimated score?

  - natural sampling schemes for reflected diffusions combine an Euler–Maryuama discretisation with a projection (Słomiński, 1994)[6] or rejection (Fishman et al., 2024)[7] step

- how to alter the training objective to avoid training data replication while maintaining statistical optimality?
  - Vardanyan et. al (2024)[8] suggest penalised Wasserstein-1-GAN generator obtained from

$$\hat{g}_n \in \arg\min_{g \in \mathcal{G}} \Big\{ \mathrm{W}_1\Big(g\sharp\,\mathcal{U}_d, \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \Big) + \lambda \min_{h \in \mathcal{H}} \int_{[0,1]^d} |h \circ g(u) - u|^2 \,\mathrm{d}u \Big\},$$

[5] Benton, Shi, De Bortoli, Delegiannidis and Doucet (2024). **From denoising diffusions to denoising Markov models**. *JRSS B*

[6] Słomiński (1994). **On approximation of solutions of multidimensional SDE's with reflecting boundary conditions**. *SPA*

[7] Fishman, Klarner, Mathieu, Hutchinson and De Bortoli (2024). **Metropolis Sampling for Constrained Diffusion Models**. *NeurIPS*

[8] Vardanyan, Hunanyan, Galstyan, Minasyan and Dalalyan (2024). **Statistically Optimal Generative Modeling with Maximum Deviation from the Empirical Distribution**. *ICML*